This is the author’s accepted manuscript.

Please do not cite this version. The published version can be found at:

<https://www.tandfonline.com/doi/abs/10.1080/09515089.2019.1688776?journalCode=cphp20>

**Haidt et al.’s Case for Moral Pluralism Revisited[[1]](#footnote-1)**

Recent work in moral psychology that claims to show that human beings make moral judgements on the basis of multiple, divergent moral foundations has been influential in both moral psychology and moral philosophy. Primarily, such work has been taken to undermine monistic moral theories, especially those pertaining to the prevention of harm. Here, I call one of the most prominent and influential empirical cases for moral pluralism into question, namely that of Jonathan Haidt and his colleagues. I argue that Haidt et al.’s argument is not as strong as it is often made out to be, given significant problems with the design of one of the key experiments used to ground the claim that there are divergent moral foundations across cultures. The flaws that I point out pose a significant challenge to Haidt et al.’s findings and have a detrimental impact on subsequent work based on this immensely influential experiment. Accordingly, I argue that both empirical and normative claims made on the basis of Haidt et al.’s findings should be treated with caution. I conclude by making some suggestions as to how some of the problems that I point out might be addressed.

Keywords: Haidt; moral pluralism; harm; Moral Foundations Theory; moral psychology moral philosophy

# 1. Introduction

Much has been made of the potential ethical implications of moral psychology’s research into the moral judgements that people make.[[2]](#footnote-2) Jonathan Haidt (2013) has gone so far as to describe moral psychology as temporarily removing ethics from the hands of philosophers and biologising it. Haidt and his colleagues’ own work has certainly been influential with some philosophers (e.g. Weinberg, et al., 2001; Prinz 2006; 2007a; 2007b; Sinnott-Armstrong and Wheatly 2012; Sinnott-Armstrong and Wheatley 2014; Antiel, Humeniuk and Tilbert 2014). One result of this influence is the prevalence of the claim that morality is disunified, based on the belief that Haidt et al.’s research shows that there are many different grounds on which people tend to make moral judgements. The claim is made that these results point to an inherent human tendency towards moral pluralism and that this finding should call into question philosophical arguments in favour of moral monism, especially when it comes to moral norms based on harm considerations, which is characterised as being informed by liberal and western biases.

Here, I want to repudiate such philosophical claims, showing that Haidt et al.’s case for moral pluralism is far from compelling. If I am right, any normative claims regarding moral pluralism based on this (enormously influential) research are significantly undermined.[[3]](#footnote-3)

1. Moral psychology and moral philosophy

As already mentioned, one of the most influential cases for the disunity of morality, both within moral psychology and beyond, is made by Jonathan Haidt and his colleagues. They claim to show that some people tend to moralise behaviours other than those pertaining to harm, such as behaviours relating to community, purity, and divinity considerations.[[4]](#footnote-4) From this, they sometimes draw a normative conclusion: There are behaviours over and above those pertaining to harm that *should* be moralised.[[5]](#footnote-5) Haidt has been advocating for a “broadened” understanding of what morality entails, which holds that “[t]here’s more to morality than harm and fairness” (e.g. Haidt & Graham 2007; Haidt 2012).

As already mentioned, it is not only moral psychologists who have been drawing normative conclusions based on Haidt's research; philosophers have been following suit, although it is not always clear what the philosophical up-shot of these arguments in favour of moral pluralism is supposed to be. Sinnott-Armstrong and Wheatley (2014) argue that their findings should encourage moral scientists to develop a “more taxonomic approach” to morality and to stop approaching it in a monolithic way. Prinz (2006; 2007a; 2007b), criticises normative ethics for abandoning our “ordinary intuitions” about what wrongness entails and for building general theories based on a limited number of cases (i.e. harm norms) (Prinz 2007a; 2007b). Generally, the idea seems to be that a normative ethics based on harm norms is illegitimate, given that the empirical research shows that humans do not tend to moralise on these grounds only. Irrespective of the question as to whether or not ethics can be naturalised and whether or not empirical considerations have normative implications, I want to dispute calls for moral pluralism based on Haidt et al.’s research, on the grounds that the empirical case in favour of their brand of moral pluralism is not as strong as it is made out to be. This is partly due to problems in the design of one of the seminal experiments used to support the contention that humans moralise on pluralistic bases, namely, “Affect, culture, and morality, or is it wrong to eat your dog” (hereafter “Affect”) (Haidt et al. 1993). Because this study has informed and set the tone for much subsequent work by Haidt et al. (and others) on the pluralistic foundations that ostensibly underlie our moralising, I contend that the problems that I highlight has had a significant detrimental impact on this research. [[6]](#footnote-6) This consideration weakens considerably any pluralist normative conclusions drawn on the basis of this, and affiliated, research.

1. The basis for the disunity thesis

The crux of Haidt et al.’s argument for moral pluralism is that their research shows that different kinds of issues are sometimes moralised by different people, which is put down to cultural differences in the cognitive foundations on the basis of which such moral judgements are made. A key piece of evidence for this claim is research undertaken into the apparent bases for moral judgements made by research participants of different socio-economic classes from both the United States of America and Brazil in “Affect” (Haidt et al. 1993). Here, apparent differences in the moral judgements made by the different socio-economic groups are found, and it is concluded that the groups have different underlying cognitive moral foundations. “Affect” is not only cited by various philosophers who make a normative case for moral pluralism[[7]](#footnote-7), but it is partly as a result of the conclusions drawn here that Haidt and various collaborators go on to propose Moral Foundations Theory (MFT) (Haidt & Joseph 2004; Haidt & Joseph 2007; Graham, Haidt et al. 2009; Iyer, Koleva et al. 2012; Graham, Haidt et al. 2013). [[8]](#footnote-8);[[9]](#footnote-9) In turn, MFT underlies many of the current pluralist assumptions within moral psychology and of the philosophical positions influenced by them (e.g. Henrich, 2010; Kim, et al., 2012; Demetriou, 2013; Weston, 2014; McKay, 2015; Silver, 2017).Moreover, as will be discussed in more detail below, Haidt and his colleagues (Koleva & Haidt, 2012; Haidt, 2015) refer to the findings in this paper in order to refute criticism of MFT by Gray et al. (Gray, Young et al. 2012; Gray, Schein et al., 2014; Schein & Gray, 2015; Schein & Gray, 2017) as well as to counter Gary et al.’s contention that their own research shows that all moral judgements are made on the basis of (perceived) harm inflicted on a patient by an agent.[[10]](#footnote-10) In essence, Haidt and his colleagues claim that the experiment reported in “Affect” shows that some cultures moralise on other bases in addition to harm, which is something that any harm-based moral theory needs to contend with. By arguing that Haidt et al. (1993) do not manage to show that participants moralise on bases other than harm, I show that this paper cannot be mustered as evidence on the basis of which to postulate multiple moral foundations, nor to refute criticism of such proposed foundations. Furthermore, I hold that the flaws in this study can be found in other research used to undergird MFT. As a result, given the far-reaching influence of MFT, I conclude that a significant part of this empirical case for multiple moral foundations is weaker than it is often made out to be.

“Affect” presents a series of vignettes to separate groups of low and high socio-economic status (SES) in three cities—two in Brazil and one on the USA. The hypothesis was that “westernised” participants would make different moral judgements from “non-westernised” participants, based on the assumption that people from industrialised, western countries tend to set greater store in autonomy and harm than those from other countries (p. 616).[[11]](#footnote-11) Research subjects were presented with stories containing actions that the researchers considered to be disrespectful or disgusting but “harmless”, since no victims were purportedly affected (i.e. “harmless but offensive”) (p. 613). The aim was to elicit moralising stances towards these actions and to look for evidence of a broader morality than that of upper-middle-class North Americans, which was construed as being harm-based. The researchers claimed to test whether or not participants did, in fact, perceive harm to be taking place in these “harmless” scenarios by presenting them with a “Harm probe”. The assumption was that participants with a harm-based morality would take a permissive stance towards the “harmless” actions. If there were those who moralised these actions, it would be taken as proof that they were forming moral judgements on bases other than harm considerations, thus proving that there are a plurality of cognitive moral foundations.

In the three “disgusting” stories i) a man masturbates with a dead chicken and then cooks it for dinner; ii) a family eat their pet dog after it dies in an accident; and iii) a brother and sister kiss each other passionately on the mouth. In the “disrespect” stories iv) a woman uses an old Brazilian/American flag to clean her bathroom; and v) a son promises his dying mother to visit her grave every week, but ends up breaking his promise. Three control stories were also presented to participants: vi) a girl wants to use a swing, so she pushes a boy off and hurts him (described as a “a prototypical moral violation, because it involves direct physical harm to an innocent victim” (p. 617); vii) a boy wears regular clothes to school, even though the school requires students to wear a uniform, and viii) a man eats all his food with his hands, in public and in private, after washing them (the last two stories are described as “prototypical conventional violations, because they involve no intrinsic harm to others” (ibid.)).

The researchers found that the majority of high-SES Philadelphians took a permissive stance toward the “harmless-offensive” stories (i-v), while in Recife (the least westernised of the Brazilian cities), the majority of low-SES subjects took a moralising stance towards these stories. There was a significant and consistent effect of social class, where high-SES groups were more permissive than low-SES groups.

Haidt et al. (1993, p. 62) conclude the following, based on their study:

The domain of morality appears to vary cross-culturally. Philadelphians of high SES exhibited a harm-based morality limited to the ethics of autonomy. Disgusting and disrespectful actions were not moralized, as long as these actions were perceived to have no harmful interpersonal consequences. But in low SES groups, and especially in Brazil, morality appears to be broader. Stories that involved disgust and disrespect were moralized,even when they were perceived to be harmless.

Essentially, the argument is that the discrepancy between the high and low SES groups indicates that they employ different moral foundations, with the high SES groups employing a harm-based foundation and the lower SES groups employing additional foundations, including authority and divinity considerations. The researchers conclude that, cross-culturally, harm is not the only basis on which behaviour is moralised. These findings do not necessarily bear up under closer scrutiny, however. In what follows, I will address questionable aspects of the research design and conclude that the researchers do not manage to show that some respondents moralised on bases other than harm considerations.

1. Interpersonal “hurting” and the assumption that offence does not constitute a harm

It should be clear from the description above that the researchers’ argument relies on plausibly establishing that the low SES participants believed the scenarios in question to be harmless even while going on the moralise them. I will argue they do not manage to do so. At base, the following three problems affect their study: i) an insufficiently precise conception of “harm” informing the empirical research (this seems mainly to vacillate between physical “hurting” and some variety of direct interpersonal psychological injury); ii) inadequate measures to test for the conception of harm that the research subjects themselves employ; and iii) an inability to account for the research subjects’ metaphysical presuppositions regarding what or who can be harmed and how.

With regard to their own conception of “harm”, the researchers clearly understand it as occurring interpersonally, as is seen in the quote given in the previous section. It is less clear from the design and discussion of the experiment what kind of interpersonal consequences they have in mind. These definitely include physical hurting, as this is presented as a “prototypical” moral violation in scenario vi). However, whether psychological and other kinds of negative interpersonal consequences are included in their conception is less clear. On the one hand, it is explicitly stated that psychological harm is possible (1993, p. 613), on the other, the research design itself excludes “offence” and “being bothered”— both candidates for psychological harm—from being harms in this context (more on this below). Significantly, their exclusive focus on harm as only entailing “interpersonal consequences”—which, ironically, they consider to be a western conception of harm—causes them to not test for the possibility that the research subjects conceived of harm that does not only take place interpersonally (i.e. between individuals). In other words, while it is possible that the research subjects may have adhered to a metaphysical worldview in which extra-personal entities such as God or the community are liable to harm, the researchers do not allow for this possibility. This omission may be due to their own metaphysical worldview and resultant biases as to what “harm” entails.

Hence, in addition to their own muddled understanding of the concept, the researchers do not establish how the research participants themselves conceive of “harm”. As a result, the researchers at most established that some of their participants moralised on grounds other than *interpersonal* harm, which is significantly different from the conclusion that they moralise on grounds other than harm. In addition, if I am right in my (more contentious) claim that they primed their participants to understand “harm” to only entail “physical hurting” in this context, they only managed to establish that some of their participants moralised on grounds other than interpersonal physical hurting.

Over and above metaphysical presuppositions as to what or who can be harmed, the researchers may also have had a narrower understanding of the concept of “harm” than some of the respondents. Whereas the researchers did not consider “offence” to constitute a harm—as is evidenced from their description of the vignettes as “harmless but offensive”—some participants may very well have done so. This possibility is strengthened when one considers the participants’ reactions to the “disgusting” scenarios.[[12]](#footnote-12) Both SES groups found the disgusting stories offensive (the mean percentage across groups was 73%) (Haidt, et al. 1993, p. 618). Perhaps more accurately, the majority of the respondents were “bothered” by the stories, as “offence” was measured with the “Bother” probe, which reads as follows: “Imagine that you actually saw someone performing [the act]. Would it bother you, or would you not care?” This opens up the possibility that both groups were, in fact, moralising on the basis of harm considerations, if some participants considered offence to be a harm. Some respondents may have moralised the actions depicted because they themselves were offended by them, especially given that the Bother probe asks them to imagine witnessing these acts. Thus, even though the researchers stipulated that there were no witnesses to the acts *in the scenarios*, participants were invited to take on the role of witness, and, as the Bother probe shows, many of them were bothered by this.[[13]](#footnote-13) To be bothered by an action is to be impacted upon psychologically in a negative manner, which could plausibly be considered to be an instance of psychological harm. If “being bothered" is conceived of as being psychologically harmful, the possibility arises that those subjects who moralised on the basis of being bothered by the actions depicted were still doing so in terms of harm considerations—they themselves were psychologically harmed. Yet, this would not be reflected in the results, as the researchers built the assumption that offence does not constitute a harm into the research design by assuming that the offensive actions depicted are nevertheless harmless.[[14]](#footnote-14)

Besides considering offence to be a harm and having themselves been offended by the actions depicted, some respondents could very well have morally condemned the vignettes on the basis that they thought that some kind of harm *had* taken place in the scenarios themselves, despite the researchers deeming the acts to be objectively harmless. The researchers claim to have tested for this possibility with the "Harm probe". No large differences were found in what they call the overall “perception of harm” between the low and high SES groups, as measured through “yes” answers to the question: “Is *anyone hurt* by what [the actor] did? Who? How?” (p. 618) [*emphasis mine*]. In contrast, there was a discrepancy in the moralising responses between the high and low SES, with the low SES groups being more prone to morally condemning the acts depicted (p.619-620). Given these discrepancies, the researchers conclude that the groups made use of different moral foundations to underpin their judgements.[[15]](#footnote-15) My contention, however, is that the Harm probe does not effectively establish that participants did *not* perceive any harm to have taken place in the scenarios presented to them. This is due to the wording of the probe, as 1) the meaning of “harm” in this context is ambiguous, and 2) its wording precludes certain metaphysical possibilities relating to what or who can be harmed. Both of these aspects undermine the researchers’ claims that they successfully established that the participants perceived the scenarios to be harmless and then still went on to moralise them.

With regard to 1), note that the probe asks whether anyone is *hurt* by the actions depicted. “Hurt” and “harm” are used synonymously here, but the synonymity of these two terms is not at all obvious. “Harm" is often taken to refer to a range of detrimental infringements that someone may incur (which includes, but is not limited to, physical hurting).[[16]](#footnote-16) "Hurt", on the other hand, can plausibly be interpreted more narrowly as mainly referring to the physical damage or pain that someone may incur, and perhaps certain kinds of emotional pain.[[17]](#footnote-17) This interpretation is even more salient in the context of the experiment, given that in the prototypical moral violation presented, the girl is described as *hurting* the boy she pushes off the swing, who is physically injured. The probe questions also serve to suggest that in the context of this experiment a distinction is being made between being “hurt” and being “bothered”, given that “being bothered” is tested for in a separate question, thus strengthening the perception that “harm” here pertains to having a detrimental physical impact rather than a psychological or more abstract one.[[18]](#footnote-18) As a result, the research subjects may have understood “hurt” here to entail physical damage rather than psychological harm, thus answering “No” to the probe question whilst still conceiving of the actions depicted as (psychologically or more abstractly) harmful in nature. Given that the research subjects were not provided with any definitions, there is no way of knowing what they took these key terms in the experiment to mean and whether or not they understood them synonymously, as they were meant to. Hence, the researchers cannot claim to have definitively shown that harm was *not* the basis on which all of the moralised judgements were made.

In addition, with regard to 2), the Harm probe is phrased so as to preclude certain background metaphysical assumptions as to what or who can be harmed and how. Note that the probe asks whether *anyone is* harmed in the scenarios, thus only allowing for the researchers’ own background metaphysical assumptions regarding the kinds of entities that can incur harm.[[19]](#footnote-19) Arguably, the wording “anyone is” points to an *individual* being hurt (or harmed) by the actions, thus reflecting a view of harm as something that occurs interpersonally, i.e. between individuals. And, as we have seen, the scenarios were designed specifically to exclude particular, identifiable individuals who could potentially suffer harm. Yet, the possibility remains that respondents who answered “no” to the probe could still have perceived of other kinds of entities as being harmed in these scenarios.

As explained above, low SES respondents generally moralised the actions depicted in the “disrespect” and “disgust” scenarios, leading researchers to conclude that they were making use of moral foundations other than harm considerations. However, while it may very well be that concerns relating to group loyalty, authority, and purity (i.e. “community” and “divinity” to use the study’s terminology) had a stronger effect on these respondents than on the high SES ones, this does not necessarily mean that these groups moralised on the basis of other moral foundations.[[20]](#footnote-20) The possibility exists that perceptions of harm ultimately underpinned all of the moral judgements made in “Affect” but that participants subscribed to different background assumptions regarding who or what can be harmed (putting aside the question of whether or not they had similar conceptions of “harm”). The case can be made that the low SES respondents did indeed focus less on considerations of autonomy and included community and/or religious considerations in their moralising judgements, although *only in as far as these pertained to harm*. In other words, it is possible that these respondents could have viewed the community as a whole and/or the divinity or divinities recognised by their religions as entities that can in themselves be harmed.[[21]](#footnote-21) With regard to the disgusting scenarios, for example, they could have perceived of “god” and/or “the community” as entities liable to offence or anger. As such, it is the communities or gods themselves that are the harmed victims that lead to the moralisation here.[[22]](#footnote-22) Alternatively, it may be that these respondents did not conceive of the community or divinities as being harmed in themselves, but that, in flouting the rules set out by these entities, the order and harmony that these entities enable would be undermined, thereby harming the individual community members that rely on this order.[[23]](#footnote-23) At base, the concern is with the harm that the immoral actions can cause, not with the fact that they break rules, or undermine authority, or violate purity norms per se.

In short, it is entirely possible that the different socio-economic groups differed with regard to their metaphysical presuppositions relating to harm. There are a whole host of metaphysical assumptions that underlie harm judgements that are not taken into account in the research design and the interpretation of the data. Ironically, given their critique of the supposed western bias underlying harm-based moral theories, the researchers themselves uncritically universalise (purported) western metaphysical presuppositions about the vehicles through which harm can take place (individual, corporeal agents) and the types of entities liable to harm (individual, corporeal patients). While it is possible that all of the research subjects did, in fact, utilise these metaphysical assumptions, we have no way of establishing whether or not this is the case, as background metaphysical assumptions are not tested for.

1. Implications

These ambiguities in Haidt et al.’s findings are especially important given that they repeatedly counter suggestions that all of their moral foundations may ultimately reduce to harm by citing “Affect” and emphasising that some respondents *by their own admission* perceived the moralised actions to be harmless. Thus, Gray et al. critique MFT in various publications by pointing out that Haidt et al. do not take into account the possibility that research participants make moral judgements based on *perceived* harm despite being presented with “objectively” harmless situations (Gray, Young et al., 2012; Gray, Schein et al., 2014; Schein & Gray, 2015; Schein & Gray, 2017).[[24]](#footnote-24) Gray et al. do not address “Affect” in any detail, however, and Haidt et al. repeatedly counter their criticism of MFT by citing this paper and arguing that here respondents *admit* to the scenarios being harmless, thus nullifying Gray et al.’s contention (Ditto, Liu et al., 2012; Koleva & Haidt, 2012; Haidt, 2015). My claim that Haidt et al. have *not* managed to definitively show that respondents did not conceive of the depicted actions as harmful in nature undermines this defence. I contend that, at best, the researchers show that respondents thought that no *interpersonal harm* had taken place in the scenarios themselves (i.e. no individuals had been depicted as being harmed). This leaves open the possibility that the actions could have been thought of as being harmful to groups, supernatural beings, or other such entities. More problematically, the researchers may only have shown that respondents did not believe that anyone described in the scenarios was physically hurt, depending on respondents’ understanding of the wording of the Harm probe. Finally, as discussed above, respondents could have considered themselves to have been harmed in being made to imagine scenarios which they found to be offensive, if they considered offence to be a harm.

In summary, it should be clear that a central weakness in “Affect”, which also affects MFT more broadly, is that it does not have a sufficiently clear conception of the meaning of “harm” that it works with.[[25]](#footnote-25) As I pointed out above, the researchers themselves do not use the term consistently. In addition, they do not take their own metaphysical presuppositions regarding the nature and scope of harm into considerations in the design of the experiment and interpretation of the data. As a result, certain conceptions regarding the nature and scope of harm are already precluded in the way that the experiment is set up. Furthermore, they do not adequately test for the conception of “harm” that the research subjects themselves employ, leaving open the possibility that the respondents do, in fact, moralise on the basis of harm considerations, rather than fundamentally different moral foundations As a result, “Affect” does not convincingly show that additional moral foundations exist among its research subjects. Hence, this study gives us no reason to accept the claim that multiple moral foundations are necessary to account for the actual moral judgements that people make, let alone the specific foundations subsequently identified by MFT.

6. Conclusion

In light of the flaws in “Affect” and the far-reaching influence of this paper on the work of Haidt et al., the empirical case for moral pluralism as set out by these researchers is ambiguous at best. As a result, theorists should be wary of making any empirical or normative conclusions based on these findings. In terms of future empirical research into possible bases for people’s moral judgements, flaws such as those identified here may be ameliorated by, among other things, paying sufficient attention to: 1) establishing clear definitions of key terminology being employed in studies and ensuring that such terminology is used consistently, 2) ensuring that research subjects share researchers’ understanding of key concepts, or at least allowing for the possibility that they may not, and 3) allowing for the possibility that researchers and research subjects may employ different background metaphysical assumptions, which may have a significant impact on the outcome of the study. This may require working more closely with experts who are familiar with the culture of particular groups being studied. Such steps should have a positive impact on the plausibility of both empirical and normative conclusions drawn on the basis of such research.

References

Antiel, R.M., Humeniuk, K.M., & Tilburt, J.C., (2014). Spanning our differences: moral psychology, physician beliefs, and the practice of medicine. *Philosophy, Ethics, and Humanities in Medicine* 9(1):17.

Dedek, J. F., (1980). Premarital Sex: The Theological Argument from Peter Lombard to Durand. *Theological Studies,* 41(4), pp. 643-667.

Demetriou, D., (2013). There’s Some Fetish in Your Ethics: A limited defense of purity reasoning in moral discourse. *Journal of Philosophical Research,* Volume 38, pp. 377-404.

Devlin, P., (1965). *The Enforcement of Morals.* Oxford: Oxford University Press.

Ditto, P.H.; Liu B.; & Wojcik S.P., (2012). Is Anything Sacred Anymore? *Psychological Inquiry,* 23(2), pp. 155-161.

Dworkin, R. M., (1966). Lord Devlin and the Enforcement of Morals. *Faculty Scholarship Series,* Volume Paper 3611.

Feinberg, J., (1984). *The Moral Limits of the Criminal Law: Harm to Others.* Oxford: Oxford University Press.

Feinberg, J., (1985). *Offense to Others: The Moral Limits of the Criminal Law, Volume 2.* New York: Oxford University Press.

Feldman, F., (2010). *What is this thing called happiness?* Oxford: Oxford University Press.

Fisher, A. & Ramsay, H., (2000). Of Art and Blasphemy. *Ethical Theory and Moral Practice,* 3(2), pp. 137-167.

Fitness, J., (2001). Betrayal, Rejection, Revenge, and Forgiveness. In: M. Leary, ed. *Interpersonal rejection.* New York: Oxford University Press, pp. 73-103.

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., Nosek, B.A., Haidt, J. Iyer, I., Koleva, S., Ditto, P.H., (2011). Mapping the Moral Domain, *Journal of Personality and Social Psychology,* 101(2):366-385.

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(5):1029.

Goodwin, G. P. (2017). Is morality unified, and does this matter for moral reasoning? In: Jean-François Bonnefon, ed. *Moral Inferences.* Oxford: Psychology Press, pp. 10-64.

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(4):1600-1615.

Gray, K., Young, L., Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*. 23(2):101-124.

Gray, K., Waytz, A., Young, L., (2012). Reply. The Moral Dyad: A Fundamental Template Unifying. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory,* 23(2), pp. 206-215.

Haidt, J., (2015). *A straw man can never beat a shapeshifter: Response to Schein and Gray (2015).* [Online]
Available at: http://www.yourmorals.org/blog/2015/10/a-straw-man-can-never-beat-a-shapeshifter/

Haidt, J., (2013) Moral psychology for the twenty-first century. J*ournal of Moral Education.* 42(3):281-297.

Haidt, J., (2012). T*he righteous mind: Why good people are divided by politics and religion*. Random House LLC, New York.

Haidt, J. & Kesebir, S., (2010). *Morality*. In, S. Fiske, D. Gilbert, & D. Lindzey (Eds.) H*andbook of Social Psychology*, 5th Edition. Wiley, Hoboken, NJ. Pp. 797-832.

Haidt, J., Graham, J., (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research.* 20(1):98-116.

Haidt, J., Joseph, C., (2007). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In: Carruthers, P., Laurence, S., Stich, S. (eds) T*he innate mind, Volume 3: Foundations and the future.* Oxford University Press, New York, pp 367-392.

Haidt, J. & Joseph, C., (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus.* 133(4):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(4):613.

Henrich, J., Heine, S., & Norenzayan, A., (2010). The weirdest people in the world? *Behavioral and Brain Science,* 33(2-3), pp. 61-83.

Holtug, N., (2002). The Harm Principle. *Ethical Theory and Moral Practice,* 5(4), pp. 357-389.

Iyer, R., Koleva, S., Graham, J., Ditto, P., and Haidt, J. (2012) Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS one* 7(8):e42366.

Johnson, D., (1997). Formal Education vs. Religious Belief: Soliciting New Evidence with Multinomial Logit Modeling. *Journal for the Scientific Study of Religion,,* 36(2), pp. 231-246.

Jost, J. T., (2012). Left and right, right and wrong. *Science*. 337, 525–526.

Kagan, S., (1998). *Normative Ethics.* Boulder, CO: Westview Press.

Kim, K. R., Yun, S. & Kang, J.-S., (2012). Moral Intuitions and Political Orientation: Similarities and Differences between South Korea and the United States. *Psychological Reports.* 111(1), pp. 173-185.

Kleinig, J., (1978). Crime and the Concept of Harm. *American Philosophical Quarterly.* 15(1), pp. 27-36 .

Koleva, S., & Haidt, J. (2012) Let's Use Einstein's Safety Razor, Not Occam's Swiss Army Knife or Occam's Chainsaw. *Psychological Inquiry,* 23(2), pp. 175-178.

Leary, M. R. et al., (1998). The causes, phenomenology, and consequences of hurt feelings. *Journal of Personality and Social Psychology,* 74(5), pp. 1225-1237.

Levenbook, B. B., (1984). Harming Someone after His Death. *Ethics,* Volume 94, pp. 407-419.

MacDonald, G., (2009). Social pain and hurt feelings. In: P. J. Corr & G. Matthews, eds. *Cambridge Handbook of Personality Psychology.* Cambridge: Cambridge University Press, pp. 541-555.

May, J., (2017). Empirical Work in Moral Psycholog*y.* Routledge Encyclopedia of Philosophy. Taylor and Francis [Online]

Available at: https://www.rep.routledge.com/articles/thematic/moral-psychology-empirical-work-in/v-2.

McKay, R. &. Whitehouse, H., (2015). Religion and morality. *Psychological Bulletin.* 141(2), pp. 447-473.

McLaren, R. M. & Steuber, K. R., (2012). Emotions, communicative responses, and relational consequences of boundary turbulence. *Journal of Social and Personal Relationships.* 30(5), pp. 606 - 626.

Nado, J., Kelly, D. & Stich, S., (2009) “Moral Judgement”. In: *Routledge Companion to the Philosophy of Psychology*, ed. by John Symons & Paco Calvo. Routledge, Milton Park

Nagel, T. (2012). The taste for being moral. *New York Review of Books*, December 6 issue, pp. 40–42.

Nichols, S., (2004) S*entimental rules: On the natural foundations of moral judgment*. Oxford University Press, Oxford

Petersen, T., (2016). No Offense! On the Offense Principle and Some New Challenges. *Criminal Law and Philosophy,* 10(2), p. 355–365.

Prinz, J.J. (2006), The emotional basis of moral judgments. P*hilosophical Explorations.* 9(1):29-43.

Prinz, J.J. (2007a) *The emotional construction of morals*. Oxford University Press, Oxford.

Prinz, J.J., (2007b) Can moral obligations be empirically discovered? *Midwest Studies in Philosophy*. 31(1):271-291.

Rosas, A. (2013) Harm, Reciprocity and the Moral Domain. In: Karakostas V, Dieks D (eds) *EPSA11 Perspectives and Foundational Problems in Philosophy of Science.* Springer International Publishing, Switzerland

Schein, C. & Gray, K., (2015) The Unifying Moral Dyad: Liberals and Conservatives Share the Same Harm-based Template. *Personality and Social Psychology Bulletin,* 41(8), pp. 1147-1163.

Schein, C., & Gray, K., (2017) The Theory of Dyadic Morality: Reinventing Moral Judgement by Redefining Harm. *Personality and Social Psychology Review.* pp. 1-39.

Schwadel, P., 2015. Explaining Cross‐National Variation in the Effect of Higher Education on Religiosity. *Journal for the Scientific Study of Religion ,* 54(2), pp. 402-418.

Sherkat, D., 1988. Counterculture or Continuity? Competing Influences on Baby Boomers' Religious Orientations and Participation.. *Social Forces, ,* 76(3), p. 1087–1114.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park,, L. (1997). The “big three” of morality (autonomy, community, divinity) and the “big three” explanations of suffering. In: Brandt, A.M., Rozin, P. (eds) *Morality and health*. Routledge, New York, pp 119-169.

Shaver, P., Schwartz, J., Kirson, D. & O'Connor, C., (1987). Emotion knowledge: further exploration of a prototype approach.*Journal of Personality and Social Psychology,* 52(6), pp. 1061-86.

Silver, J. R. &. Silver, E., 2017. Why are conservatives more punitive than liberals? A moral foundations approach. *Law and Human Behavior,* 41(3), pp. 258-272.

Sinnott-Armstrong, W., (2008), Is Moral Phenomenology Unified? *Phenomenological Cognitive Science*. 7:85-97.

Sinnott-Armstrong, W. & Wheatley, T., (2014). Are moral judgments unified? *Philosophical Psychology.* 27(4):451-474.

 Sinnott-Armstrong, W. & Wheatley, T., (2012). The disunity of morality and why it matters to philosophy. *The Monist*. 95(3):355-377.

Stich, S., Fessler, D.M.T., & Kelly, D., (2009) On the Morality of Harm: A response to Sousa, Holbrook and Piazza. *Cognition*. 113:93-97.

Shweder, R. A., Mahapatra, M. & Miller, J. G., (1990). Culture and Moral Development. In: J. Kagan & S. Lamb, eds. *The Emergence of Morality in Young Children.* Chicago and London: University of Chicago Press, pp. 1-82.

Shweder, R. A. & Miller, J. G., (1985). The Social Construction of the Person: How Is it Possible?. In: K. Gergen & K. Davis, eds. *The Social Construction of the Person.* New York: Springer, pp. 41-69.

Thomson, J. J., (1986). Feinberg on Harm, Offense, and the Criminal Law: A Review Essay. *Philosophy & Public Affairs,* 15(4), pp. 381-395.

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

Turiel, E., Killen, M. & Helwig, C. C., (1990). Morality: It's Structure, Functions, and Vagaries. In: J. Kagan & S. Lamb, eds. *The Emergence of Morality in Young Children.* Chicago and London: University of Chicago Press, pp. 155-244.

Vangelisti, A. L., Young, S. L., Carpenter-Theune, K. E. & Alexander, A. L., (2005). Why Does It Hurt?: The Perceived Causes of Hurt Feelings. *Communication Research,* 32(4), pp. 443 - 477 .

Wainryb, C., Brehl, B. A. &. Matwin, S., (2005). Being Hurt and Hurting Others: Children's Narrative Accounts and Moral Judgments of Their Own Interpersonal Conflicts. *Monographs of the Society for Research in Child Development,* 70(3), pp. 1-114.

Weinberg, J. M., Nichols, S. & Stich, S., (2001). Normativity and Epistemic Intuitions. *Philosophical Topics,* 9(1 & 2), pp. 429-460.

Weston, W., 2014. A Deeper Durkheimian Society for Haidt’s Righteous Minds. *Society,* 51(6), p. 686–691.

1. I would like to thank the anonymous reviewers of this manuscript for their thoughtful and valued comments and recommendations. [↑](#footnote-ref-1)
2. For a useful overview see May (2017). [↑](#footnote-ref-2)
3. It should be noted that Haidt et al.’s work is not the only account of innate moral pluralism in the moral psychology literature, although it is one of the most influential ones. My critique here is aimed at this account specifically and should not be understood as calling all empirical work on moral pluralism into question. Nevertheless, I do go on to suggest that some of the problems that I identify here occur in other moral psychologists’ work as well. (See footnote 25.) [↑](#footnote-ref-3)
4. These postulated moral categories are partly derived from Shweder’s CAD theory (Shweder, Much et al. 1997) where Shweder and his colleagues propose that there are three distinct ethical systems that cultures use to approach and resolve moral issues: the ethics of community, autonomy, and divinity.   [↑](#footnote-ref-4)
5. In fairness to Haidt et al., they are, at times, careful to note that their claims are descriptive, not prescriptive. Nevertheless, they do sometimes make explicit normative claims. This is especially prevalent in Haidt (2012), as is pointed out in reviews by Jost (2012) and Nagel (2012). A few examples of such normative claims include the following: In Haidt and Graham (2007), with regard to “conservative” moral values, they state: “Traditions and institutions which have been vested with authority over the ages should be given the benefit of the doubt; they should not be torn down and rebuilt each time one group has a complaint against them” (112). Elsewhere, Haidt (2012) states, “But anyone who tells you that all societies […] should be using one particular moral matrix, resting on one particular configuration of moral foundations, is a fundamentalist of one sort or another” (316). And even while claiming to make descriptive points about human morality, value-laden language sometimes creeps in:

This functionalist approach allows psychology to move from moral *parochialism* (i.e., the belief that there is one universal moral domain that happens to include the values most prized by the secular academics who defined the domain) to moral pluralism (i.e., the belief that there are multiple incompatible but *morally defensible* ways of organizing a society… (Haidt & Kesebir 2010: 800). [*Emphasis mine*] [↑](#footnote-ref-5)
6. The impact of this study is highlighted by Goodwin’s (2017) claim that it is a key reason that Turiel’s (1983) influential contention that moral judgements are characterised by harm considerations has been called into question in moral psychology. (Also see footnote 8.) [↑](#footnote-ref-6)
7. See for example, Sinnott-Armstrong (2008); Nado, Kelly & Stich (2009); Stich, Fessler and Kelly (2009); Sinnott-Armstrong and Wheatley, 2012; Sinnott-Armstrong and Wheatley, 2014. [↑](#footnote-ref-7)
8. Haidt and Graham (2007) cite the findings in this paper as one of the principle reasons for developing their broadened moral theory, viz. Moral Foundations Theory. Significantly, Graham, Nosek, Haidt et al. (2011) also cite this paper as one of the bases for developing the Moral Foundations Questionnaire (MFQ), which in turn is taken to provide empirical support for MFT. I contend that the problems I highlight in “Affect” go on to contaminate MFQ and subsequent research based on it.

 [↑](#footnote-ref-8)
9. According to the MFT, there are (at least) five innate foundations that underpin human morality. These are the Care/harm foundation, the Fairness/cheating foundation, the Loyalty/betrayal foundation, the Authority/subversion foundation and the Sanctity/degradation foundation (Graham, Haidt et al. 2013). MFT is said to explain the “moral matrices” underlying human morality, by presenting the apparently differing moral permutations between different cultures in terms of differing configurations of these foundations. [↑](#footnote-ref-9)
10. In their various critiques, Gray et al. suggest ways in which the data that MFT is based on may be flawed. They do not elaborate on this criticism in great detail, however, nor do they address “Affect” or its design in detail. My critique here is entirely independent of theirs (see footnotes 13 and 24). [↑](#footnote-ref-10)
11. Westernisation and relative affluence are conflated here. [↑](#footnote-ref-11)
12. My discussion on “offence” here most plausibly applies to the “disgust” scenarios, which are also the scenarios where the greatest divergence in moralisation between the two groups occurred. My claim is that many of those who morally condemned the disgusting acts did so on the basis of perceiving them to be offensive and thus harmful. It is less likely that those who morally condemned the authority and purity violations did so on the basis of offence. In these instances, it is my contention that the acts depicted were conceived of as being harmful in ways not envisioned by the researchers (more on this below). [↑](#footnote-ref-12)
13. Gray, Waytz, et al. (2012) raise a related point with reference to their work on the phenomenon of dyadic completion (see footnote 24). My suggestion differs from theirs, however, in that I do not assume that we have a moral cognitive template that causes us to perceive victims where there may in fact be none. My suggestion is that the respondents were *caused* to conceive of themselves as witnesses *by the research design* which led them to feel offended. This possibility does not rely on the veracity of the dyadic model. [↑](#footnote-ref-13)
14. It is not immediately obvious on what basis the researchers assume that “offence” does not constitute a harm. They may have been influenced by the liberal legal tradition, where offence is generally not considered to be a harm (see Feinberg, 1985) (although this position is contested (see Thomson, 1986 and Petersen, 2016). (This raises the possibility that the higher SES groups who did not moralise the offensive acts did so precisely because they were “westernized” and thus influenced by the liberal legal tradition).Yet, even in this tradition, other adverse psychological impacts, particularly ones that affect our “welfare interests”, are considered to be harms (see Feinberg, 1984, p. 37-38; 46).

As mentioned above, it may be that the researchers assumed that the acts are harmless in the context of the vignettes seeing that there were no witnesses in the vignettes themselves. Hence, even if they did consider offence to be a harm, they still assumed that the vignettes were harmless since there is no one depicted who could be offended. Nevertheless, the design of the experiment did not guarantee that the participants would disregard their own feelings of offence in their moralising judgements. Essentially, this expectation required quite a sophisticated bit of moral reasoning on the part of participants, where their habitual moral responses would have needed to have been overridden on the basis of unusual factual circumstances where an offensive act is both “private” (in the context of the vignette) and under scrutiny (by the research subjects themselves). They would have had to realise that even if the acts depicted are usually harmful, they are not harmful under these particular circumstances. It is not at all obvious that moral intuitions elicited under such unusual circumstances are generalisable to all instances of moral reasoning. [↑](#footnote-ref-14)
15. More accurately, they hold that harm probably underlies all cultures’ morality, but that some cultures moralise on the basis of other considerations as well, here glossed as divinity or authority considerations. They need to make this concession, as the majority of respondents moralised the “prototypically moral” Swings story where an unambiguous case of interpersonal harm takes place (Haidt et al. 1993, p. 619-620).    [↑](#footnote-ref-15)
16. There does not seem to be much in the literature by way of empirical research into the folk conception of the concept of “harm”. In terms of the academic literature more genrally, the definition of “harm” is most comprehensively discussed in legal theory. In this context, Kleinig (1978), for example, points out that the use of the concept in English traditionally extends beyond physical hurting and is taken to include mental anguish, losses (of which one can remain unaware, such as damage to one’s reputation), and, in the context of legal theory, the violation of legally protected interests, such as when one’s property is trespassed upon. Similarly, Feinberg (1984) points out that in ordinary usage “harm” is employed in contexts where there is a “setting back of interests”, where these interests are components of someone’s well-being. These include welfare interests, a subset of which are freedom from physical and emotional pain. In addition, in legal theory, he argues, “harm” has a distinctively normative sense where it refers to wronging someone and violating her rights. In a similar vein, philosophers tend to define “harm” in terms of someone’s well-being being negatively impacted upon (e.g. Kagan, 1998; Holtug, 2002; and Feldman, 2010. In none of these analyses is “harm” taken to be synonymous with “hurt”. [↑](#footnote-ref-16)
17. An anonymous reviewer points out that we do speak of someone’s feelings being “hurt” and that the term thus has a wider meaning than that which pertains to physical infringement only. While this is true, I contend that the psychological impact associated with “hurt” in such usage only extends to that of causing someone psychological “injury”, usually by saddening him or by mortifying him and negatively thus impacting on his sense of self. Again, there does not seem to be research into the folk conception of “hurt”. Moreover, there is very little discussion of the meaning of the concept in the academic literature. Nevertheless, my construal of the meaning of the term is borne out by empirical work such as that of Wainryb, et al. (2005) where children tend to use the concept to describe physical injury or feelings of sadness and emotional pain, usually as a function of social rejection. The term is employed in a similar fashion elsewhere in the psychology literature (e.g. Shaver, et al., 1987; Leary, et al., 1998; Fitness, 2001; Vangelisti, et al. 2005; MacDonald, 2009; McLaren & Steuber, 2012). Being caused to experience emotional pain and feelings of sadness is still only a subset of the kinds of ways in which one can be negatively impacted upon psychologically speaking. Arguably, one can be psychologically affected in ways that do not entail hurt feelings in this sense, e.g. being traumatised or being caused to feel anxiety, distress, terror, outrage, misery, and grief. As Feinberg (1984, p. 46) puts it, there are several “unhappy” mental states over and above hurt feelings, and these may still constitute harm. Thus, “hurt” has connotations of (physical or emotional) injury and pain, which is a subset of the harms one may suffer (see footnote 16). [↑](#footnote-ref-17)
18. Support is given to my contention in that Goodwin (2017) also suggests that “hurt” may in this context be taken to refer to violent, physical harms only, although he does not elaborate on his reasons for this claim. [↑](#footnote-ref-18)
19. I thank an anonymous reviewer for pointing this out. [↑](#footnote-ref-19)
20. In this vein, Nichols (2004) and Rosas (2013) argue that the in-group/loyalty, authority/hierarchy and purity/ sanctity domains postulated in MFT are subordinate to the harm and fairness domains. I want to go further and suggest that it is entirely possible that *only* harm is constitutive of the moral domain, and that all of Haidt’s foundations reduce to a conception of harm. This possibility is certainly not precluded by the results of “Affect”, given the limitations in its design.   [↑](#footnote-ref-20)
21. A similar point is made by (Turiel, et al., 1990) in response to Shweder et al.’s (1990) claim that their cross-cultural research points to divergent grounds for moralisation in the United States and India. Shweder’s (Shweder & Miller, 1985; Shweder, et al., 1990) own research shows that Indian subjects consider unobservable entities such as souls, the deceased, the gods, and the natural order to be liable to harm and offence, which in turn results in harm to those who harm and offend them. One of the ways in which harm to these entities takes place is through the violation of “natural law” through breaking food and sexual taboos. Similar conceptions of harm exist in the context of Christianity, where God, the natural order, and one’s soul can be harmed through one’s actions, irrespective of there being (earthly) witnesses to the acts (see, for example, Dedek, 1980 and Fisher & Ramsay, 2000). Given that the respondents in Affect were from Brazil and the United States of America, some of them may presumably have adhered to a Christian worldview. This raises the possibility that some of their moralising of the given vignettes was due to considerations similar to those of Shweder’s research subjects, i.e. the harm that the actions caused the unobserved entities that form part of their metaphysical worldview. Moreover, extra-individual entities thought to be liable to harm need not be restricted to otherworldy or religious entities. In the context of western legal theory, for example, provision is made for the possibility of “harms to society” and “the social fabric”, even through actions that occur in private (e.g. Devlin, 1965 and Dworkin, 1966). Hence, some of the research subjects in Haidt et al. may very well thought that such extra-personal harm had occurred in the vignettes that they moralised.

The question as to *why* the different socioeconomic groups may have adhered to differing metaphysical presuppositions relating to harm is an empirical one, and I can only speculate here as to possible reasons for such a discrepancy. A possible reason could be the effect that education levels have on religious belief. Some research suggests that higher education levels are correlated with lower levels of religious belief (Sherkat, 1988; Johnson, 1997; Schwadel, 2015). Since the (adult) high SES-groups in Affect were primarily students, it may be that these groups were less religious than their low-SES counterparts and thus held more secular metaphysical worldviews where God and the community as such are less likely to feature as entities liable to harm. [↑](#footnote-ref-21)
22. This construal is borne out by the results from the control scenarios presented to participants (Haidt et al. 1993, p. 619). With regard to the violations of convention (not wearing a uniform to school and eating with one’s hands), the low SES respondents judged the uniform violation more harshly than the violation of etiquette (eating with one’s hands). These findings correlate with my postulation that the community and/or an authority figure (here, a school) is thought to be harmed in having its prescriptions violated. In the case of the breaching of etiquette, it is less clear that harm to the community or an authority figure occurs, as etiquette is not prescribed by a clearly identifiable authority and breaches of etiquette do not obviously undermine an authority figure. [↑](#footnote-ref-22)
23. See footnote 21. [↑](#footnote-ref-23)
24. Gray et al. base their criticism on their own model of what moral judgements entail. According to this model, our cognitive moral template consists of a dyad, where an intentional moral agent causes suffering to a moral patient (Gray, Waytz, et al. 2012; Gray, Young, et al. 2012; Gray, Schein et al., 2014; Schein & Gray, 2015; Schein & Gray, 2017). This model predicts that in cases where apparently harmless violations take place, respondents who moralise the actions depicted will automatically “complete” the dyad by perceiving a victim, even where there is none. The actions are moralised because they are taken to generally cause harm (Schein & Gray, 2017). In this vein, Schein and Gray (2017) suggest that the MFT “foundations” are, instead, values that tend to be moralised, based on their perceived link to harm. My argument does not presuppose the accuracy of Gray et al.’s dyadic model in that I claim that the wording of the probes and the way in which respondents were asked to think about the vignettes undermine the researchers’ findings, not the possibility that respondents intuitively perceived victims to be involved due to their cognitive constitution. [↑](#footnote-ref-24)
25. This failing is not limited to MFT. For example, Gray et al.’s own conception of “harm” is also ambiguous (e.g. Gray, Young, et al., 2012; Gray, Waytz, et al. 2012; Gray, Schein, et al. 2014). They contend that “perceived suffering” (which they equate with “harm”) is core to all moral judgements, but they are not clear on what such suffering entails. So, while they allow for broader metaphysical possibilities as to what or who can be harmed than MFT, they rely on a broad, equally ambiguous conception of harm, which is taken to include, *inter alia*, “bodily injury, emotional damage, or even spiritual destruction” (Gray, Young et al. 2012: 105). In recent work, they address this issue by defining “harm” as involving “… two perceived and causally connected minds, *an intentional agent causing damage to a vulnerable patient*” (Schein & Gray, 2017, pp. 1-2). However, here the term becomes shorthand for referring to a specific cognitive process (i.e. the making of moral judgements in terms of a dyadic template) rather than denoting specific kinds of states in entities that the term pertains to. The real work in this definition is being done by the term “damage”, which in itself requires definition. In order to support the thesis that morality is harm-based, these theorists need to indicate how the kinds of things that people moralise can plausibly be considered to be “damages”, otherwise, they run the risk of “Procrusteanism”, to borrow a criticism from Graham, Haidt et al. (2013). [↑](#footnote-ref-25)