natureresearch

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Situational factors shape moral judgments in the trolley dilemma in Eastern, Southern, and Western countries in a culturally diverse sample
Corresponding author name(s): Bence Bago

Reviewer Comments & Decisions:

Stage 1, pre-study:

Decision Letter, initial version:

18th July 2019

Dear Dr Bago,

Thank you once again for submitting your registered report, entitled "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgments</p>" to Nature Human Behaviour and for your patience during the peer review process.

Your manuscript has now been evaluated by 3 reviewers, whose comments are included at the end of this letter. We also solicited signed comments by the authors of the original research your work proposes to replicate, which are attached. Please note that signed comments do not function as reviews and do not form a basis for our editorial decision. The authors of signed comments are asked to comment solely on methodological aspects of your work, as well as on the extent to which their work is accurately represented in your manuscript. Decisions on whether to reject a manuscript or invite a revision are solely based on the feedback of our independent reviewers.

You will see that, although the reviewers find your protocol to be of interest, they also raise some important concerns. We are interested in the possibility of proceeding further with your submission in Nature Human Behaviour, but would like to consider your response to these concerns in the form of a revised manuscript before we make a decision on in principle acceptance and Stage 2 submission.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team to both identify key priorities that should be addressed in revision and overrule referee requests that are deemed beyond the scope of the current study. In your case, we would like to emphasize the following:

1. Please expand on the rationale for replicating this study provide justified and clear hypotheses about the cross-cultural similarities/differences you expect to find. Please note that we overrule any ambiguity in R3's comment "These questions are left for resolving after the phenomena have been established. It can be done that way round. But for the phenomena to be theoretically relevant, it needs to be done sometime. I think now would be a good time, so the theory can be tested with the advantage of a registered report. True, establishing phenomena is part of science, and the finding being replicated is a central one. So I leave this to the authors." and ask that you provide clear hypotheses and interpretations of the expected (and alternative) results.

2. Please address the concerns raised about the sample size estimation. We ask that, rather than basing your effect size estimates on past studies, including the ones listed by Greene, you need to use the smallest theoretically informative effect size as a basis for your a priori power estimation and/or sampling strategy (https://media.nature.com/original/nature-cms/uploads/ckeditor/attachments/4825/RegisteredReportsGuidelines_NatureHumanBehaviour.pdf). Please do this following the advice given by R3, including the analysis strategy detailed in the last paragraph.

3. Please include additional scenarios in the study design, as suggested by R1 and also mentioned by the author of the original study to ensure that your results are not dependent on the specific stimuli used but rather are likely to generalize to moral dilemmas generally.

4. We agree with the author of the original study that familiarity with the trolley problem poses a significant potential influence. If you do decide to use familiarity as a co-variate, please provide evidence that the results from these analyses would not likely differ from those from a dataset large enough that you can split familiar participants and unfamiliar participants and analyze these groups separately. The important concern here is that we ensure that any results you report are not likely to be driven by familiarity and rather represent true opinions.

In sum, we invite you to revise your Stage 1 Registered Report taking into account reviewer and editor comments. Please highlight all changes in the manuscript text file.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

* In your cover letter, please include the following information:
--An anticipated timeline for completing the study if your Stage 1 submission is accepted in principle.

--A statement of the number of labs who have received IRB approval for the study and updated sample size (per lab) for data collection.
--A statement confirming that you agree to share your raw data, any digital study materials, computer code (if relevant), and laboratory log for all eventually published results.
--A statement confirming that, following Stage 1 in principle acceptance, you agree to register your approved protocol on the Open Science Framework (https://osf.io/) or other recognised repository, either publicly or under private embargo, until submission of the Stage 2 manuscript.
--A statement confirming that if you later withdraw your paper, you agree to the Journal publishing a short summary of the pre-registered study under a section Withdrawn Registrations.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Human Behaviour is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,

Mary Elizabeth

Mary Elizabeth Sutherland
Editor
Nature Human Behaviour

Reviewer expertise:

Reviewer #1: moral psychology

Reviewer #2: moral psychology

Reviewer #3: Bayesian statistics, registered reports

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
In this registered report the authors propose to replicate the findings of Greene et al 2008, studies 1a and 2a. At the time they were published, the original studies added to an existing literature implicating both the doctrine of double effect and some version of "personal force" as factors that influence the moral judgment of welfare tradeoff dilemmas. The novel contribution of the Greene et al 2008 study was to show a statistical interaction between these influences, in the form of a superadditive effect on moral condmenation. Moreover, the authors intended to conduct this replication attempt across dozens of different sample populations around the world, assessing the degree to which it is culturally moderated or, instead, universal. This is an ambitious and interesting proposal and there are several strong points. At the same time I also have some substantial reservations.

First, the proposal does little to motivate why these effects—the DDE, personal force, and their interaction—are of particular interest. Nor is it obvious why they should be. Although the 2008 study is well cited, not very much subsequent work has attempted to provide a rich theoretical account of why these particular factors matter, or why they interact (although there are at least few publications that do). Nor does the study outline any compelling hypotheses about why there might or might not be cross cultural variability, or how these possibilities bear on current theory. In order for any outcome of this study to be appropriate for Nature Human Behavior I think the authors have to do more to motivate why this is an interesting or important effect to understand, including in cross cultural context, from the standpoint of improving our current theories of moral psychology.

Second, the proposal seeks to characterize cross-cultural variability in two general moral principles, but it attempts to do this by conducting experiments on a very limited set of vignettes — eight in total, all of them variants of the trolley problem. If the authors find cross cultural variability it will not be clear whether this is due to underlying variability in the principles and their application, or instead variability in something very specific about the way those principles are applied to one particular kind of case. Even the original 2008 paper noted this shortcoming of Studies 1a and 2a, and in a subsequent study the authors reanalyzed a larger dataset comprising several different kinds of welfare tradeoff dilemmas. This subsequent analysis confirmed several findings from Studies 1a and 2a but showed other ways in which the more limited analysis of the trolley problem was idiosyncratic. In the years since the publication of that 2008 paper it has become more-or-less mandatory in the field to show that patterns of moral judgment generalize across diverse cases, and I worry that it would be something of a step backwards for this study to rely on contrasts among trolley problems alone.

Reviewer #2:
Remarks to the Author:
This paper represents a major project to examine whether a foundational effect in the moral dilemma literature replicates in many samples around the world. The authors will examine whether a key finding from Greene and colleagues will replicate in three cultural clusters. That finding remains foundational for the field, and like the authors I cannot think of a direct replication; moreover they are right to note that the few crosscultural studies suggest some systematic differences, though whether those differences pertain to the current question remain untested. In addition, this large diverse sample allows for testing moderation by a number of key factors such as religiosity and relationship to the 2D model. All in all, there is strong potential to learnsonething of value from these studies. Indeed, this paper would seem poised to become a landmark publication for the field, regardless of the results. Therefore, I am favorably disposed towards publication.

I have few serious concerns; mostly I suggest the authors refine the conceptual language they use to describe dilemma judgments.

The authors describe dilemmas as a clash between "deontological perspectives (i.e., emphasizing rules, individual rights and duties)" and utilitarian perspectives. This is largely right but this wording can be misleading, as dilemma judgments in themselves do hot directly measure philosophical thinking. I suggest rephrasing to emphasize the distinction between the judgment itself and the philosophy. For example, refusing to cause harm is consistent with deontological ethics. This wording avoids theimplication that such judgments are caused by committment to deontological principles, a claim wisely debunked (e.g., see Kahane et al., 2015, 2018; Conway et al., 2018). Likewuse, I suggest clarifying the phrase "use of deontological versus utilitarian considerations" as this implies that the ethics or philosophy form the psychology directly, when instead phrasing like "judgments consistent with deontological versus utilitarian considerations" avoids this implication. Likewise with phrasing like dilemmas involving "decisions between two main conflicting moral principles." Although past work has described it this way Common this phrasing is misleading as it makes the untested assumptions that the principles themselves are the root cause of the psychological experience behind the judgements. However Common even the paper they cite by Greene and colleagues describe the psychology involved in terms of a dual process model describing psychological processes such as effective reactions to home they are not themselves the principles. Such processes ml judgements that happen to be consistent with those principles.

In a related problem common the authors describe those who arrive in judgements that are consistent with deontology as "deontological reasoners" but this description makes the untested assumption that judgements to reject causing harm reflect reasoning period although this terminology is coming it is incorrect. In fact, the due process model the authors site suggest that such judgements reflect affective or intuitive processes. although it is true that recent work suggests that some types of reasoning can lead to judgements consistent with deontological principles, considerable evidence still supports the role of affective and emotional processes. Therefore, it is safer conceptually to describe such people using neutral language rather than language that assumes the process, such as 'deontological decision-makers'.

Otherwise, I fibd the paper clear, proposing good methods, and citing appropriately. I will be eagerly awaiting the results of this project.

Reviewer #3:
Remarks to the Author:
This is a generally well written and formulated proposal, as one would expect for something to come out of the psychological accelerator, and from the lead team. I think it would be an excellent contribution to the journal.

Theory tested:
The experiment is an attempt to replicate a phenomenon (a finding) across peoples from different demographics. Beyond this, there is no theory. What theory of moral intuitions is put on the line by potential results of the study? What theory would predict differences according to demographics and in what direction? These questions are left for resolving after the phenomena have been established. It can be done that way round. But for the phenomena to be theoretically relevant, it needs to be done sometime. I think now would be a good time, so the theory can be tested with the advantage of a registered report. True, establishing phenomena is part of science, and the finding being replicated is a central one. So I leave this to the authors.

Hypothesis tests:
The authors suggest separate tests for the key phenomena for Western clusters and the others. That is fine. Bear in mind that evidence for H1 in the Western clusters and for H0 in the others would not legitimate concluding a difference between the clusters. Try explaining that pattern of results without assuming and hence claiming a difference between the clusters! Hence I recommend explicitly testing the difference between the clusters (using the same Bayesian t-tests).

"As we have no clear prior expectations on certain tests, we will use a non-informative prior for the alternative model: a two-tailed Cauchy distribution with scale r =0.707"

A Cauchy with any scale factor is not non-informative. It claims there is a 50% probability of getting an effect between 0 and the scale factor. The function of the prior is to represent the predictions of the theory; I am not sure what non-informative could mean in that context, or why it would be a virtue. (A non-informative prior in estimation is different; that means the posterior is is determined completely by the likelihood. In fact, that is the type of estimation prior I typically use.) So the key question is whether a Cohen's d of 0.7 is the right scale for the questions asked. Later on the authors say "The original study reported effect sizes for effects of interest to be a standardized mean difference of 0.46 (personal force) and 0.47 (means - side effect distinction) respectively. We used a standardized mean difference of 0.3 (2/3 of the original effect) in our simulations." So now we are told the effect size of interest: 0.3 Cohen's d units. So there is a conceptual discrepancy between what

effect is claimed to be roughly the right scale, and what effect is used the Bayes factor. It is easy to remove the discrepancy by using r = 0.3 for the Bayes factors. A further peculiarity can be likewise ironed out. The authors say "Here, the predictions of the H1 model are represented with a normal distribution centered on zero and with a SD of Cohen's d = 0.5." Why now jump to a Cohen's d of 0.5? Just use the same one as for the Cauchy so that the only difference is Cauchy versus normal. Further the authors say they will use an informed prior: "This model assigns the standardized effect size a t distribution centered on 0.35 with a scaling factor of 0.102 and three degrees of freedom. " There is nothing informed about this. This is one person's subjective prior for a completely different effect in social psychology. The authors go on to say they think it may well apply broadly. But they already get that sort of prior, and better, by using a Cauchy or Normal scaled by 0.3. These latter models are truly informed by the very effect they wish to test.
Now all this is admittedly academic. With N = 13,000, fiddling with the precise prior will make no difference. Their power calculation shows this for even a considerably smaller N. So really what I am saying is all a matter of presentation. But it is important see what is actually going on with the testing. So I suggest the authors use the motivated priors- Cauchy or normal with a scale factor based on Greene. Then calculate robustness by means of a suggestion of one of the lead authors, Aczel Balazs, a robustness region for the Cauchy and Normal. With N = 13,000 that region is likely to be broad, likely showing that any plausible scale factor would have led to the same conclusion.

* NOTE * signed comments are attached as a pdf:

Thanks very much for giving me the opportunity to comment on the plan for "Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgments", by Bago and colleagues. First, I wish to express how deeply honored I am that such a large and distinguished group of researchers has agreed to invest their precious time and resources in this project. I am very excited to see the results, and I am happy to offer whatever help I can as this project moves forward.

You have asked me to "comment on methodological aspects of this submission and the extent to which your work is accurately represented and interpreted". It is clear that the authors have taken a great deal of care in developing this research plan. The overall strategy makes sense, and the authors have anticipated many important issues. I believe that the plan, at the most general level, is sound.

I do, however, have some significant concerns about the plan, which I outline below. I also suggest ways of addressing these problems.

1. Familiarity with trolley dilemmas

To their credit, the authors have anticipated this problem. Nevertheless, I worry that this will be a much larger problem than the authors have anticipated.

When the original data were collected in the mid 2000s, few members of the general public had heard of the "Trolley Problem" and its now well-known variants, such as the standard *footbridge* and *switch* cases. Today, these dilemmas are not only widely known in academia, but are known to the general public through popular TV shows (such as *The Good Place* and *The Unbreakable Kimmy Schmidt*), popular books (e.g. *Would you Push the Fat Man?* by David Edmonds), and countless internet "memes". A *New York Times* article begins: "You probably know the dilemma of the trolley problem: …" (Orin Kerr, 1/27/16). This author's assumption is backed up by data. In my lab's most recent experiment using a version of the *footbridge dilemma*, we asked our MTurk participants, "Have you encountered this kind of dilemma before?" 60% said "Yes".

Familiarity with such cases does not necessarily preclude their use in research. It depends on the research question and design. In the aforementioned experiment, the *footbridge* dilemma was used in both the control and experimental condition, where the latter involved an additional cognitive manipulation. In this case we did not exclude participants reporting familiarity, and we obtained a positive effect, which we then replicated with several different dilemma types. (See Huang, Greene, and Bazerman: https://psyarxiv.com/fcqpt/) But, unfortunately, familiarity with trolley dilemmas is likely to be an acute problem for the present research.

Here, the problem is that the key pattern of judgment, whereby participants judge moral sacrifices to be worse when the action is more physically direct, is a pattern that participants tend to disavow once they are made aware of it (Cushman et al., 2006). And familiarity with the with the standard cases can produce this kind of awareness. Consider a participant in the *footbridge switch* condition. She's asked whether it's acceptable to save the five people by hitting a switch on the footbridge that opens a trap door, dropping the trolley-stopping victim onto the tracks. Suppose that this person knows of the original *footbridge* dilemma, in which one can push the man off the footbridge. And suppose that she, like most people, thinks it's clearly wrong to push the man off the footbridge. With the original *footbridge* case in mind, is she now going to say that killing this person becomes acceptable if the physical mechanism is changed from pushing with one's hands to using a switch-operated trap-door?

As this example illustrates, familiarity with trolley dilemmas undermines the study's between- subjects design. The participants may only see one dilemma on the screen, but if other relevant dilemmas come to mind, then they are essentially putting themselves in a different experimental condition. This is especially problematic here because we already know that people respond differently to these cases when they are considered together. More specifically, when people encounter less emotionally salient versions (such the original *switch* case) after encountering more emotionally salient versions (such as the original *footbridge* case), the behavioral gap between the two cases shrinks, with people giving more deontological responses to the less emotional versions (e.g. Schwitzgebel & Cushman, 2012; Barak-Corren et al., 2017). This gap is exactly what this study is trying to understand.

While the authors have anticipated the familiarity problem, I'm afraid that their plans to address it are not adequate. They have said that they will exclude philosophers and philosophy students from their samples. While that may have been adequate in the past, today one rightly expects readers of mainstream newspapers to be familiar with these dilemmas. This expectation applies even more to the psychology students who populate departmental study pools, some of which will be used for the present research. The authors have also said that they will include a question about familiarity at the end of the study and that they will use this measure as a covariate in an exploratory analysis of the effect of familiarity. I agree that asking about familiarity is important, but I don't think this can be handled with an exploratory analysis, as it's central to the integrity of the study.

Depending on how widely familiar these dilemmas are, and how much familiarity reduces the target effect, there may not be a good solution. But if the study is going to move forward, the authors should run their main analyses both including and excluding participants who report being familiar with the testing materials. The authors propose to ask: "Are you familiar with moral dilemmas of this kind, in which you can save more people by causing the death of one person? Please rate it in a scale from 1 (absolutely not familiar) to 5 (absolutely familiar)". This seems to me like a good way to ask the question. The analysis excluding familiar participants should exclude participants who answer 4 or 5. The analysis that includes familiar participants should include the familiarity variable, not as a simple covariate (aimed at finding a main effect), but rather as a moderator (i.e. interacting with the effects of interest based on dilemma type).

As the foregoing suggests, this could have a large impact on the unfiltered sample sizes needed, with a majority of participants being excluded from some sub-samples, and with exclusion rates differing across cultural groups. The authors are currently expecting to exclude only 10% of participants. For this reason (and others—See below), this estimate seems to me far too optimistic.

To be clear, I'm not saying that the authors shouldn't also analyze the complete data set, as they have planned. But I don't think that the authors should draw firm conclusions from results that depend on including participants who report being familiar with the testing materials.

Again, it's not truly a between-subjects design if the participants are thinking about dilemmas similar to those used in other conditions. If there is a discrepancy between the results with and without exclusion for familiarity, this needs to be acknowledged prominently in the reporting of the main results. If it turns out that familiarity has no effect (neither a main effect nor an interaction with dilemma type), then it would make sense to collapse across familiar and unfamiliar participants. This would restore a great deal of statistical power.

2. Attention and comprehension

Unfortunately, there is an additional need for more stringent exclusion criteria. In order to exclude participants who engage in "careless responding" the authors intend to use the following "yes"/"no" items: "I was born on February 30th", "I've travelled to the Moon three times", and, "I can read and write". The authors plan to exclude only those participants who give incorrect answers to *all three* questions. I don't see how this could be an adequate filter for "careless responding". A participant who answers all three questions *randomly* will have an 87.5% chance of passing.

To begin, the authors should simply exclude participants who respond incorrectly to *any one* of these items. (We don't want any participants who claim to have visited the moon three times!) But I don't think that's enough, as responding correctly to these items requires very little effort and attention, far less than the main dilemma task. The authors should, in addition, use a more demanding attention filter such that participants who are paying minimal attention will be unlikely to pass. More importantly, though, the study should include comprehension checks.

For example, the participants could be asked about the key manipulations concerning the physical mechanism of harm (e.g. in Study 1a, they could select the nature of the physical action: pushing with hands, pushing with pole, hitting switch on footbridge, hitting switch far from footbridge).

On Mturk, and depending on the study, my lab expects to exclude 30%-45% of participants for failed comprehension. Our combined total exclusion rate (using an attention filter, exclusion for duplicate Mturk IDs from previous studies, technical failures, and comprehension) can be as high as 50%. That may seem excessive, but this is the price one pays for having on-demand access to thousands of participants at modest pay rates.

It’s worth pausing to consider the combined effect of these two exclusion issues. If 60% of participants report being familiar with trolley dilemmas (i.e. 40% unfamiliar), and an uncorrelated 50% of participants need to be excluded for other reasons, then the total exclusion rate (for the analyses excluding familiar participants) will be 80%. Fortunately, it should be possible for this large research team to apply these appropriately strict criteria, at least for some cultural clusters. But the authors should know that their statistical power will be nothing close to what they are currently expecting.

3. Planned analyses for Experiments 1a-b

Overall, I think the plan for testing Hypotheses 1a-b makes sense. However, I think that it would be a big improvement to include an additional dilemma, the *standard switch* case. And I think that it would be better to include all four of the original Experiment 1a dilemmas in the first wave of data collection, so as to avoid violating random assignment (See below).

First, although the *standard switch* dilemma (a version of which appears in the introduction to the authors’ research plan) was regrettably not included in the original paper, this dilemma and the standard *footbridge* cases are the two best known and widely used trolley dilemmas.

Critically, these two dilemmas are the anchors for the broader “Trolley Problem”, i.e. the problem of explaining (or, for some philosophers, justifying) the different responses that people tend to give to these cases. Given the massive effort going into this project, it seems worthwhile to cover the whole Trolley Problem, from *footbridge* to *switch*, and not just the *footbridge* variations. That way, however the results turn out, the data will be able to say something about the overall shape of the Trolley Problem across cultures. I would be happy to supply a matching version of the *standard switch* dilemma, which I have from pilot testing of the original research.

In addition to broadening the interest of this work, having data for this dilemma could be very helpful for interpretation, especially if the personal force effect fails to replicate. In that case, there are two possibilities: The first is that the empirical Trolley Problem remains in full force, such that the gap between the *footbridge* and *switch* cases is as large as ever. If that’s the case, but the *personal force* effect is not found, having data for the *standard switch* case will provide clues about which factors do matter. Another possibility, in the absence of the *personal force* effect, is that the *personal force* effect is absent because the Trolley Problem itself is shrinking, i.e. the gap between *switch* and *footbridge* has diminished or (unlikely) disappeared. In that case, the shrinking of the Trolley Problem could be due to familiarity, as explained above, but it could also be eroded from the other side, with responses to the *footbridge* case becoming more utilitarian. There is some evidence that people’s judgments are becoming more utilitarian over time, specifically because people are less emotionally sensitive to dilemmas such as the *footbridge* case (Hannikainen, Machery, and Cushman, 2018).

Second, the authors plan to begin by running only the *footbridge pole* and *footbridge switch* dilemmas. I understand the reason for wanting to focus specifically on the "personal force" contrast, as defined by these two cases. However, as explained below, I think it would be better to run all four of the original dilemmas (plus the *standard switch* dilemma) at the same time to avoid violating random assignment.

4. Planned analyses for Experiments 2a-b

With respect to hypotheses 2a-b, the analysis plan is somewhat inconsistent, both internally and with the original paper. The authors describe Hypothesis 2 as: "The means – side effect distinction should have an effect on moral acceptability ratings, when personal force is applied." This would seem to refer to the simple contrast between the *obstacle push* and *obstacle collide* cases. But later the analysis protocol states:

> "To test Hypothesis 2a and 2b, we will compare the obstacle push (means and personal force) with the three other dilemmas from Study 2, namely loop, obstacle collide, and loop weight. Importantly, the obstacle push dilemma is predicted to receive the lowest level of moral acceptability rating, and it should differ from all the other experimental conditions combined."

Neither formulation is exactly right. The first formulation is accurate, but incomplete, while the second does not completely accurately reflect the conclusions drawn from Studies 2a-b of the original paper.

As the authors correctly state, there should be an effect of intention in the context of personal force, meaning that there should be a difference between *obstacle push* and *obstacle collide*. But the original paper does not claim that there should necessarily be a significant difference between the *obstacle push* case, on the one hand, and the *loop* and *loop weight* cases, on the other. Rather, the claim is that the difference between the *obstacle push* and *obstacle collide* cases should be larger than the difference between *loop* and *loop weight* cases—a "difference of differences" constituting an interaction between the personal force factor and the intention factor.

The analysis proposed in the second formulation above is consistent with how the data happened to turn out in the original Experiment 2a (shown in Figure 4). But not all features of this pattern were replicated in the original Experiment 2b, and this is reflected in the conclusions of the paper. To get it right, the authors need to break this analysis down into three parts.

1. compare *obstacle push* to *obstacle collide* (as suggested in the first formulation above)

2. compare *loop* to *loop weight*

3. compare the sizes of the two effects above (predicting 1 > 2).

To illustrate the difference between the original paper's conclusion and what the authors have proposed, consider the following set of outcomes:

loop: 4.5

loop weight: 4.5

obstacle push: 4.0

obstacle collide: 6.0

Suppose that *obstacle push* differs from *obstacle collide*, that this effect is significantly larger than the (non-existent) difference between *loop* and *loop weight*, and that *obstacle push* does not differ significantly from the mean of the other three dilemmas. This set of outcomes would be perfectly consistent with the conclusions of the original paper, and yet this would (incorrectly) be considered a failure to replicate, based on the current analysis plan.

I am not sure how to do the "difference of differences" analysis required by #3 within a Bayesian framework, but I assume that there's a way.

5. Assumptions in Bayesian analyses

A strength of the present research is its Bayesian approach, which attempts to quantify the relative strength of the evidence for the null and alternative hypotheses, rather than enforcing a binary "significant"/"not significant". In keeping with this Bayesian approach, If the main results are inconclusive, they should be reported as "inconclusive" and not as "failure to replicate" or "not significant", as the latter descriptions can easily be misinterpreted as reflecting positive evidence for the null. While I am enthusiastic about this analysis strategy, I have some questions/concerns about some of the specific assumptions made in the analysis plan.

Selection of priors: For the primary analysis, the authors intend to use a non-informative prior for the alternative hypothesis, i.e. the hypothesis that there is a positive effect. I understand that the effect size is 0 on the null model and that there is an "objective Bayes" argument for using a noninformative prior when using Bayesian analysis to estimate an effect size. But in an analysis that is aimed at contrasting a null hypothesis with an existing, *directional* hypothesis, why formulate the existing hypothesis as 'there's some effect, but it's as likely to be negative as positive'? Instead, the main analysis should use a prior distribution for the alternative hypothesis centered on a reasonable positive value.

As it happens, the authors are already planning to do this in one of their robustness checks. Here, they plan to model the alternative hypothesis with a prior distribution centered at +0.35, based on general effect sizes in the field. My strong recommendation is that this approach be adopted for the main analysis, as it makes no sense to ignore the direction of the effects reported in the original paper.

This recommendation is especially important because the "Bayesian t-test" approach that the authors have planned is likely to be especially sensitive to the choice of priors. More specifically, a limitation of this approach is that it does not distinguish between evidence for the null due to the absence of an effect from evidence for the null due to a poor specification of the prior for the effect. An alternative approach, recommended by Kruschke (2013), is to avoid distilling results down to a single number (such as a Bayes factor) and instead describe the posterior distribution in more detail. I am not expecting the authors the change their analysis plan in light of these recommendations. But if they are going to stick with the "Bayesian t-test" approach, they need to pay special attention to robustness of their result to the selection of different priors.

Putting the robustness analysis to use: The authors write "The robustness analysis will only be added to the supplementary material, and no inference will be drawn from it." If the selection of priors substantially influences the conclusions, then this should change the way that the main results are reported. Again, I recommend using a distribution centered at +0.35 for the alternative hypothesis in the main analysis. And then the robustness checks should shift this positive value up and down, for example, with one alternative prior centered on the original effect size (moving thing up) and one centered on half the original effect size (moving things down). The authors can also vary the shapes of these distributions as they suggested, using a two-tailed Cauchy distribution, a normal distribution, and *t* distribution.

If all, or nearly all, of these variations point to the same conclusion, that's good news. If main conclusions vary substantially with the selection of priors (among those suggested above), then this should be presented prominently in the reporting of the main results. If the robustness checks waver between support for the alternative hypothesis and "inconclusive", then the main results should be described as something like "tentative support" for the alternative hypothesis. Likewise, if they waver between support for the null and "inconclusive", they should be presented as providing "tentative support" for the null. If the selection of priors (among those suggested above) shifts things from support for the alternative to support for the null and vice versa, then something has gone wrong, and the results should be reported as simply "inconclusive", pending a diagnosis and mutually agreeable remedy for the problem.

6. Minimum effect sizes for sample size estimation

The authors intend to use an effect size estimate of .3, which is described as 2/3 of the original effect size (.46 for study 1a and .47 for study 2a). First, it's not clear to me how the authors obtained these effect size numbers (.46 and .47). In the paper (pg. 367) we report an effect size of *d*=.40 for the personal force effect (*remote footbridge* vs. *footbridge pole*). I'm not sure which exact contrast is supposed to have a value of .47, but perhaps this is *obstacle collide* vs. *obstacle push*? Or *obstacle collide* vs. the other three? Were these values calculated independently from the reported means, standard errors, and Ns?

Second, the authors propose to set the minimum effect size of interest at 2/3 of the original, reflecting the tendency for effects that replicate to be smaller than those originally reported. While I agree that this adjustment is a good idea, I think it doesn't go far enough. For example, if there were a significant effect at half of the original effect size, that would certainly be of interest, and it would most naturally be regarded as a successful direct replication—with caveats concerning the reduced effect size.

To see how much this matters for sample size estimation, I asked a member of my lab who is fluent in both R and Bayesian statistics to download the code for the sample size estimation (very helpfully provided by the authors) and try the same analysis using an effect size of half of the original (.23, assuming .46 is correct for the original). If I understand the results correctly, this dramatically increases the suggested sample size. With an effect half as large as the original, the 95% chance of making all 6 inferences correctly drops to 54% with their original sample size of 700 per condition per cluster. Fortunately, because of the Bayesian analysis strategy, this doesn't have a big effect on the likelihood of making a *false* inference. But it does make it much more likely that the results will be inconclusive.

7. Other issues

Prediction regarding intention in the absence of personal force: In discussing the conclusions of the original paper, the authors write, "Hence, [Greene et al.] concluded that people are only sensitive to the doctrine of double effect when they have to use their own physical force." This is not what we concluded. It is true that Experiment 2a found an effect of intention (i.e., the means/side-effect factor, consistent with the "doctrine of double effect") only in the presence of personal force. But Experiment 2b showed a main effect of intention, which included a relatively weak, but significant, effect of intention even in the absence of personal force.

Random assignment and prioritization of dilemmas: To maximize power for the key analyses, the authors intend to initially test the *footbridge switch*, *footbridge pole*, *loop*, *obstacle push*, *loop weight*, and *obstacle collide* dilemmas. And then they will test the *standard footbridge* and remote *footbridge dilemmas*, if resources allow. While I appreciate the reason for doing this, it has the unfortunate effect of violating random assignment to condition, as participants who get the latter two dilemmas will be drawn from different populations from those receiving the original four.

Instead of planning to violate random assignment for Experiment 1, I think it makes more sense to first run all of the Experiment 1 dilemmas (*standard footbridge*, *remote footbridge*, *footbridge pole*, and *remote footbridge*). And, as noted above, I strongly recommend adding the *standard switch* dilemma so that, however the results turn out, they can be interpreted in the context of the boarder Trolley Problem. I recommend subsequently running the Experiment 2 dilemmas (*loop*, *loop weight*, *obstacle push*, and *obstacle collide*) if and when an adequately powered sample is available. I think it makes sense to give priority to Experiment 1 because that experiment documents the effect of personal force. Showing that the personal force effect is weaker in the absence of intention is, I think, of secondary interest.

Oxford Utilitarianism Scale: The authors intend to use the Oxford Utilitarianism Scale to "help us address potential critiques that suggested caveats for using the trolley problem as a direct measure of utilitarian responding". I do not recommend using the Oxford Utilitarianism Scale for this purpose. The best response to the critique in question (made by Kahane and colleagues) is to point out that it is conceptually flawed, based on a misunderstanding of how I and others have used the term "utilitarian judgment". As explained at great length elsewhere (Conway et al., 2018), Kahane and colleagues have mistakenly insisted that a "utilitarian judgment" can only mean "a judgment made by a utilitarian" rather than "a judgment required by utilitarianism". (This is like saying that an Italian meal can only be made by an Italian.) The Oxford Utilitarianism Scale, then, is Kahane et al.'s solution to a non-existent problem—the problem of identifying the true utilitarians in the general population, even though no one (myself included) believes that general population includes more than a tiny fraction of true utilitarians.

If the authors would like to assess individual differences in philosophical orientation— and in a way that is relevant to ordinary people's trolley judgments—they are much better off using Conway's Process Dissociation measure (Conway & Gawronski, 2013; Conway et al., 2018), which has been remarkably effective at relating dilemma judgments to other individual difference variables. (For an excellent recent example, see Patil et al.: https://psyarxiv.com/q86vx/).

I wish to be clear that this is just a suggestion and that the decision to use a "utilitarianness" measure, and which one to use, is orthogonal to the present replication effort.

Follow-up questions from original experiments: The original experiments include three follow- up questions. The first asks about confusion. I think this is worth keeping, and using for exclusion as in the original paper. The second and third questions were designed to address a deflationary hypothesis according to which people's answers simply reflect confounding beliefs about the realism of the scenarios. The original paper showed very little effect of this kind, and as far as I know there's not been much discussion of this idea since. Of course, there's no harm in including these additional measures. But if the protocol is overcrowded, I think it would be fine to drop them.

Dilemma labels: The authors didn't comment on this in their plan, but they should not include the dilemma labels (e.g. "remote footbridge") in the actual testing materials.

Name of agent: In the original experiments, the agent is always named "Joe". Should this be adjusted for cultural context?

My apologies for the length of these comments. But given the great effort and expense that will be applied to this project, I prefer to err on the side of being thorough. Once again, I am grateful and honored that the authors have chosen to embark on this project. As I've explained, I do have some serious concerns about the current research plan, but I also believe that these concerns can be addressed. Most notably, I believe that the widespread familiarity of these dilemmas, combined with the need for strong quality control, will dramatically reduce the effective sample size. Nevertheless, I see no reason why this large research group cannot obtain a truly adequate sample size. Likewise, I believe that the other issues I've raised concerning the analysis plan and data collection can be addressed as well.

Thanks very much for giving me the opportunity to comment on this work. I am excited to see the results.

Yours truly,

Joshua D. Greene

Professor, Department of Psychology
Center for Brain Sciences

Harvard University

References

Barak-Corren, N., Tsay, C. J., Cushman, F., & Bazerman, M. H. (2017). If you're going to do wrong, at least do it right: considering two moral dilemmas at the same time promotes moral consistency. *Management Science*, *64*(4), 1528-1540.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, *104*(2), 216.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition*, *179*, 241-265.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082-1089.

Hannikainen, I. R., Machery, E., & Cushman, F. A. (2018). Is utilitarian sacrifice becoming more morally permissible?. *Cognition*, *170*, 95-101.

Huang, K., Greene, J. D., & Bazerman, M. (2019). Veil-of-Ignorance Reasoning Favors the Greater Good. https://psyarxiv.com/fcqpt/

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., ... & Cushman, F. (2018). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*(2), 135-153.

Author Rebuttal to Initial comments

We would like to thank the editorial team, the anonymous reviewers, and Prof. Greene for providing us with thoughtful comments. We agreed and complied with all comments, and have implemented the following changes:

1. **We provide clear hypotheses on how culture is expected to moderate the effect of intention and personal force on moral dilemma judgements.**
2. **We recalibrated the priors in our Bayesian analysis based on theoretical considerations and supplemented it with a robustness analysis.**
3. **Now, participants will be presented with two dilemmas, instead of one.**
4. **We changed the analysis strategy of Hypothesis 2. Instead of the originally proposed *t* test, now we propose to use linear regression models and directly test the interaction of intention and personal force, following the original experiment.**
5. **In accordance with Prof. Greene's suggestion, we added 1 more filtering question to the survey to measure the participants' attentiveness.**
6. **From all analyses, we will exclude the data of participants who had been familiar with the trolley problem prior to participating in the experiment.**
7. **In accordance with the suggestion by Reviewer 2, we updated the use of conceptual language we used to explain moral dilemma judgements.**

Response to reviewers

Reviewer #1:

Remarks to the Author:

In this registered report the authors propose to replicate the findings of Greene et al 2008, studies 1a and 2a. At the time they were published, the original studies added to an existing literature implicating both the doctrine of double effect and some version of "personal force" as factors that influence the moral judgment of welfare tradeoff dilemmas. The novel contribution of the Greene et al 2008 study was to show a statistical interaction between these influences, in the form of a superadditive effect on moral condemnation. Moreover, the authors intended to conduct this replication attempt across dozens of different sample populations around the world, assessing the degree to which it is culturally moderated or, instead, universal. This is an ambitious and interesting proposal and there are several strong points. At the same time I also have some substantial reservations.

First, the proposal does little to motivate why these effects—the DDE, personal force, and their interaction—are of particular interest. Nor is it obvious why they should be. Although the 2008 study is well cited, not very much subsequent work has attempted to provide a rich theoretical account of why these particular factors matter, or why they interact (although there are at least few publications that do). Nor does the study outline any compelling hypotheses about why there might or might not be cross cultural variability, or how these possibilities bear on current theory. In order for any outcome of this study to be appropriate for Nature Human Behavior I think the authors have to do more to motivate why this is an interesting or important effect to understand, including in cross cultural context, from the standpoint of improving our current theories of moral psychology.

Response:

Thank you for the suggestion. We have now added three hypotheses specifying similarities and differences among cultures. For details, please see our answer to the editor above. We further specified that the DDE and intention are both hypothesized to affect moral judgment through eliciting emotional responses (guilt, shame). Cultural differences in these social emotions predict differences among collectivist and individualist countries.

Second, the proposal seeks to characterize cross-cultural variability in two general moral principles, but it attempts to do this by conducting experiments on a very limited set of vignettes — eight in total, all of them variants of the trolley problem. If the authors find cross cultural variability it will not be clear whether this is due to underlying variability in the principles and their application, or instead variability in something very specific about the way those principles are applied to one particular kind of case. Even the original 2008 paper noted this shortcoming of Studies 1a and 2a, and in a subsequent study the authors reanalyzed a larger dataset comprising several different kinds of welfare tradeoff dilemmas. This subsequent analysis confirmed several findings from Studies 1a and 2a but showed other ways in which the more limited analysis of the trolley problem was idiosyncratic. In the years since the publication of that 2008 paper it has become more-or-less mandatory in the field to show that patterns of moral judgment generalize across diverse cases, and I worry that it would be something of a step backwards for this study to rely on contrasts among trolley problems alone.

Response:

We agreed to use other problems as well from Experiment 1b and 2b from Greene et al. Participants will be presented with these problems after they are presented with the trolley dilemmas. Please see our response to the editor and our manuscript for detailed description.

Reviewer #2:

Remarks to the Author:

This paper represents a major project to examine whether a foundational effect in the moral dilemma literature replicates in many samples around the world. The authors will examine whether a key finding from Greene and colleagues will replicate in three cultural clusters. That finding remains foundational for the field, and like the authors I cannot think of a direct replication; moreover they are right to note that the few crosscultural studies suggest some systematic differences, though whether those differences pertain to the current question remain untested. In addition, this large diverse sample allows for testing moderation by a number of key factors such as religiosity and relationship to the 2D model. All in all, there is strong potential to learnsonething of value from these studies. Indeed, this paper would seem poised to become a landmark publication for the field, regardless of the results. Therefore, I am favorably disposed towards publication.

I have few serious concerns; mostly I suggest the authors refine the conceptual language they use to describe dilemma judgments.

The authors describe dilemmas as a clash between "deontological perspectives (i.e., emphasizing rules, individual rights and duties)" and utilitarian perspectives. This is largely right but this wording can be misleading, as dilemma judgments in themselves do hot directly measure philosophical thinking. I suggest rephrasing to emphasize the distinction between the judgment itself and the philosophy. For example, refusing to cause harm is consistent with deontological ethics. This wording avoids the implication that such judgments are caused by commitment to deontological principles, a claim wisely debunked (e.g., see Kahane et al., 2015, 2018; Conway et al., 2018). Likewise, I suggest clarifying the phrase "use of deontological versus utilitarian considerations" as this implies that the ethics or philosophy form the psychology directly, when instead phrasing like "judgments consistent with deontological versus utilitarian considerations" avoids this implication. Likewise with phrasing like dilemmas involving "decisions between two main conflicting moral principles." Although past work has described it this way Common this phrasing is misleading as it makes the untested assumptions that the principles themselves are the root cause of the psychological experience behind the judgements. However Common even the paper they cite by Greene and colleagues describe the psychology involved in terms of a dual process model describing psychological processes such as effective reactions to home they are not themselves the principles. Such processes ml judgements that happen to be consistent with those principles.

In a related problem common the authors describe those who arrive in judgements that are consistent with deontology as "deontological reasoners" but this description makes the untested assumption that judgements to reject causing harm reflect reasoning period although this terminology is coming it is incorrect. In fact, the due process model the authors site suggest that such judgements reflect affective or intuitive processes. although it is true that recent work suggests that some types of reasoning can lead to judgements consistent with deontological principles, considerable evidence still supports the role of affective and emotional processes. Therefore, it is safer conceptually to describe such people using neutral language rather than language that assumes the process, such as 'deontological decision-makers'.

Otherwise, I find the paper clear, proposing good methods, and citing appropriately. I will be eagerly awaiting the results of this project.

Response:

We would like to thank the reviewer for pointing this out. We changed the “reasoner” labels to “decision-makers”, and used the term “judgments consistent with deontological considerations” instead of “deontological judgments” everywhere in the text. We also clarified that there is a distinction between the investigated judgment and the underlying philosophy.

Reviewer #3:

Remarks to the Author:

This is a generally well written and formulated proposal, as one would expect for something to come out of the psychological accelerator, and from the lead team. I think it would be an excellent contribution to the journal.

Theory tested:

The experiment is an attempt to replicate a phenomenon (a finding) across peoples from different demographics. Beyond this, there is no theory. What theory of moral intuitions is put on the line by potential results of the study? What theory would predict differences according to demographics and in what direction? These questions are left for resolving after the phenomena have been established. It can be done that way round. But for the phenomena to be theoretically relevant, it needs to be done sometime. I think now would be a good time, so the theory can be tested with the advantage of a registered report. True, establishing phenomena is part of science, and the finding being replicated is a central one. So I leave this to the authors.

Response:

Thank you for the suggestion. Now, we present three hypotheses on the cultural universality of these effects and also on how individualism/collectivism is expected to affect moral judgements. Please see our response to the editor and our manuscript for further details.

Hypothesis tests:

The authors suggest separate tests for the key phenomena for Western clusters and the others. That is fine. Bear in mind that evidence for H1 in the Western clusters and for H0 in the others would not legitimate concluding a difference between the clusters. Try explaining that pattern of results without assuming and hence claiming a difference between the clusters! Hence I recommend explicitly testing the difference between the clusters (using the same Bayesian t-tests).

Response:

We will only use cultural clusters to test the universality of the effects of personal force and intention. Individualism/collectivism will be added as a continuous predictor to our model.

"As we have no clear prior expectations on certain tests, we will use a non-informative prior for the alternative model: a two-tailed Cauchy distribution with scale r =0.707"

A Cauchy with any scale factor is not non-informative. It claims there is a 50% probability of getting an effect between 0 and the scale factor. The function of the prior is to represent the predictions of the theory; I am not sure what non-informative could mean in that context, or why it would be a virtue. (A non-informative prior in estimation is different; that means the posterior is is determined completely by the likelihood. In fact, that is the type of estimation prior I typically use.) So the key question is whether a Cohen's d of 0.7 is the right scale for the questions asked. Later on the authors say "The original study reported effect sizes for effects of interest to be a standardized mean difference of 0.46 (personal force) and 0.47 (means - side effect distinction) respectively. We used a standardized mean difference of 0.3 (2/3 of the original effect) in our simulations." So now we are told the effect size of interest: 0.3 Cohen's d units. So there is a conceptual discrepancy between what effect is claimed to be roughly the right scale, and what effect is used the Bayes factor. It is easy to remove the discrepancy by using r = 0.3 for the Bayes factors. A further peculiarity can be likewise ironed out. The authors say "Here, the predictions of the H1 model are represented with a normal distribution centered on zero and with a SD of Cohen’s d = 0.5." Why now jump to a Cohen's d of 0.5? Just use the same one as for the Cauchy so that the only difference is Cauchy versus normal. Further the authors say they will use an informed prior: "This model assigns the standardized effect size a t distribution centered on 0.35 with a scaling factor of 0.102 and three degrees of freedom. " There is nothing informed about this. This is one person's subjective prior for a completely different effect in social psychology. The authors go on to say they think it may well apply broadly. But they already get that sort of prior, and better, by using a

Cauchy or Normal scaled by 0.3. These latter models are truly informed by the very effect they wish to test.

Now all this is admittedly academic. With N = 13,000, fiddling with the precise prior will make no difference. Their power calculation shows this for even a considerably smaller N. So really what I am saying is all a matter of presentation. But it is important see what is actually going on with the testing. So I suggest the authors use the motivated priors- Cauchy or normal with a scale factor based on Greene. Then calculate robustness by means of a suggestion of one of the lead authors, Aczel Balazs, a robustness region for the Cauchy and Normal. With N = 13,000 that region is likely to be broad, likely showing that any plausible scale factor would have led to the same conclusion.

Response:

Thank you for these suggestions. We have revised our use of priors and sample size estimation strategy. For details, please see response to the editor above.

Decision Letter, first revision:

4th December 2019

Dear Bence,

Thank you once again for your manuscript, entitled "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgements</p>," and for your patience during the peer review process.

Your manuscript has now been evaluated by the Reviewers #2 and #3 (from the first round of reviews), and by Joshua Greene who has provided signed comments. Unfortunately, Reviewer #1 was not able to comment again. The reviews and Dr Greene's comments included at the end of this letter. Based on all the feedback we've received, we ask you to conduct a final round of revisions before we make a final decision on Stage 1 in principle acceptance for your Registered Report.

The most salient remaining concern is the ultimate sample size you can achieve in your study, and whether this will provide you with sufficient power to detect the targeted effect size, i.e., ⅔ of the original effect sizes. Please respond to the comments made by Dr Greene regarding the estimate of the effect sizes you used in this calculation. The rate of drop-out because of familiarity across different cultures is currently unknown, although we expect that exclusions due to familiarity will be very limited beyond the US sample. Similarly, predictions about comprehension failures are attached with some uncertainty. To address the referees' concerns and make use of the comprehensive comments provided by Dr Greene, we ask you to follow Reviewer #2's suggestion to preregister two analyses, one including participants who indicate familiarity, and one without these participants. Your sampling strategy must be to collect data until you have reached the BF criterion in the sample that excludes participants familiar with the paradigm (and of course those excluded due to comprehension failure).

In sum, we invite you to revise your Stage 1 Registered Report taking into account reviewer and editor comments. Please highlight all changes in the manuscript text file.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

* Include a “Response to reviewers” document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the reviewers along with the revised manuscript.

* In your cover letter, please include the following information:
--An anticipated timeline for completing the study if your Stage 1 submission is accepted in principle.
--A statement confirming that you agree to share your raw data, any digital study materials, computer code (if relevant), and laboratory log for all eventually published results.
--A statement confirming that, following Stage 1 in principle acceptance, you agree to register your approved protocol on the Open Science Framework (https://osf.io/) or other recognised repository, either publicly or under private embargo, until submission of the Stage 2 manuscript.
--A statement confirming that if you later withdraw your paper, you agree to the Journal publishing a short summary of the pre-registered study under a section Withdrawn Registrations.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know. We will be happy to consider your revision so long as nothing similar has been accepted for publication at Nature Human Behaviour or published elsewhere.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Human Behaviour is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best wishes,
Marike

Reviewers' Comments:

Reviewer #2:
Remarks to the Author:
Overall it seems the authors have been responsive to the comments and suggestions raised by reviewers. I am especially pleased to see the edits to the wording as I think this version is much clearer regarding terminology, and to see the addition of the Greene dilemmas following the ones in the main document--although this is more work, it is likely to increase the scientific value of this project. Likewise with the clarifications of the hypotheses, which helps this paper have better impact. However, I do think the authors could lean a little further in this direction. The paper still reads like the cultural hypotheses were tacked on at the end rather than an organic reason to do the study.

For one thing they appear late in the paper and are worded in such passive voicing that they stand out from the rest of the text,

For example "Our third hypothesis is based on the belief that even though personal force and intention are cultural universals, certain cultural differences are expected to moderate their effect" could get reworded as something like "Even though we anticipate that personal force and intention will emerge universally across cultures, we nonetheless expect cultural differences to moderate these effects. In particular..." Note that such wording is easier for readers to process, making your claims seem intuitive. As is, this portion remains cognitively demanding for readers and hence less impactful than might be hoped. Small edits will help.

The abstract was another place that jumped out at me as worth improving. For example, the abstract focuses primarily on the question of generalizability, with a large portion of words dedicated thus:

"As their study was conducted with US samples, our knowledge is limited concerning the universality of this effect, in general, and the impact of culture on the situational and psychological factors of moral judgements, in particular. Here, we empirically test the universality of intent and personal force on moral dilemma judgements by testing the replicability of the experiments "

Meanwhile, the cultural hypothesis portion of the abstract gets only "We test three hypotheses regarding the effect of culture on moral dilemma judgements."

Consider that the majority of readers will not advance beyond the abstract--they are left without any sense of what cultural differences are getting tested and what direction the effects might be. The authors would gain a lot by slightly trimming the words focused on limited knowledge about generalizability etc., and instead clarifying the actual theoretical meat, such as:

"We hypothesize that intent and personal force will universally increase the unacceptability of outcome-maximizing harm in dilemmas, this effect will be stronger in collectivistic than individualistic cultures due to increased guilt and shame "

Such wording is only a rough draft and no doubt the authors can improve on it, but this sort of language is I think what the other reviewer was hoping for when they suggested clarifying the theoretical reasons for doing the study (I sentiment I obviously agree with). Such wording helps organically help readers envision interesting potential outcomes at a glance, and makes your hypotheses seem clear and integral to the project.

SImilarly, it strikes me that the title is a bit generic and does not fully communicate all the key ideas in play--nor it does not convey a sense of the direction and nature of the hypotheses. Right now it reads:

"Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements"

but think of all the future people searching by keywords and such who might skip this paper based on the title--for example anyone interested in individualism-collectivism. A few tweaks might buy you more information in the title (you could update the in the eventual version with data)

"Personal force and intention universally increase the unacceptability of sacrificial harm, especially in collectivistic cultures: A Global Study"

Again this is only a suggested draft title, but the point is I think the authors can improve clarify, precision, and tone especially in the title, abstract, and the new hypothesis section, to make sure this paper ends up finding and communicating effectively to the audience it deserves.

I only have one other outstanding issue: I am somewhat concerned by the response to Dr. Greene's concerns over familiarity as a decision to exclude all people who score 4 or 5. There are two problems with this: over the past decades trolley style dilemmas have grown so famous (even featuring in mainstream American TV such as The Good Place) that this decision risks excluding a very high percent of the sample (i.e., possibly upwards of 60% in North America), hence ruining power. Second, its highly likely that this decision will systematically throw out more people from some groups than others (i.e., those most familiar with such cultural touchstones), skewing this otherwise fairly representative global sample. It seems like there might be wiser alternatives to this extreme decision. For example, the authors should pledge a priori to specific analyses. such as:

-the core analysis with and without controlling for familiarity, and documenting the difference in effect size
-the core analysis with and without excluding people who score 4 or 5 on familiarity, and documenting the difference in effect size

In other words, rather than cutting off a bug chunk of the full sample, arm the reader with all the information to reach their own conclusions over this factor. Moreover, such analyses would add to the value of this paper. People debate how much familiarity impacts dilemma judgments; with a sample this large you may be able to help put a tight confidence interval around that effect.

On the other hand, I am perfectly content with excluding all people who fail the dilemma attention check question described in Supplementary Material B.

Reviewer #3:
Remarks to the Author:
I am happy with the revisions the authors have made. They will use models of H1 informed by the original study. (As the function of a model of H1 is to represent the predictions of the theory - given evidence depend son how well data are predicted - the authors are right that in this context it is the predicted effect size, and its range, that is important. When power is used in frequentist statistics, by contrast, it is, as the editor says, the minimal meaningful effect size that must be used, because in this case the aim of power is to guard against missing out on any effect of interest. As the authors say they will make inferences with respect to the Bayesian analyses, they in fact do not have to justify a minimal effect. As an aside, this is one reason I find Bayes factors so useful!) The authors also now consider robustness explicitly. The authors have also added some theory to the introduction - which, as the edtior says, does make better science.

Signed comments:
See attached:

Dear Dr. Schiffer,

Thanks very much for giving me the opportunity to comment further on the research plan of Bago et al. I'm delighted to see the many improvements made to this revised plan, and I appreciate the authors' openness to feedback. And, once again, I am honored that so many researchers have agreed to devote their precious time and resources to this project. I am excited to see the results, assuming the study goes forward.

Unfortunately, I still have some major concerns that remain unresolved from my last set of comments related to statistical power and sample sizes. These involve a set of component concerns about participant exclusions and anticipated effect sizes. Indeed, my reading of the revised research plan has only deepened these concerns. After working through these issues, I have concluded that the authors should drop Study 2, as this seems to me the only way to ensure adequate statistical power. I explain my reasoning in detail below.

I was not expecting to reach this conclusion, and before I did, I drafted a set of detailed comments about the design of Study 2. I have included most of these comments here, but in light of my recommendation regarding statistical power, I hope that these comments will be unnecessary.

In the first section below, I lay out my concerns about statistical power, breaking them down by component concerns that bear on the sample size calculations. Most of these points re-visit points that were made in my original set of comments.

I then make a quick but important point about centering the prior distribution of the alternative hypothesis.

The next two sections, respectively, concern the new secondary dilemmas planned for Study 2 and the rationale and analyses planned for the primary Study 2 dilemmas.

I then present a set of more circumscribed points organized by page.

In addition to providing input at this stage, I request that I have the option to write a commentary, to be published alongside the main paper, if this project moves forward.

The big issue: statistical power and sample sizes

It may seem strange to raise concerns about statistical power for a project involving over 100 labs intending to collect data from over 17,000 participants. Indeed, when I first considered this research plan, I assumed that statistical power would be a non-issue. Nevertheless, I believe that it's a major problem, partly due to the ambition of the experimental design (6 conditions in 3 cultural clusters), partly due to the size of the original effects (especially Study 2), partly due to the low quality of data from online samples, partly due to the need to detect relatively small effects, and partly due to the particular challenges of testing with trolley dilemmas in 2019- 2020 vs. 2006-2008. As I will explain below, these problems are *multiplicative.*

According to the authors' power calculations, they need 700 participants per condition per cluster for Study 1 and 1,100 per condition per cluster for Study 2. The labs are committed to collecting a minimum of 17,035 participants. As I'll explain, it is difficult to reconcile these numbers once all of the relevant factors are accounted for.

<u>6 conditions x 3 cultural clusters requires a lot of power</u>

To begin with the most obvious point, the ambition of the project quickly multiplies the necessary sample size. Studies 1 and 2 will be run simultaneously. Study 1 has two conditions, and Study 2 has four conditions, for a total of six conditions. And then there are three cultural clusters, for a total of 18 conditions across the clusters. Dividing 17,035 by 18 yields 946 participants per cluster (on average), which is more than the 700 participants required for Study 1, but less than the 1,100 participants said to be required for Study 2.

I understand that the authors have factored this 3 x 6 design into their power analyses, and I am not suggesting that we substitute my back-of-the envelope calculation for their more complex simulation-based calculations. My point here is simply to dispel the immediate impression (which I certainly had) that recruiting 17,000 participants is obviously going to be enough for this project. The proposed project requires approximately *nine times* the sample size of a basic experiment comparing two conditions within a single cultural context.

If the 3 x 6 design were the only thing driving up the sample size, things might be fine. But this is just the beginning. The calculation above does not account for:

1. Excluding large numbers of participants due to for failed comprehension

2. Excluding large numbers of participants due to familiarity with the testing materials

3. Accurately incorporating the original effect sizes

4. Using sample sizes powered to detect "<u>the smallest theoretically informative effect size</u>" (as per the editor)

<u>Exclusion for failed comprehension</u>

In my original review, I emphasized the need to test of for comprehension of the moral dilemma, beyond a general test of attentiveness and non-serious responding. I pointed out that, in my experience with

Mturk studies, we typically need to exclude about 40% of participants for comprehension/attention related problems. (See supplementary materials for our recently published *PNAS* paper "Veil-of-Ignorance Reasoning Favors the Greater Good" for current exclusion rates.)

To their credit, the present authors are now planning to use a comprehension question and to exclude non-comprehending participants. Unfortunately, the authors make no mention of how these expected exclusions are factored into their sample size calculations. The authors cite a minimum total sample size of 17,035. If this is the number of participants whose data are expected to be analyzed (after exclusions for failed comprehension and familiarity), there is no mention of the much larger set of participants who will need to be tested. As indicated below, a corresponding pre-exclusion sample size could be over 70,000 participants.

Because there is no mention of testing such a large number of participants, my assumption is that the figure of 17,035 participants is based on what the participating labs have committed to running, meaning that this is the figure before participants have been excluded for failed comprehension and familiarity. In that case, the authors' power analyses are way off.

Let's start with exclusion for failed comprehension. If 40% fail the comprehension check, then the total post-exclusion sample size should be about 10,221, and that yields a sample size of 568 per cell per cluster—well below the 700 needed for Study 1 and the 1,100 said to be needed for Study 2.

And this is before we consider exclusion for familiarity.

<u>Possible exclusion for familiarity with canonical trolley dilemmas</u>

In my original review I wrote at length about the problem of familiarity with the canonical trolley dilemmas. I noted that, since the original studies were run, trolley dilemmas have become widely known to the general public, and even more widely known to participants on Mturk and to psychology students. The most telling example is a *New York Times* article that begins "You probably know the dilemma of the trolley problem: …" (Orin Kerr, 1/27/16) signaling an expectation that most *Times* readers are familiar with these dilemmas.

While familiarity with trolley dilemmas is not necessarily a problem for all research designs, I explained in my last set of comments why it is likely to be a problem in this case: Participants in the no-personal-force condition (using the *footbridge pole* dilemma) are likely to be reminded of the familiar *footbridge* dilemma (which does involve personal force), thus undermining the studies' between-subjects design.

To their credit, the authors are now planning to include a question about familiarity. And they plan to exclude participants who report being familiar. There are two related problems here.

First, as above with exclusions for comprehension, the authors make no mention of how exclusions for familiarity are factored into their sample size calculations. As I explained in my original review, these exclusions could be huge. In our most recent study using the *footbridge* dilemma, 60% of participants answered “Yes” to the question “Have you encountered this kind of dilemma before?”. This 60% figure may be an *underestimate*, at least for the Western cluster, but let’s consider its implications for sample size.

The possibility of having to exclude 60% of participants cuts the initial total set from 17,035 to 6,814, which leaves approximately 379 participants per cell—far less than the 700 needed for Study 1 and the 1,100 said to be needed for Study 2. But these figures ignore the *combined* effect of excluding for comprehension *and* for familiarity. If we start with 10,221 participants following exclusion for comprehension, and then exclude 60% for familiarity, we are left with a total sample of 4,088 and only 227 participants per cell per cluster—far, far below the 700 participants per cell needed for Study 1, never mind Study 2.

This brings me to second problem related to familiarity. if I understand the current research plan, the authors are simply planning to exclude participants who report being familiar with trolley dilemmas. This would be appropriately conservative if one could be confident that relatively few participants would be excluded. But that’s definitely not the case here. In my original comments I recommended analyzing the data both ways, with and without exclusion for familiarity. I understand the rationale for constraining the analysis plan, and thus reducing experimenter degrees of freedom. And I would be fine with designating the more exclusive analysis as the primary analysis provided that it is adequately powered. But whatever the sample size, it’s important to acknowledge that circumstances have changed dramatically since the original studies were run, making it unclear which analysis plan is most faithful to the original study: Is it one that assumes that nothing has changed, or one that attempts to adapt to changes? Just as positive results should be checked for robustness, negative results should as well. If the results replicate when all participants are included, that’s important to know, and it should be reported prominently.

In my initial set of comments, I emphasized the implications of excluding for both comprehension and familiarity. I wrote:

“It’s worth pausing to consider the combined effect of these two exclusion issues. If 60% of participants report being familiar with trolley dilemmas (i.e. 40% unfamiliar), and an uncorrelated 50% of participants need to be excluded for other reasons, then the total exclusion rate (for the analyses excluding familiar participants) will be 80%. [76% if one uses the 40% estimate for exclusion for comprehension and other problems] Fortunately, it should be possible for this large research team to apply these appropriately strict criteria, at least for some cultural clusters. But the authors should know that their statistical power will be nothing close to what they are currently expecting."

In retrospect, I regret expressing optimism about how the power analysis would work out, but I was not expecting the project to move forward based on my guess. I was expecting the authors to present a new power analysis based on the assumption that large numbers of participants (very likely a majority) would need to be excluded for failed comprehension and/or familiarity with the testing materials. As far as I can tell, the authors have not incorporated these considerations into their sample size calculations.

<u>Incorporation of the original effect sizes</u>

The authors' power calculations depend on effect size estimates based on the original studies. In my first set of comments I raised questions about how the authors identified the original effect sizes. I wrote:

"The authors intend to use an effect size estimate of .3, which is described as 2/3 of the original effect size (.46 for study 1a and .47 for study 2a). First, it's not clear to me how the authors obtained these effect size numbers (.46 and .47). In the [original] paper (pg. 367) we report an effect size of *d*=.40 for the personal force effect (*remote footbridge*[1] vs. *footbridge pole*). I'm not sure which exact contrast is supposed to have a value of .47, but perhaps this is *obstacle collide* vs. *obstacle push*? Or *obstacle collide* vs. the other three? Were these values calculated independently from the reported means, standard errors, and Ns?"

[1In my earlier comments I mislabeled this dilemma. The correct contrast is *footbridge switch* vs. *footbridge pol*e. But the effect size of *d*=.40 is correct for the effect of personal force.]

It's still not clear to me where the effect size of .46 for Study 1 comes from. Again, it seems to me that the right number is *d*=.4. Assuming an effect 2/3 of the original, that would be an anticipated effect of .27 rather than .3.

The bigger question here concerns the new figure of .55 for the interaction of personal force and intention in Study 2. I do not know where this number came from or what effect size measure this is using. In the original Study 2a, the effect size for the interaction is a partial eta squared of .02. By my calculation, this corresponds to a *d*=.285.

Thus, as far as I can tell, the authors have overestimated the original effect size for Study 1 (by a little) and may be substantially overestimating the effect size for the interaction in Study 2.

<u>Use of the "the smallest theoretically informative effect size"</u>

The authors make their sample size calculations using an effect size of 2/3 of the original. As I noted in my original comments, this is a step in the right direction. However, as the editor affirmed in her most recent response, the sample size estimate should be based on "the smallest theoretically informative effect size". I think it's reasonable to use 2/3 of the original as a prior for the alternative hypothesis in a Bayesian analysis (see below), but I don't think it's reasonable to say that anything less than 2/3 of the original effect size is uninteresting and equivalent to zero. This is certainly true for Study 1, where the original effect is d=.4. A statistically significant effect of d=.2 (half the original effect size) should not count as a "failed replication". Thus, in my original comments, I suggested that the sample size calculations use half of the original effect sizes. On this point, I presented some additional calculations using half the original effect size for Study 1, taking the .46 figure as given:

"To see how much this matters for sample size estimation, I asked a member of my lab who is fluent in both R and Bayesian statistics to download the code for the sample size estimation (very helpfully provided by the authors) and try the same analysis using an effect size of half of the original (.23, assuming .46 is correct for the original). If I understand the results correctly, this dramatically increases the suggested sample size. With an effect half as large as the original, the 95% chance of making all 6 inferences correctly drops to 54% with their original sample size of 700 per condition per cluster."

In the authors' most recent response, they write:

"With this number of participants, we will still be able to detect half of the original effect size with 63% probability for Study 1, and 58% probability for Study 2."

These numbers, while consistent with my earlier estimate, are not very encouraging. And one must bear in mind that these discouraging estimates do not (as far as I can tell) account for potentially massive exclusions (up to 70-80%) due to failed comprehension and/or familiarity.

Summary and recommendation regarding statistical power issues

At this point it's worth taking stock and considering the combined weight and implications of the problems outlined above. As I acknowledged at the outset, it is surprising—it certainly was to me—that a group of 100 labs aiming to collect 17,000+ participants would lack the power to conclusively replicate (or fail to replicate) two studies. And, yet, that is what the foregoing suggests, accounting for the use of six conditions in each of three cultural clusters using convenience samples that are likely to have high levels of failed comprehension and familiarity with the testing materials—plus the need to detect effects substantially smaller than the original effects.

What follows from this? Am I saying, in effect, that the original results can never be put to the test? No. I'm saying that the authors are trying to do too much at once.

If the authors want to move forward with this project, my recommendation is that they focus on the key contrast of Study 1 and drop Study 2.

About half of the labs are testing from the Western cluster, so let's focus on that. Let's assume that this accounts for half of the 17,305 participants, which is 8,517. Focusing on the key personal force contrast for Study 1 requires two dilemmas, *footbridge pole* and *footbridge switch*. (The authors can also run the corresponding "speedboat" dilemmas). This requires only two cells, so we have 4,259 participants per cell, prior to exclusion. If 40% fail the comprehension checks, we have 2,555 per cell. If 60% report being familiar with the testing materials (and it could be higher), that leaves 1,022 per cell. (Again, I think that the results should be prominently reported with and without exclusion for familiarity) That's more than the 700 the authors were aiming for, but this gives the researchers more power to detect a smaller effect. I haven't done all the calculations, but it sounds to me like this could work.

As for the cross-cultural element, this can, at the very least, be done in an exploratory way. And it's possible that there will be, after the necessary exclusions, enough power to compare all three clusters with full control for multiple comparisons. The authors will have fewer initial participants in the non-Western clusters, but they are less likely to be familiar with the testing materials and, if they are not being recruited through Mturk or similar, they will likely have higher rates of comprehension.

By dropping Study 2, the number of cells to be tested gets cut by 2/3. And the interaction reported in Study 2 is a much weaker effect, requiring a much larger sample size to detect it. Also, as explained below, there are major problems with the current plan for the design of Study 2—at least for the secondary dilemmas.

Finally, Study 1 is more important than Study 2. The field of moral cognition recognizes two papers as properly documenting the effect of "personalness" in responses to moral dilemmas. Cushman et al. (2006) documented the effect of "physical contact", and this paper is rightly more influential than my 2009 paper (which includes Cushman as a co-author). Study 1 of Greene et al. then refined the "personalness" factor, replacing "physical contact" with "personal force". Study 2, showing the interaction between personal force and intention, is secondary, despite its prominence in the title. It's still an interesting result, but nothing much turns on whether the effects of these two factors are multiplicative rather than additive.

In sum, I think the best way forward is to drop Study 2. Or, to put the point more strongly, I doubt that it will be possible to achieve adequate power without dropping Study 2.

My main regret about the more focused plan that I am recommending is that it focuses too much on the specific personal force contrast and not enough on the overall trolley problem. The ideal plan would be to test the *standard footbridge* and *standard switch* cases as well, so that we can see how the effect of personal force fits into the larger picture.

The scientific problem is to explain why people tend to give utilitarian responses cases like *standard switch*, but not to *standard footbridge*. Personal force, then, is a partial answer to this question. So far, the data have supported the idea that this is an important factor (e.g. See meta-analysis by Feltz & May, 2017), but if the present study says otherwise, then it will be important to know what *does* account for the difference between the two standard cases, and having the two standard cases would provide important clues, and perhaps a clear answer (e.g. an effect of physical contact: *standard footbridge* vs. *footbridge pole*).

Including the *standard switch* and *standard footbridge* cases in the study would mean doubling the sample size, beyond what is needed for *footbridge switch* and *footbridge pole*. But this might be possible, and would definitely be worth it if it is possible. I recommend that the authors conduct power analyses to evaluate the feasibility of running the two standard dilemmas in addition to *footbridge switch* and *footbridge pole*, all a part of a single study.

But whatever is decided about this, the take-home message is: For adequate power, drop Study 2.

The center of the prior distribution for the alternative hypothesis

In my original comments I made the point that the prior for the alternative hypothesis should be centered, not at zero, but based on the original effect size. The authors write:

"We will use informed priors for the alternative model: a one-tailed Cauchy distribution with scale r = 0.3 (Hypothesis 1a and 1b) and r = 0.37 (Hypothesis 2a and 2b) on the standardized effect size using the BayesFactor package in R[43] for the analysis."

I appreciate these clear statements about the scale factor. However, I think it's important to be explicit about where the prior distribution for the alternative hypothesis will be centered, as the default assumption may be that it's centered at zero. Following the logic of the original plan for robustness checks, I think it makes sense to center the prior at 2/3 of the original effect size. I ask that the analysis plan be revised to reflect this.

The use of the "burning" dilemmas in Study 2

Once again, I recommend dropping Study 2 to provide adequate statistical power. Nevertheless, I feel that I should comment on some significant problems with the revised design of Study 2.

Study 2 asks whether there is an interaction between the effect of intention (means vs. side- effect) and the effect of personal force. There are four trolley cases used to examine this interaction:

standard switch: NO personal force, NO harm as intended means
standard footbridge: YES personal force, YES harm a intended
means *obstacle collide*: YES personal force, NO harm as intended
means *loop*: NO personal force, YES harm as intended means

If Study 2 were to move ahead, I think it would be fine to use these dilemmas. This design substitutes *standard switch* for *loop weight* in the original and likewise substitutes *standard footbridge* for *obstacle push*. I think that these substitutions have value, and that they could be more clearly justified in the paper. But overall this design makes sense.

The problem is with two of the secondary dilemmas that are designed to re-create this 2 x 2 design using non-trolley cases. These two are the "burning" dilemmas.

The design for the secondary dilemmas of Study 2 looks like this:

standard switch—speedboat version- : NO personal force, NO harm as intended means
standard footbridge—speedboat version: YES personal force, YES harm a intended
means *obstacle collide—burning version*: YES personal force, NO harm as intended
means *loop—burning version*: NO personal force, YES harm as intended means

(Note: I'm using the dilemma labels as they appear in the supplementary materials, where the speedboat and burning dilemmas are named with reference to the trolley cases that they are intended to match.)

The burning dilemmas come from Cushman et al., 2006 and were among the nineteen dilemmas used in Study 2b of Greene et al., 2009. These dilemmas involve a fireman trying to save five children. He must break a large window in order to save them, and breaking this window will involve killing a man. In one case the harm occurs with personal force and as a side-effect (as in *obstacle collide*). In the other case the harm is supposed to occur without personal force and as a means (as in *loop*).

The authors chose the "burning" dilemmas because they exhibited the largest effect of intention (means vs. side-effect) in Cushman et al., 2006. However, this is not a good criterion for selection because the focus of Study 2 is not on the main effect of intention, but rather on the effect of personal force and its *interaction* with intention. As it happens, the burning dilemmas show no sign of exhibiting the predicted interaction in Cushman et al., 2006.

Table 1 of Cushman et al. breaks down the effects for each pair of contrasting dilemmas. The burning dilemmas exhibit a whopping effect of intention (d=.7), more than double that of any other contrast set. The burning dilemmas are also unusual in that they are the only dilemma set that have the intention variable fully crossed with physical contact variable. (Recall that in Cushman et al. the dilemmas are coded in terms of "physical contact" rather than "personal force", but for the burning cases these two variables line up.) This allows Cushman et al. to report on two effects of personal force for the burning dilemmas, one for the pair in which the harm is intended as a means and one for the pair in which the harm is side-effect. (See Cushman et al., Table 1, bottom)

For the two burning cases where the harm is intended as a means, the effect size for the personal contact/force variable is d=.17—pretty small. However, for the two burning cases in which the harm is a side-effect, the effect size for personal force/contact is d=.3—notably larger. Thus, the burning dilemmas *do not* exhibit the predicted interaction. They exhibit a main effect of personal force, and if there is an interaction, it goes the other way. That is, the effect of personal force is larger when the harm is a *side-effect*.

At this point, one might wonder how this is possible, given that these dilemmas were used in Study 2b of Greene et al., 2009, which did find the predicted interaction. The answer is that the burning dilemmas were just four of the nineteen dilemmas used in Study 2b. And, thus, it seems that the interaction observed in Study 2b of Greene et al. was *in spite of* the contributions of these particular dilemmas.

In the context of a replication effort, it doesn't make sense single out a sub-set of dilemmas that actually *worked against* the effect that was observed with the full set. Thus, the authors should not use the burning dilemmas.[2,3]

[2 This raises the question: Why don't the burning cases exhibit the predicted pattern? A closer look at these dilemmas suggests why they don't fit with the others. The burning case that is supposed to be equivalent to the *loop* case involves throwing a piece of burning debris at someone. This is a case that is supposed to NOT involve personal force. Technically, this doesn't qualify as personal force because the burning debris is thrown with the expectation that the person will jump out of the way, and thus won't be physically impacted by the debris. But throwing burning debris at someone is pretty different form hitting a switch, as in the *loop* case.]

[3 The question at the end of the other burning dilemma reads as follows: "Is it morally acceptable for John to hit the man with his fist in order to avoid the deaths of the five children, causing the death of the single man instead?" This is supposed to be a side-effect case (equivalent to *obstacle collide*), but the "in order to" makes it sound as if hitting the man is a means rather than a side-effect. If one were to use this case, the question would need to be re- worded. But, again, I don't think the burning cases should be used if Study 2 is to go forward, despite my recommendation to drop it.]

A further problem with using the burning cases is that these two dilemmas form a diagonal in the design matrix. This means that all of the simple effects are confounded with all of the incidental differences between the speedboat scenarios and the burning scenarios. For example, to isolate the effect of personal force within the side-effect cases, one would compare the burning version of obstacle collide to the speedboat version of standard switch. Likewise, the other three meaningful pairwise contrasts compare a speedboat case to a burning case.

So, what should the authors do? Once again, my recommendation is drop Study 2 entirely in order to conserve statistical power. Nevertheless, before I had adequately thought through the statistical power issues, I drafted some additional "speedboat" dilemmas for Study 2 that could be used to replace the "burning" dilemmas. I can share these dilemmas if necessary, but I hope that they are no longer of interest.

Justifying and making full use of the *standard switch* and *standard footbridge* dilemmas in Study 2

The authors are currently planning to use the *standard switch* and *standard footbridge* dilemmas in Study 2, instead of the original *loop weight* and *obstacle push* dilemmas. I hope that this is a moot point, but I will make some comments about this just in case.

I had originally recommended including the two "standard" dilemmas as part of Study 1, so as to help with the interpretation of the results, especially if the effect of personal force does not replicate. Once again, the rationale is as follows: Given the longevity of the trolley problem, it is very unlikely the difference between *standard switch* and *standard footbridge* will disappear.

The cases that test for the effect of personal force (*footbridge switch* and *footbridge pole*) are both intermediate between the two "standard" cases. Thus, if the effect of personal force does not replicate (in some or even all cultural clusters), having data for the two standard cases would enable one to diagnose what happened: If the gap between *footbridge pole* and *remote footbridge* disappears, but the canonical gap between *standard footbridge* and *standard switch* persists, then having those two anchors can tell us where the missing variance has gone, and thus which factors do explain the main scientific puzzle of the trolley problem.

Again, I recommend dropping Study 2. But if Study 2 using the two standard cases goes ahead, I have two recommendations. First, the *standard footbridge* case should be compared to *footbridge pole* form Study 1 to look for an (unexpected) effect of physical contact. Second, the *standard* switch case should be compared to the *footbridge switch* case from Study 1 to confirm the effect of the means/side-effect factor. (This contrast involves a possible effect of distance, previously ruled out by the *remote footbridge* case). Third, it should be explained that these dilemmas are being included to facilitate these comparisons.

Other issues, organized by page

Pg. 6 "The first and second hypotheses, that the effects of personal force and intention are present in different cultures, comes from the view that their effect on moral judgements is attributed to emotional processes[9,19,20], and specifically social emotions (such as guilt, shame or regret)[20,21].

It's not clear to me how a process' being emotional *per se* make it more likely to be present across cultures. A specific phrase or gesture might be considered extremely rude in a culture, and thus people in that culture might respond to it emotionally. But it does not follow from the emotional nature of this response that it will be universal.

A better explanation for the apparent universality of this response is that it's related to something that all societies regulate, namely basic interpersonal violence. See Cushman on "simulating murder" (Emotion, 2012) and his general reinforcement learning framework in his 2013 PSPR paper. (See also Chapter 9, "Alarming Acts", of Moral Tribes)

I also don't think it's established that the emotional response to the idea of committing intentionally harmful actions using personal force is one of guilt, shame or regret. These strike me as primarily retrospective emotions. We may not have a good name for the feeling that an action one is contemplating would be wrong or otherwise bad. Perhaps the closest thing is anxiety. (See Perkins et al., JEPG, 2013).

Pg. 7. "In the third part, we are going to test our hypothesis according to which collectivism moderates the effect of intention and personal force (Hypothesis 3)." It's not clear to me why this is presented as being in a "third part", when this is an aspect Studies 1-2, rather than the subject of a third study.

Pg. 8. "for example, in the *standard footbridge* scenario both intention and physical contact are required to push the man off the bridge"

It's not clear to me why physical contact is being mentioned here since the studies here focus on personal force (although the two usually coincide).

"All the materials are presented in the Supplemental material B at https://osf.io/h9xpe/." This links to the older set of supplementary materials from April.

Pg. 12. Description of Study 2 hypotheses. (I hope this is a moot point, but I am including it just in case.) The way Hypotheses 2a and 2b are described (pg. 12) is not fully consistent with the analysis plan. Here it says that the hypothesis is that there will be an effect of the means/side- effect factor when personal force is applied. This is consistent with the prior results, but, strictly speaking, this hypothesis does not require an interaction between the means/side-effect factor and the personal force factor. That is, one can have an effect of means/side-effect in the presence of personal force simply by having a main effect of means/side-effect and a main effect of personal force. On pg. 13 the authors state that they will test for this interaction, and this should be reflected in the way H2a and H2b are described. More specifically, it should state the expected direction of the interaction: the means/side-effect factor is expected to be larger when personal force is present, as compared to when personal force is absent.

Pg. 14. Data collection procedures: (Also a moot point, I hope) The authors outline a sequential process for data collection (pg. 14), where the authors may stop or continue depending on the first set of results. But earlier (pg. 12) it says, “We will conduct no analyses on our data before we stop data collection and we will not collect new data after conducting our analyses.“ I assume that the plan is to go with the sequential procedure, and, if so, then the statement on pg. 12 should be removed or clarified. The authors also write “On Study 2, we will conduct the analysis only when we exhausted our resources.” Does this mean that they won’t be using the sequential procedure for Study 2? How is this possible if Study 1 and Study 2 are being run at the same time?

Pg. 15. “We will test whether collectivism interacts with personal force and intention (Hypothesis 3), as explained in the introduction.”

It’s not clear to me what this means. Is this a test for an interaction between individualism- collectivism and the personal force factor and then, separately, an interaction between individualism-collectivism and the means/side-effect factor? Or are the authors looking for a three-way interaction? All three interactive effects? (FWIW, I would not expect a three-way interaction)

Pg. 29 of supplemental materials:

The attention check question presented in the supplemental materials says:

“Which option describes the most accurate the situation that you were first presented with?”

I would say, “Which description below most accurately describes the *first* dilemma that you saw?”

And I would add an attention check question for the second dilemma, too. This will not only help with quality control for the second dilemma, but it will also help people keep straight which came first and which came second.

For the attention check options, a bit of polishing would help. For example, the first option reads:

“a. Joe had to run through the footbridge to hit a switch, but by doing so, he had to collide with a single workman who fell off of the bridge and would die. (obstacle collide)”

I would say “run across the footbridge”. And the wording here assumes that Joe chose to perform the action, but the dilemma simply presents it as a possibility.

My apologies for the length of these comments. But given the resources being invested in this project and the weight that will be attached to its conclusions, I would like to be as thorough as possible.

Once again, I am grateful for the opportunity to provide input, and I appreciate the great effort that has gone into this project already, and the even larger effort to come, assuming it moves forward. I am very much looking forward to seeing the results.

With best wishes,
jdg

Author Rebuttal, first revision:

We would like to thank you, Prof. Greene ,and the 2 anonymous reviewers for the comments on our Stage 1 Registered Report. Below, we addressed all the points you have raised.

Reviewers' Comments:

<u>Reviewer #2:</u>

Remarks to the Author:
Overall it seems the authors have been responsive to the comments and suggestions raised by reviewers. I am especially pleased to see the edits to the wording as I think this version is much clearer

regarding terminology, and to see the addition of the Greene dilemmas following the ones in the main document--although this is more work, it is likely to increase the scientific value of this project. Likewise with the clarifications of the hypotheses, which helps this paper have better impact. However, I do think the authors could lean a little further in this direction. The paper still reads like the cultural hypotheses were tacked on at the end rather than an organic reason to do the study.

For one thing they appear late in the paper and are worded in such passive voicing that they stand out from the rest of the text,

For example "Our third hypothesis is based on the belief that even though personal force and intention are cultural universals, certain cultural differences are expected to moderate their effect" could get reworded as something like "Even though we anticipate that personal force and intention will emerge universally across cultures, we nonetheless expect cultural differences to moderate these effects. In particular..." Note that such wording is easier for readers to process, making your claims seem intuitive. As is, this portion remains cognitively demanding for readers and hence less impactful than might be hoped. Small edits will help.

The abstract was another place that jumped out at me as worth improving. For example, the abstract focuses primarily on the question of generalizability, with a large portion of words dedicated thus:

"As their study was conducted with US samples, our knowledge is limited concerning the universality of this effect, in general, and the impact of culture on the situational and psychological factors of moral judgements, in particular. Here, we empirically test the universality of intent and personal force on moral dilemma judgements by testing the replicability of the experiments "

Meanwhile, the cultural hypothesis portion of the abstract gets only "We test three hypotheses regarding the effect of culture on moral dilemma judgements."

Consider that the majority of readers will not advance beyond the abstract--they are left without any sense of what cultural differences are getting tested and what direction the effects might be. The authors would gain a lot by slightly trimming the words focused on limited knowledge about generalizability etc., and instead clarifying the actual theoretical meat, such as:

"We hypothesize that intent and personal force will universally increase the unacceptability of outcome-maximizing harm in dilemmas, this effect will be stronger in collectivistic than individualistic cultures due to increased guilt and shame "

Such wording is only a rough draft and no doubt the authors can improve on it, but this sort of language is I think what the other reviewer was hoping for when they suggested clarifying the theoretical reasons for doing the study (I sentiment I obviously agree with). Such wording helps organically help readers envision interesting potential outcomes at a glance, and makes your hypotheses seem clear and integral to the project.

Response: Thank you. We followed your advice and updated these parts of the introduction and abstract.

SImilarly, it strikes me that the title is a bit generic and does not fully communicate all the key ideas in play--nor it does not convey a sense of the direction and nature of the hypotheses. Right now it reads:

"Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements"

but think of all the future people searching by keywords and such who might skip this paper based on the title--for example anyone interested in individualism-collectivism. A few tweaks might buy you more information in the title (you could update the in the eventual version with data)

"Personal force and intention universally increase the unacceptability of sacrificial harm, especially in collectivistic cultures: A Global Study"

Again this is only a suggested draft title, but the point is I think the authors can improve clarify, precision, and tone especially in the title, abstract, and the new hypothesis section, to make sure this paper ends up finding and communicating effectively to the audience it deserves.

Response: We agree, however, at this moment the title cannot refer to results as we do not have them yet. We propose to have the opportunity to change the title before Stage 2 submission, so it can better describe the results.

I only have one other outstanding issue: I am somewhat concerned by the response to Dr. Greene's concerns over familiarity as a decision to exclude all people who score 4 or 5.

There are two problems with this: over the past decades trolley style dilemmas have grown so famous (even featuring in mainstream American TV such as The Good Place) that this decision risks excluding a very high percent of the sample (i.e., possibly upwards of 60% in North America), hence ruining power. Second, its highly likely that this decision will systematically throw out more people from some groups than others (i.e., those most familiar with such cultural touchstones), skewing this otherwise fairly representative global sample. It seems like there might be wiser alternatives to this extreme decision. For example, the authors should pledge a priori to specific analyses. such as:

-the core analysis with and without controlling for familiarity, and documenting the difference in effect size
-the core analysis with and without excluding people who score 4 or 5 on familiarity, and documenting the difference in effect size

In other words, rather than cutting off a bug chunk of the full sample, arm the reader with all the information to reach their own conclusions over this factor. Moreover, such analyses would add to the value of this paper. People debate how much familiarity impacts dilemma judgments; with a sample this large you may be able to help put a tight confidence interval around that effect.

On the other hand, I am perfectly content with excluding all people who fail the dilemma attention check question described in Supplementary Material B.

Response: We followed your advice and pre-registered additional analyses regarding familiarity (see response to Editor). This is what we write:

"A potentially large number of participants will be excluded due to familiarity with the trolley dilemma, and there is a possibility that this exclusion criterion will affect the data from some countries or cultural clusters more than others. To avoid this potential sampling bias, we will compute all above-listed analyses on moral dilemmas (confirmatory and exploratory) on the full sample in which we do not exclude the participants who were familiar with the trolley problem. Second, we will compute all analyses specifically on data coming from people who were familiar with the trolley problem in order to compare the results of "familiar" and "unfamiliar" participants."

Reviewer #3:

Remarks to the Author:
I am happy with the revisions the authors have made. They will use models of H1 informed by the original study. (As the function of a model of H1 is to represent the predictions of the theory - given evidence depends on how well data are predicted - the authors are right that in this context it is the predicted effect size, and its range, that is important. When power is used in frequentist statistics, by contrast, it is, as the editor says, the minimal meaningful effect size that must be used, because in this

case the aim of power is to guard against missing out on any effect of interest. As the authors say they will make inferences with respect to the Bayesian analyses, they in fact do not have to justify a minimal effect. As an aside, this is one reason I find Bayes factors so useful!) The authors also now consider robustness explicitly. The authors have also added some theory to the introduction - which, as the edtior says, does make better science.

Response: Thank you for your nice words and support.

Response to the Original author

[The lengthy comments are summarised below.]

Statistical power

[The original author suggested that we might not have enough participants to reliably detect the interaction effect in Study 2, hence, Study 2 should be dropped and the focus should be on Study 1.]

Response: As explained to the editor, we are willing to prioritise Study 1 over Study 2. That is, we will only launch Study 2 after we reached one of the Bayes Factor thresholds in Study 1. However, in the case that we do not need to exclude high proportion of participants in Study 1, or that the effect is easily detectable, it would be a waste not to use the remaining participants to collect data for Study 2. This is what we write:

"As the data are planned to be collected globally, our knowledge is insufficient concerning the noise of the measurement and the rate of exclusion in the various samples, which are needed for an accurate sample size estimation. For this reason, we propose a sequential data acquisition. That is, first, we will launch Study 1 (Hypotheses 1a and 1b), and collect data in sequences from 500 participants per cluster per condition; from 3,000 participants altogether (after all exclusions). We will stop data collection after each sequence. At these stops, we will conduct our planned Bayesian analyses. Should the BF reach the preset thresholds in a given cluster, we will stop data collection for that cluster. If, in a cluster, the BF thresholds were not reached, we will continue data collection with 200 additional participants per cluster per condition, and then re-analyse the data, repeating this procedure until one of the BF thresholds is reached, or the participant pool is exhausted. Should we not reach this limit with our present capacity of ~19,000 participants, we would extend the data collection to a new semester. In the unlikely case

that we do not reach our evidence threshold within 12 months, we would report our final results, acknowledging the limited strength of the findings. We will launch Study 2 data collection in a given cluster only when the analysis of Study 1 is conclusive. In Study 2, we will conduct the analysis only when we have exhausted our resources."

Incorporation of the original effect sizes

[The original author argued that we miscalculated the effect sizes in the original study]

Response: We calculated our target effect sizes based on the results of Study 1a and 1b together, as well as Study 2a and 2b together (Figures 2 and 4). The effect sizes of 1b and 2b are generally larger than 1a and 2a and we wanted to base our estimate on the most available data. For example, the ⅔ of the effect size of Study 2a is 0.19, while this number is only ⅙ of the effect size observed in Study 2b, making it a likely underestimate. However, with this number of participants, using priors based on the ⅔ of the effects of Studies 1a and 2a, will not affect the results in any way. Hence, as our most important goal is to replicate Study 1a and 2a, we decided to go with the suggestions of the reviewer, and use these numbers (⅔ of the original effect sizes, that is, 0.26 in Study 1 and 0.19 in Study 2).

The center of the prior distribution for the alternative hypothesis

Response:The reviewer seems to suggest to calculate a Bayes factor with a data-driven model of H1 (cf., Replication Bayes factor by Ly, Etz, Marsman, & Wagenmakers, 2018) rather than a Bayes factor that assesses the level of evidence for a theory-driven model of H1 against H0. We do not agree with the assumption of the data-driven model of H1 that the most likely effect size is the one that was found by the original study. We believe that small effect sizes are more likely (under H1) than large ones. Hence, we opted for a half-Cauchy distribution with a mode of zero.

The use of the "burning" dilemmas in Study 2

[The original author suggested that we should not use the "burning" moral dilemmas we proposed, but instead we should use the speedboat dilemmas.]

Response: Thank you for providing us with the Speedboat versions of these dilemmas. We will use these versions instead of the "burning man" dilemmas.

Justifying and making full use of the standard switch and standard footbridge dilemmas in Study 2

[The original author suggested that we should provide a better description of the justification behind our dilemma choice in the manuscript.]

Response: We pointed out more directly why these dilemmas are being used here, and detailed the additional analyses. This is what we write:

"The third difference is that for Study 2, we will use different moral dilemmas than those that were used by Greene et al.; the standard switch and footbridge dilemmas will be used instead of the loop weight and obstacle push dilemmas, respectively. These dilemmas are not different from the ones used by Greene et al. in their structural characteristics, only on surface characteristics. That is, in the standard switch the harm is unintended and no personal force is required, while in the standard footbridge dilemma, the harm is intended and requires personal force. By including the standard switch and standard footbridge scenarios instead of the original ones, we gain further insight into the data. Imagine for example, that the personal force effect does not replicate in one of the cultural clusters. One explanation for this is that people are simply not sensitive to the effect of personal force in that cluster. However, it might also be the case that utilitarian response rates to similar dilemmas increase over time[44]*. If so, we should see that the replicated difference between the standard footbridge and switch dilemmas is shrinking or disappeared. Furthermore, by comparing the standard footbridge to the footbridge pole dilemmas, we can test the effect of physical contact, and by comparing the standard switch case to the footbridge switch case to confirm the effect of intention."*

"Effect of physical contact and intention. *With this set of items, we will be able to assess the effect of physical contact, by comparing the standard footbridge and footbridge pole dilemmas. We will also assess the effect of intention by comparing the standard switch case with the footbridge switch case. These analyses will be done in every cluster, and we will use Bayesian t-tests for these comparisons. We will use the same prior we use for the assessment of the effect of physical force (d = 0.26). This analysis will be done separately for the trolley and speedboat dilemmas.*

Comparing the standard switch and standard footbridge dilemmas. *For the reasons explained earlier, we are planning to compare the standard footbridge and standard switch dilemmas, in each cultural cluster. For this, we will conduct a Bayesian t-test, with the same prior previously used for the assessment of the effect of physical force (d = 0.26). This analysis will be done separately on the trolley and speedboat dilemmas."*

Other issues, organized by page

Pg. 6 “The first and second hypotheses, that the effects of personal force and intention are present in different cultures, comes from the view that their effect on moral judgements is attributed to emotional processes, and specifically social emotions (such as guilt, shame or regret).

It’s not clear to me how a process’ being emotional per se make it more likely to be present across cultures. A specific phrase or gesture might be considered extremely rude in a culture, and thus people in that culture might respond to it emotionally. But it does not follow from the emotional nature of this response that it will be universal.

A better explanation for the apparent universality of this response is that it’s related to something that all societies regulate, namely basic interpersonal violence. See Cushman on “simulating murder” (Emotion, 2012) and his general reinforcement learning framework in his 2013 PSPR paper. (See also Chapter 9, “Alarming Acts”, of Moral Tribes)

Response: Thank you for pointing this out. We updated our discussion on this issue accordingly. This is what we write:

“The first and second hypotheses, that the effects of personal force and intention on moral judgements are culturally universal, come from their relatedness to interpersonal violence. People seem to exhibit a general tendency to avoid causing violent harms (e.g., murder)[19,20], and they are more likely to perceive actions as violent or harmful when they supposed to use personal force or intention[3]. As a result, people are more likely to behave in a deontological way when personal force or intention is present in the dilemma. As all cultures regulate interpersonal violence[21], we expect to find that both intention and personal force, as well as their interaction, have an effect on moral judgements in different cultures.”

I also don’t think it’s established that the emotional response to the idea of committing intentionally harmful actions using personal force is one of guilt, shame or regret. These strike me as primarily retrospective emotions. We may not have a good name for the feeling that an action one is contemplating would be wrong or otherwise bad. Perhaps the closest thing is anxiety. (See Perkins et al., JEPG, 2013).

Response: We included this idea in the discussion of the hypothesis. This is what we write:

"Even though we anticipate that the effect of personal force and intention will emerge universally across cultures, we nonetheless expect cultural differences to moderate these effects. The effect of personal force on moral judgement has been attributed to emotional processes[9,22–24], specifically social emotions (such as guilt, shame or regret)[23,25]; the potential use of personal force makes people feel guilt or shame before making a decision, and, therefore, they will rate actions that use personal force as morally less acceptable. There is a convincing argument that these social emotions are universal[26–28], with some cultural variation in their intensity and the social contexts in which they are experienced[26–28]. It has been argued that shame and guilt are more important in interdependent, collectivistic cultures (as their function is argued to be social control); people living in East Asian countries have reported experiencing them more frequently and more intensively[26–28]. Other findings suggest that it is anxiety that mediates the effect of intention and personal force[24], but anxiety (social anxiety in particular) has also been positively associated with collectivism[29], pointing to the same direction. Hence, we hypothesize that people living in collectivistic cultures will judge actions that involve personal force and intention as morally less acceptable than people in individualistic cultures. Utilitarian responding in moral dilemma judgements has also been associated with low levels of empathic concern[30]. People living in collectivistic cultures have also been suggested to exhibit higher levels of empathy[31,32]. Hence, we predict that individualism/collectivism factor will also have a main effect on utilitarian responding: collectivists will be less utilitarian in general, due to their higher levels of empathic concern."

Pg. 7. "In the third part, we are going to test our hypothesis according to which collectivism moderates the effect of intention and personal force (Hypothesis 3)." It's not clear to me why this is presented as being in a "third part", when this is an aspect Studies 1-2, rather than the subject of a third study.

Pg. 8. “for example, in the standard footbridge scenario both intention and physical contact are required to push the man off the bridge” It’s not clear to me why physical contact is being mentioned here since the studies here focus on personal force (although the two usually coincide).

“All the materials are presented in the Supplemental material B at https://osf.io/h9xpe/.” This links to the older set of supplementary materials from April.

Pg. 12. Description of Study 2 hypotheses. (I hope this is a moot point, but I am including it just in case.) The way Hypotheses 2a and 2b are described (pg. 12) is not fully consistent with the analysis plan. Here it says that the hypothesis is that there will be an effect of the means/side effect factor when personal force is applied. This is consistent with the prior results, but, strictly speaking, this hypothesis does not require an interaction between the means/side-effect factor and the personal force factor. That is, one can have an effect of means/side-effect in the presence of personal force simply by having a main effect of means/side-effect and a main effect of personal force. On pg. 13 the authors state that they will test for this interaction, and this should be reflected in the way H2a and H2b are described. More specifically, it should state the expected direction of the interaction: the means/side-effect factor is expected to be larger when personal force is present, as compared to when personal force is absent

Pg. 15. “We will test whether collectivism interacts with personal force and intention (Hypothesis 3), as explained in the introduction.” It’s not clear to me what this means. Is this a test for an interaction between individualism/collectivism and the personal force factor and then, separately, an interaction between individualism-collectivism and the means/side-effect factor? Or are the authors looking for a three-way interaction? All three interactive effects? (FWIW, I would not expect a three-way interaction)

Response: Thank you. We made all these minor changes and added further information wherever was necessary based on your suggestions.

Pg. 29 of supplemental materials: The attention check question presented in the supplemental materials says: “Which option describes the most accurate the situation that you were first presented with?” I would say, “Which description below most accurately describes the first dilemma that you saw?” And I would add an attention check question for the second dilemma, too. This will not only help with quality control for the second dilemma, but it will also help people keep straight which came first and which came second. For the attention check options, a bit of polishing would help. For example, the first option reads: “a. Joe had to run through the footbridge to hit a switch, but by doing so, he had to collide with a single workman who fell off of the bridge and would die. (obstacle collide)” I would say “run across the footbridge”. And the wording here assumes that Joe chose to perform the action, but the dilemma simply presents it as a possibility.

Response: We added a second attention check question, and made the changes on the previously existing one. We also updated our exclusion criteria as follows:

"Moreover, we will introduce two additional questions (posed right after the moral dilemmas), asking participants about the specifics of the trolley and speedboat scenarios that they had been presented with, in order to test whether they were paying attention when reading the scenario. Specifically, they will be asked to select the option which most accurately describes the situation that they had been presented with. Each option will describe the nature of the physical action; the key manipulation in the experiment. As the attention on the trolley and speedboat dilemmas will be measured by different questions, when analysing the responses on the trolley dilemma, we will exclude the data of those giving an incorrect response to the trolley attention check question, and same for the speedboat dilemma. For example, people who gave a correct response on the trolley attention check question, but not on the speedboat version, will be included when analysing the trolley dilemma, and excluded when analysing the speedboat version."

Decision Letter, second revision:

30th January 2020

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Bence

Thank you once again for submitting your revised Stage 1 Registered Report, entitled "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgements</p>." Everything is in order and I am delighted to say that we can offer acceptance in principle. You may progress to Stage 2 and complete the study as approved.

As you know, a condition of in-principle-acceptance is that the authors agree to deposit their Stage 1 accepted protocol in a repository, either publicly or under embargo until Stage 2 acceptance and publication. We are very keen to showcase our in-principle accepted protocols, so that our readers, reviewers, and potential authors can gain insight into the requirements of the format as well as an idea of the types of projects that are suitable for publication in Nature Human Behaviour. We have set up a space on figshare (https://springernature.figshare.com/registered-reports_NHB) to host all of our in-principle accepted protocols, which can either be made public or kept under embargo until Stage 2 acceptance (depending on author preference). This gives you the opportunity to have your work publicly associated with Nature Human Behaviour, and of course we will be very pleased to showcase your report if you agree to share it publicly.

Depositing the work on our figshare space does not preclude deposition of your Stage 1 protocol on other depositories – your protocol can also be posted on OSF, Dataverse, Dryad or any other public repository of your choice. You also do not need to do anything – if you agree with posting your protocol on our figshare space, we will upload your protocol on your behalf and either set it public or place it under embargo, depending on your choice. Your protocol will be licensed under a CC BY license (Creative Commons Attribution 4.0 International License). The CC BY license allows for maximum dissemination and re-use of open access materials and is preferred by many research funding bodies. Under this license users are free to share (copy, distribute and transmit) and remix (adapt) the contribution including for commercial purposes, providing they attribute the contribution in the manner specified by the author or licensor (read full legal code: http://creativecommons.org/licenses/by/4.0/legalcode) Please note that any use of https://springernature.figshare.com will be subject to the Figshare terms of use. Figshare has the right to enforce these terms and conditions where applicable. Use of third party services and sites will be subject to the relevant terms of use and will apply if we act on your behalf in this regard. Do let me know if you would like to take up this option or if you have any questions regarding the protocol deposition requirement.

Following completion of your study, we invite you to resubmit your paper for peer review as a Stage 2 Registered Report. Please note that your manuscript can still be rejected for publication at Stage 2 if the Editors consider any of the following to hold:

• The results were unable to test the authors' proposed hypotheses by failing to meet the approved outcome-neutral criteria
• The authors altered the Introduction, rationale, or hypotheses, as approved in the Stage 1 submission
• The authors failed to adhere closely to the registered experimental procedures
• Any post hoc (unregistered) analyses were either unjustified, insufficiently caveated, or overly dominant in shaping the authors' conclusions
• The authors' conclusions were not justified given the data obtained

We encourage you to read the complete guidelines for authors concerning Stage 2 submissions at http://go.nature.com/2hKPe06. Please especially note the requirements for protocol deposition, data sharing, and that withdrawing your manuscript will result in publication of a Retracted Registration.

When you are ready, please use the following link to access your home page and submit your Stage 2 Registered Report:

[REDACTED]

*This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first.

We expect your Stage 2 Registered Report to be submitted by the date specified in your latest cover letter. If unforeseen circumstances prevent submission by that date, please contact us as soon as possible to discuss any changes to the submission time-frame.

Thank you again for offering us this work and we look forward to receiving your Stage 2 Registered Report.

Best wishes,
Marike

Anne-Marike Schiffer
Editor
Nature Human Behaviour

Stage 2, post-study:

Decision Letter, initial version:

4th March 2021

Dear Bence

RE: "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgements</p>"

Thank you for submitting your Stage 2 Registered Report.

Before we can send the work out to peer-review, we ask you to revise the manuscript so that it complies with our Registered Report format; this request is motivated by the fact that a manuscript that complies with editorial policies at this stage will require fewer rounds of peer-review.

I highlight in particular 5 issues that we'd ask you to address:

1) Please present the sections in the right order: Abstract, Introduction, Results, Discussion, Methods, References. Please note that you may have to revise the Results to allow the reader to better follow your presentation without reference to the Methods section.

2) Please revise how you describe the strength of evidence derived from BFs. It's journal policy to adopt the categories presented in Schönbrodt, F.D., Wagenmakers, E. Bayes factor design analysis: Planning for compelling evidence. Psychon Bull Rev 25, 128–142 (2018) https://rdcu.be/b6uOC ; this means that the BF = 9.5 (and similar) must be described as moderate evidence, not "weak" or "nearly strong" as it is currently the case.

3) Please avoid the use of bar graphs. Instead, please use Figures that show the nature of the distributions (e.g., dotplots, violin plots).

4) Please report Credible Intervals.

5) In your description of how it was decided that the sampling criteria may be amended, you may want to reference that the distribution of responses at the time (in terms of samples and exclusions per country) was factored into the decision.

Before we can send the manuscript back to our reviewers, we ask that you revise it to ensure that it complies fully with our policies and is formatted according to our requirements. I have attached another copy of our checklist, as well as the template document that exemplifies our formatting and policy requirements. If you are uncertain as to how to address any of the points in the checklist, please don't hesitate to contact me.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you in advance for attending to these requests and I look forward to receiving your revised manuscript.

Best wishes,
Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

Decision Letter, first revision:

26th May 2021

*Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors.

Dear Bence,

Thank you once again for submitting your Stage 2 Registered Report, entitled "<p>Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements</p>," and for your patience during the re-review process.

Your manuscript has now been evaluated by Reviewers 1, and 2 from the previous rounds of review, and a new statistics expert (Reviewer #5), whose comments are included at the end of this letter. In light of the changes you have made and the comments made by the referees, we will require substantive revisions of the presentation of the manuscript, and additional analyses of the main effect of intention as suggested by the referees, before we can move forward with your manuscript.

In your revision, you have removed substantive sections from the Methods (previously called 'Analysis plan'); this is not permitted and it has prevented Reviewer #5 from evaluating whether your statistical analyses are sound and were conducted as preregistered (or with permissible, clearly explained deviations).
A Stage 2 Registered Report may not contain any deviation in text from the Stage 1 protocol as accepted in principle, with the exception of changes in tense (in the Introduction) and additional passages in the Methods that explain deviations from the analysis plan. You cannot remove any parts of the analysis plan. The Stage 1 protocol does not constitute a separate text that can be referred to as 'the preregistration'. Stage 1 and Stage 2 are separate versions of the same text, and all information, what you intended to do and how you eventually implemented the protocol, must be fully contained in the Stage 2 manuscript.

Once we have received a suitable manuscript, we will confer with Reviewer #5 to ascertain whether the analyses were implemented as preregistered, and whether any deviations are justified. As you implement revisions in response to the remarks made by Reviewers #1 and #2, please keep in mind that the outcome of exploratory analyses that did not form part of the Stage 1 protocol must not be discussed at length and be presented as tentative, requiring further confirmatory research.

One of the main reasons for delays in eventual acceptance is failure to fully comply with editorial policies and formatting requirements. To assist you with finalizing your manuscript for publication, I attach a checklist that lists all of our editorial policies and formatting requirements.

Please attend to *every item* in the checklist and upload a copy of the completed checklist with your submission. I have highlighted in the checklist items that require your attention. I also mention here a few points that are frequently missed and can cause delays:

1) Please insert the following Protocol Registration information in your manuscript:

The stage 1 protocol for this Registered Report was accepted in principle on 30th January 20. The protocol, as accepted by the journal, can be found at https://doi.org/10.6084/m9.figshare.11871324.v1

2) Ensure that all corresponding authors have linked their ORCID to their account on our online manuscript handling system. This is very frequently missed and invariably causes delays in formal acceptance.

3) Ensure that you provide all of the materials requested in the attached checklist and below with your final submission.

Nature Human Behaviour offers a transparent peer review option for new original research manuscripts submitted from 1st December 2019. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

We hope to hear from you within 1 month; please let us know if the revision process is likely to take longer.

To submit your revised manuscript, you will need to provide the following:
• Cover letter
• Point-by-point response to the reviewers (if applicable)
• Manuscript text (not including the figures) in .docx or .tex format
• Individual figure files (one figure per file)
• Extended Data & Supplementary Information, as instructed
• Reporting summary
• Editorial policy checklist
• Third-party rights table (if applicable)
• Suggestions for cover illustrations (if desired)

Consortia authorship:
For papers containing one or more consortia, all members of the consortium who contributed to the paper must be listed in the paper (i.e., print/online PDF). If necessary, individual authors can be listed in both the main author list and as a member of a consortium listed at the end of the paper. When submitting your revised manuscript via the online submission system, the consortium name should be entered as an author, together with the contact details of a nominated consortium representative. See https://www.nature.com/authors/policies/authorship.html for our authorship policy and https://www.nature.com/documents/nr-consortia-formatting.pdf for further consortia formatting guidelines, which should be adhered to prior to acceptance.

Forms:
Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish

your work. Once your paper is accepted, you will receive an email in approximately 10 business days providing you with a link to complete the grant of rights. If you choose to publish Open Access, our Author Services team will also be in touch at that time regarding any additional information that may be required to arrange payment for your article.

Please note that you will not receive your proofs until the publishing agreement has been received through our system.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

Please use the following link for uploading these materials:

[REDACTED]

If you have any further questions, please feel free to contact me.

With best regards,

Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

Reviewer #1:
Remarks to the Author:
This manuscript presents the results of a preregistered report testing whether patterns of moral judgment previously identified by Greene et al (2009) are shared cross-culturally. The impressive research undertaking was to test for these effects in 45 culturally diverse countries. The authors find strong evidence for cross-cultural consistency in the sensitivity of moral judgment to "personal force". They find inconclusive evidence for cross-cultural consistency in the sensitivity of moral judgment to the interaction between personal force and "intention" (i.e., the distinction between harm as an instrumental means and harm as a foreseen side-effect).

I was a reviewer of the original registration document. This manuscript closely tracks the original plan, and so I have few comments to make. I commend the authors on an impressive and informative piece of science.

My one major suggestion is that the authors add some statistical analysis of the "main effect" of intention—i.e., not its interaction with personal force, but rather it's simple, unmoderated effect. Inspecting the data plots, it looks to me as if even in cases where there is no interaction between intention and personal force, there is evidence for a main effect of intention. If that is correct, I think this finding would be of significant interest to the field, warranting the reporting of the statistical effects even if these were not included in the original preregistration. After all, there are many studies other than Greene's that report simple effect of "intent" (in the author's sense of the term) on moral judgment, and even Greene 2009 provided some evidence that there may be an effect of intent in the absence of personal force—specifically, in the reanalysis of data originally presented by Cushman et al 2006.

I just have one more suggestion, which is minor. When the authors first report in the manuscript that they are excluding 80% of participants, it might assuage the reader's concerns to immediately address why this number is so high, and why it should not be taken as a reason to question the validity of the entire project.

Reviewer #2:
Remarks to the Author:
This manuscript, Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements, presents the results of a preregistered report studying how a large cross cultural sample of people respond to the combination of intention to harm as a direct versus side effect with personal versus impersonal force in versions of the trolley and speedboat dilemmas.

Overall, this is a solid paper and the results are largely clear and sensible. I do not see much here that changes my opinion of the original preregistered report, which was favorable. I appreciate the large sample, diverse sample, the use of Baysian analyses, etc. Overall, I think this paper should be published. I do have a few minor suggestions for the authors, however.

It seems fairly clear the interaction comes out cleanly for North America and the personal force effect cross culturally. I note that the authors spend most of the analysis focusing on these two effects, and rarely touch on the main effect of intention itself. I suggest that the main analysis, being a 2x2 should include a slightly more robust nod to this missing main effect. How well does intention by itself influence responding in the cultural clusters? Please clarify this point for the reader.

The authors briefly mention the largely null findings with individualism collectivism. I shared their surprise that this measure did not do more. For what it is worth, I recently included this same measure in a recent study (with a smaller and more limited sample) yet found the exact same pattern—largely null effects with a modest effect of increased sacrificial acceptability for only vertical individualism. No changes requested on this point but I thought the authors might take some small comfort in noting they are not the only lab to produce this result.

The authors also included the OUS scale in their study. I understand they relegated this analysis to the supplement to save space and whatnot, but I wondered whether these results might warrant a brief nod in the main text to at least give the reader a hint and motivate those interested to seek out the supplement. Moreover, the supplement only provides the OUS scale means for the three country clusters. What I, and perhaps other readers, really want to know is how each subscale correlated with overall utilitarian responding, perhaps within each country cluster. Kahane and colleagues (e.g., 2018) argue that IH shows a higher correlation than IB with utilitarian responding; others suggest a different pattern (e.g., Koerner et al., 2020). Briefly presenting these results would help speak to an important debate and hence further boost the value of this paper to the current discussion. Hopefully this would just involve adding one table to the supplement combined with a brief paragraph commenting on the results in the supplement, and an even briefly nod to this conclusion in the main text.

The authors argued they might be underpowered to detect effects for the Southern cluster. Specifically, they write, "analysis of the Southern cluster in Study 2. Specifically, if the scale of the prior 418 distribution had been r = 0.21 or higher (instead of r = 0.19), we could have concluded that 419 there was strong evidence for the effect (instead of saying that the test is inconclusive). This 420 exception might indicate that there was an effect in the Southern cluster, but the sample size 421 after exclusions was not large enough to detect this with enough certainty." This naturally led me to wonder what the sample size for this cluster was? It was unclear in the text. I hunted around, finally finding this information in Table 2, where apparently there was around 600 people in this cell. I was hoping the authors would clarify this point in the text, and expand on why they think this number would result in low power? It seems like a rather large sample? I think this point warrants a slightly expanded discussion, perhaps a clarification of the power levels the authors achieved.

When aiming to explain weaker effects for the speedboat than trolley dilemma, the authors argue that "in experiments presenting multiple moral dilemmas in a row, after the first dilemma people tend to judge subsequent moral dilemmas less morally acceptable53" citing a paper by Haidt & Baron (1996). There are a couple problems with this argument. First, that paper presents lying cases and other problems that vary considerably from the classic type of sacrificial dilemma examined here, making me question this use of evidence and the point in general. Second, this statement vastly oversimplifies the effect. The actual distribution of responses on subsequent dilemmas varies considerably depending on the identity of participants and their social goals (e.g., Uhlmann et al., 2011), and the specific combination of dilemmas to the degree that people transfer to the inherent causal model form one to the other (Wiegmann & Waldmann, 2014)—that is starting with a more emotion-laden dilemma appears to reduce utilitarian judgments on a second less emotion laden dilemma as people transfer over the emotional reaction to harm in the first dilemma. Reversing the order has less impact as the emotional rection to the less emotion-laden harm in the first dilemma fails to transfer over to the second. Note that both these findings and theory directly contradict the arguments the authors make. I suggest revising this section bearing in mind this possibility, and removing or toning down the simplistic argument that subsequent dilemmas show a blanket movement towards increased utilitarian responding.

The authors write "being familiar with the trolley problem makes people more likely to apply certain moral rules when making moral judgements (e.g., utilitarian principles, regardless of context)." I agree with the general point the authors are making about possible mechanisms that may explain the familiarity effect, but note this specific example makes no sense as it directly contradicts the pattern in the data—applying utilitarian principles across context should eliminate differences between scenarios rather than highlight them. I suggest rephrasing the example to match the data pattern lest readers get confused.

The authors mention they eschewed analyzing dichotomous data as per their pre-registration. It would be nice to clarify for the reader, most of whom will not look at the pre-registration, why they made this decision. Was it an issue of power? Did they expect different patterns for dichotomous versus continuous results? I can see arguments for these points but note that most papers using both tend to find similar effects (e.g., Bartels, 2008). It would be nice to see a clearer rationale briefly mentioned.

Finally, the authors used continuous measures essentially pitting harm rejection against outcome-maximization in something of a tug of war. Although fairly common in the field, including the Greene work the authors expand on, this approach has also been criticized for insensitivity for more complex patterns uncovered by modelling approaches such as process dissociation or the CNI model (e.g., Conway & Gawronski, 2013; Gawronski et al., 2017). For example, some people show increased tendencies to both reject harm and maximize outcomes using these models, effects which should cancel out here resulting in null effects (i.e., statistical suppression, Conway et al., 2018; Koerner et al., 2020). Could it be that some of the weak patterns in the current data may reflect such suppression, and that use of modelling might uncover more complex patterns? It might be nice to add a line or so of speculation in the discussion.

In sum, those are my suggestions. In my view, they are all relatively minor as they should require only modest adjustments to the text rather than any substantive changes. Therefore, I can congratulate the authors on a fine piece of (lots of) work and hope they find these comments helpful in refining the final version of the manuscript.

Reviewer #5:
Remarks to the Author:
You section on "Statistical Analysis", starting on
line 790, is just too superficial to be helpful.
However, I did find some of your scatter plots informative.

Author Rebuttal, first revision:

Reviewer #1:

Remarks to the Author:

This manuscript presents the results of a preregistered report testing whether patterns of moral judgment previously identified by Greene et al (2009) are shared cross-culturally. The impressive research undertaking was to test for these effects in 45 culturally diverse countries. The authors find strong evidence for cross-cultural consistency in the sensitivity of moral judgment to “personal force”. They find inconclusive evidence for cross-cultural consistency in the sensitivity of moral judgment to the interaction between personal force and “intention” (i.e., the distinction between harm as an instrumental means and harm as a foreseen side-effect).

I was a reviewer of the original registration document. This manuscript closely tracks the original plan, and so I have few comments to make. I commend the authors on an impressive and informative piece of science.

My one major suggestion is that the authors add some statistical analysis of the “main effect” of intention—i.e., not its interaction with personal force, but rather it’s simple, unmoderated effect. Inspecting the data plots, it looks to me as if even in cases where there is no interaction between intention and personal force, there is evidence for a main effect of intention. If that is correct, I think this finding would be of significant interest to the field, warranting the reporting of the statistical effects even if these were not included in the original preregistration. After all, there are many studies other than Greene’s that report simple effect of “intent” (in the author’s sense of the term) on moral judgment, and even Greene 2009 provided some evidence that there may be an effect of intent in the absence of personal force—specifically, in the reanalysis of data originally presented by Cushman et al 2006.

Thank you for your suggestion. We added an analysis testing the main effect of intention on (p. 14), see response to the Editor.

I just have one more suggestion, which is minor. When the authors first report in the manuscript that they are excluding 80% of participants, it might assuage the reader’s concerns to immediately address why this number is so high, and why it should not be taken as a reason to question the validity of the entire project.

Now, we have added the following sentence explaining that we also conducted the analysis on the full sample, without any exclusion, and found the same results (p. 11):

“We collected data from 27,502 participants out of 45 countries. Due to our exclusion criteria, we had to exclude 80.6% of the sample for the main analysis (see Table 2 for the various exclusion criteria). Note that, as we registered, we conducted the analysis without excluding the data of the participants who were familiar with the trolley problem (36.2% exclusions), and we also conducted a post-hoc explorative analysis in which we applied no exclusion criteria.”

Reviewer #2:

Remarks to the Author:

This manuscript, Moral thinking across the world: Exploring the influence of personal force and intention in moral dilemma judgements, presents the results of a preregistered report studying how a large cross cultural sample of people respond to the combination of intention to harm as a direct versus side effect with personal versus impersonal force in versions of the trolley and speedboat dilemmas.

Overall, this is a solid paper and the results are largely clear and sensible. I do not see much here that changes my opinion of the original preregistered report, which was favorable. I appreciate the large sample, diverse sample, the use of Baysian analyses, etc. Overall, I think this paper should be published. I do have a few minor suggestions for the authors, however.

It seems fairly clear the interaction comes out cleanly for North America and the personal force effect cross culturally. I note that the authors spend most of the analysis focusing on these two effects, and rarely touch on the main effect of intention itself. I suggest that the main analysis, being a 2x2 should include a slightly more robust nod to this missing main effect. How well does intention by itself influence responding in the cultural clusters? Please clarify this point for the reader.

Thank you for this suggestion. We have added an analysis testing the main effect of intention. See response to Editor.

The authors briefly mention the largely null findings with individualism collectivism. I shared their surprise that this measure did not do more. For what it is worth, I recently included this same measure in a recent study (with a smaller and more limited sample) yet found the exact same pattern—largely null effects with a modest effect of increased sacrificial acceptability for only vertical individualism. No changes requested on this point but I thought the authors might take some small comfort in noting they are not the only lab to produce this result.

Thank you for noting this, this indeed further increases our confidence in the null findings.

The authors also included the OUS scale in their study. I understand they relegated this analysis to the supplement to save space and whatnot, but I wondered whether these results might warrant a brief nod in the main text to at least give the reader a hint and motivate those interested to seek out the supplement. Moreover, the supplement only provides the OUS scale means for the three country clusters. What I, and perhaps other readers, really want to know is how each subscale correlated with overall utilitarian responding, perhaps within each country cluster. Kahane and colleagues (e.g., 2018) argue that IH shows a higher correlation than IB with utilitarian responding; others suggest a different pattern (e.g., Koerner et al., 2020). Briefly presenting these results would help speak to an important debate and hence further boost the value of this paper to the current discussion. Hopefully this would just involve adding one table to the supplement combined with a brief paragraph commenting on the results in the supplement, and an even briefly nod to this conclusion in the main text.

This would be interesting, but we believe this is beyond the scope of our current work. The reason that we delegated that part into the Supplementary Information is because it does not really connect to the main scope of the paper, and would only confuse the reader. However, for the sake of completeness, we report these correlations in the Supplementary Information and added a footnote to the main text that analysis regarding the Oxford Utilitarian Scale can be found in the Supplementary Information (shortly explaining the effect). This is the most we can do without disrupting the flow of the main text. This is what we write (p. 15):

"As we preregistered, we added a figure showing the distribution of responses of both subscales of the Oxford Utilitarianism Scale for each country claster, and also reported means and 95% confidence intervals, as registered. Moreover, we also added a post-hoc analysis correlating each subscales of the OUE with moral acceptability ratings of the moral dilemmas. We found that moral acceptability ratings correlate higher with the "instrumental harm" sub-scale (r = 0.40 - 0.45) than with the "impartial beneficence" sub-scale (r = 0.05 - 0.20) - with this latter correlation exhibiting somewhat larger cultural variations. Details can be found in the Supplementary Analysis."

The authors argued they might be underpowered to detect effects for the Southern cluster. Specifically, they write, "analysis of the Southern cluster in Study 2. Specifically, if the scale of the prior distribution had been r = 0.21 or higher (instead of r = 0.19), we could have concluded that there was strong evidence for the effect (instead of saying that the test is inconclusive). This exception might indicate that there was an effect in the Southern cluster, but the sample size after exclusions was not large enough to detect this with enough certainty." This naturally led me to wonder what the sample size for this cluster was? It was unclear in the text. I hunted around, finally finding this information in Table 2, where apparently there was around 600 people in this cell. I was hoping the authors would clarify this point in the text, and expand on why they think this number would result in low power? It seems like a rather large sample? I think this point warrants a slightly expanded discussion, perhaps a clarification of the power levels the authors achieved.

According to our preregistered sample size planning, we did not reach the designated sample size for the desired power. The sample size target for reaching 95% power was 1,800 participants per condition per cluster. Now, we have added the following clarification to the manuscript (p. 13):

"Here, we would like to stress that we did not reach our registered sample size in this cluster for Study 2, (we registered that for 95% power, we would need 1,800 participants in each cluster of which we only reached 690 - see the Methods for details on sample size estimation). This could explain why our results did not reach our evidence thresholds and remained inconclusive."

When aiming to explain weaker effects for the speedboat than trolley dilemma, the authors argue that "in experiments presenting multiple moral dilemmas in a row, after the first dilemma people tend to judge subsequent moral dilemmas less morally acceptable53" citing a paper by Haidt & Baron (1996). There are a couple problems with this argument. First, that paper presents lying cases and other problems that vary considerably from the classic type of sacrificial dilemma examined here, making me question this use of evidence and the point in general. Second, this statement vastly oversimplifies the effect. The actual distribution of responses on subsequent dilemmas varies considerably depending on the identity of participants and their social goals (e.g., Uhlmann et al., 2011), and the specific combination of dilemmas to the degree that people transfer to the inherent causal model form one to the other (Wiegmann & Waldmann, 2014)—that is starting with a more emotion-laden dilemma appears to reduce utilitarian judgments on a second less emotion laden dilemma as people transfer over the emotional reaction to harm in the first dilemma. Reversing the order has less impact as the emotional reaction to the less emotion-laden harm in the first dilemma fails to transfer over to the second. Note that both these findings and theory directly contradict the arguments the authors make. I suggest revising this section bearing in mind this possibility, and removing or toning down the simplistic argument that subsequent dilemmas show a blanket movement towards increased utilitarian responding.

You are right indeed. We removed this argument from the text. This is what we write instead (p. 15):

"Our general observation is that the size of the interaction was smaller on the speedboat dilemmas in every cultural cluster. It is yet unclear whether this effect is caused by some deep-seated (and unknown) differences between the two dilemmas (e.g., participants experiencing smaller emotional engagement in the speedboat dilemmas that changes response patterns), or is caused by some unintended experimental confound (e.g., order effect of the presented dilemmas)."

The authors write "being familiar with the trolley problem makes people more likely to apply certain moral rules when making moral judgements (e.g., utilitarian principles, regardless of context)." I agree with the general point the authors are making about possible mechanisms that may explain the familiarity effect, but note this specific example makes no sense as it directly contradicts the pattern in the data—applying utilitarian principles across context should eliminate differences between scenarios rather than highlight them. I suggest rephrasing the example to match the data pattern lest readers get confused.

Indeed. This is what we write now (p. 16):

"This increase could be due for at least two reasons: (1) familiarity is not the main reason behind the change in response patterns: familiarity correlates with an as yet unknown underlying variable, which induces a selection bias (e.g., educational background); and (2) familiarity is the main reason behind the change in response patterns: for example, being familiar with the trolley problem might have caused people to exhibit a lower emotional response to the problem or caused them to apply different reasoning that ended up affecting their responses. Our results cannot differentiate between the above described explanations (which are not necessarily mutually exclusive)."

The authors mention they eschewed analyzing dichotomous data as per their pre-registration. It would be nice to clarify for the reader, most of whom will not look at the pre-registration, why they made this decision. Was it an issue of power? Did they expect different patterns for dichotomous versus continuous results? I can see arguments for these points but note that most papers using both tend to find similar effects (e.g., Bartels, 2008). It would be nice to see a clearer rationale briefly mentioned.

The main motivation behind our decision was that the hypothesis Greene et al. put forward focused on the continuous data too (and did not report findings regarding the dichotomous outcome). Now, we have added this motivation to the Methods section.

Finally, the authors used continuous measures essentially pitting harm rejection against outcome-maximization in something of a tug of war. Although fairly common in the field, including the Greene work the authors expand on, this approach has also been criticized for insensitivity for more complex patterns uncovered by modelling approaches such as process dissociation or the CNI model (e.g., Conway & Gawronski, 2013; Gawronski et al., 2017). For example, some people show increased tendencies to both reject harm and maximize outcomes using these models, effects which should cancel out here resulting in null effects (i.e., statistical suppression, Conway et al., 2018; Koerner et al., 2020). Could it be that some of the weak patterns in the current data may reflect such suppression, and that use of modelling might uncover more complex patterns? It might be nice to add a line or so of speculation in the discussion.

It is unclear to us what "weak patterns" the reviewer is referring to that could be explained by the "statistical suppression" effect. Nevertheless, we indeed find the process dissociation method interesting and noted in the Discussion section that it could be beneficial for future research to use such methods to further understand the effects of physical force and intention on moral judgments. This is what we write (p. 17):

"Another limitation of our study might come from the fact that we used a single continuous measure of deontological/utilitarian tendences. Although common in the field, such an approach has been criticized for being overly simplistic and not being able to pick up on more complex response patterns [46,47]***. For example, maximizing outcome and rejecting harm are not necessarily symmetrical (as our***

continuous measure suggests). Hence, an interesting direction for future research could be to identify whether personal force and intention increase reliance on deontological rules or decrease reliance on consequentialist thinking. Methodological approaches, such as process dissociation, are promising in this regard[41].”

In sum, those are my suggestions. In my view, they are all relatively minor as they should require only modest adjustments to the text rather than any substantive changes. Therefore, I can congratulate the authors on a fine piece of (lots of) work and hope they find these comments helpful in refining the final version of the manuscript.

We would like to thank the reviewer for the effort and the useful comments.

Reviewer #5:

Remarks to the Author:

You section on "Statistical Analysis", starting on

line 790, is just too superficial to be helpful.

However, I did find some of your scatter plots informative.

We hope that the present update is of satisfactory standards.

Decision Letter, second revision:

Dear Dr Bago,

RE: "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgements</p>"

Thank you for submitting your revised manuscript and for all your work on the revision. Please accept my sincere apologies for the unconscionable delay in contacting you about your revision and thank you for your patience. I am writing to you instead of Dr Schiffer, as she is currently out of the office.

As we noted in our previous decision letter, a requirement for Stage 2 Registered Reports is that the Stage 1 portion remains unaltered in its entirety at Stage 2 (except of changes in tense). This means that the full Stage 1 submission needs to be incorporated word-for-word, without any changes, in the Stage 2 manuscript. Specifically, no wording can be changed (except tenses), no sections can be added or removed, no phrasing should be altered, no references added or removed, and no aspects of the analytical plan should change.

Unfortunately, the Stage 1 portion of your Stage 2 manuscript substantially deviates from the accepted protocol, including in substantive ways - for instance, the addition of the following to the analysis plan:

"Individual-level horizontal and vertical individualism-collectivism. Triandis and Gelfand45 defined individualistic and collectivistic cultural tendencies along four dimensions: vertical individualism, vertical collectivism, horizontal individualism, and horizontal collectivism. We added these into the Bayesian linear regression models. The predictive power of all four measures was assessed separately."

I have attached the output of a document comparison that highlights all changes. The Stage 1 portion of the manuscript needs to be reverted completely to be identical to the accepted protocol, which I also attach. The individualism/collectivism variables must be removed from the regression models you ran, which should be reported as they were originally planned. If you would like to add these variables to the regression models, you can do it in the results section, as an exploratory analysis (to be reported in addition to your originally planned analysis). For any deviations from the protocol, the original wording must remain and the deviation will then need to be detailed - for instance, you intended to have 146 laboratories participating, but the final count was 140. The original wording with indication of Stage 1 status ('The data collection team at Stage 1 included 146 laboratories from 52 countries.') must remain and then a note of how many laboratories finally participated should be added: 140 laboratories actually participated in data collection.

When you have addressed the comments above, please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you in advance for attending to these requests and we look forward to receiving your revised manuscript. If you have any questions regarding the implementation of these changes, please don't hesitate to contact me.

Sincerely,

Stavroula Kousta

Stavroula Kousta, PhD
Chief Editor
Nature Human Behaviour

Decision Letter, third revision:

17th November 2021

Dear Bence,

Thank you for submitting your revised manuscript "<p>Moral thinking across the world:Exploring the influence of personal force and intention in moral dilemma judgements</p>" (NATHUMBEHAV-19046333G).

It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision. Please note that Reviewer #1 initially didn't receive your replies to their comments. We resolved this later, and they confirmed that your replies addressed the issues they raised. Reviewer #5, who vetted the statistics, likewise informed us that they had no significant concerns.

In light of this feedback, we will be able to accept your Stage 2 Registered Report for publication, as soon as you have undertaken some minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour

Reviewer #1 (Remarks to the Author):

This manuscript does not seem to have changed very substantively since the last time I reviewed it. Or, in any event, the authors' "reply to reviewers" doesn't engage with any points I made last time. But my overall evaluation of this manuscript is positive and I think it should be published.

Reviewer #2 (Remarks to the Author):

I am satisfied with these changes and the authors' explanations. Although these edits do slightly move away from the original wording, I agree with the authors that they help clarify ambiguous points and refine understanding that was clouded in the original version, while remaining true to the spirit of the original. I do not have any problem moving forward with these edits.

Final Decision Letter:

Dear Bence,

We are pleased to inform you that your Registered Report "Situational factors shape moral judgments in the trolley dilemma in Eastern, Southern, and Western countries in a culturally diverse sample", has now been accepted for publication in Nature Human Behaviour.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors whose manuscript was submitted on or after January 1st, 2021, may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. IMPORTANT NOTE: Articles submitted before January 1st, 2021, are not eligible for Open Access publication. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. For submissions from January 2021, if your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route our standard licensing terms will need to be accepted, including our self-archiving policies. Those standard licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Before your manuscript is typeset, we will edit the text to ensure it is intelligible to our wide readership and conforms to house style. We look particularly carefully at the titles of all papers to ensure that they are relatively brief and understandable.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

Best wishes,

Marike

Marike Schiffer, PhD
Senior Editor
Nature Human Behaviour