Check for updates

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Reply

Psychological Science
1–5
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797619865236
www.psychologicalscience.org/PS
SAGE

Should Trolleys Be Scared of Mice? Replies to Evans and Brandt (2019); Białek, Turpin, and Fugelsang (2019); Colman, Gold, and Pulford (2019); and Plunkett and Greene (2019)

Dries H. Bostyn and Arne Roets
Department of Developmental, Personality, and Social Psychology, Ghent University

Received 3/11/19; Revision accepted 7/1/19

In our original article (Bostyn, Sevenhant, & Roets, 2018), we reported on an experiment that investigated how people behave when confronted in real life with a sacrificial, trolley-style moral dilemma (i.e., the mouse dilemma) and how this behavior relates to their judgments on hypothetical dilemmas. We concluded that participants react differently to real-life dilemmas than to hypothetical dilemmas on the basis of three findings: (a) Participants were more likely to give consequentialist responses on the real-life than on the hypothetical dilemma, (b) traditional philosophical dilemmas predicted responses on the hypothetical mouse dilemma but not on its real-life variant, and (c) individual-differences measures that predict responses to hypothetical dilemmas did not predict responses to the real-life dilemma. Our article drew multiple Commentaries. We would like to thank the authors of all these Commentaries for their valuable feedback.

Reply to Evans and Brandt (2019)

In our original article, we analyzed the data from the real-life-dilemma sample separately from the data of the hypothetical-dilemma sample. We reported that there was (a) a significant association between preference for consequentialism and participants' decisions on the hypothetical mouse dilemma and (b) a nonsignificant association between preference for consequentialism and participants' decisions on the real-life dilemma. Building on these results, Evans and Brandt correctly point out that one cannot infer a significant difference from our findings reporting a difference in significance. To do so, one would need to explicitly test whether the association is moderated by sample type. Evans and Brandt ran the required test and found that the interaction was not significant.

We consider Evans and Brandt's Commentary to be informative because it adds an important nuance to our findings and calls for more research. At the same time, we want to warn against drawing strong conclusions on the basis of this additional interaction test. We did not power the study with the aim of testing this interaction. In hindsight, we acknowledge this is perhaps a missed opportunity because, as a result, the interaction test proposed by Evans and Brandt is unfortunately dramatically underpowered. When designing the original study, our primary focus was on the real-life behavior and the association of such behavior with responses to traditional trolley dilemmas. Our study was designed to be sufficiently powered for that purpose. Indeed, a power-sensitivity test shows that our real-life study had 80% power to detect effects with a magnitude (odds ratio, or *OR*) of 1.74 (roughly comparable with an r of .15) and no less than 97% power to detect associations of the same strength as those we found in the hypothetical sample ($OR = 2.14$, $r = \sim.21$). Thus, our article provided good evidence for the lack of an association between the dilemma battery and behavior in the real-life dilemma.

In addition to the real-life-dilemma study, we presented a hypothetical version of the mouse dilemma to a second, smaller sample. This was intended as a basic

Corresponding Author:
Dries H. Bostyn, Ghent University, Department of Developmental, Personality, and Social Psychology, Henri Dunantlaan 2, B-9000, Ghent, Belgium
E-mail: dries.bostyn@ugent.be

validity check to ensure that participants' responses on the dilemma were still meaningful if traditional human targets were replaced with animals. This was confirmed by the expected, significant association between the responses on the trolley-dilemma battery and the hypothetical mouse dilemma.

So what does the lack of a significant difference in the size of the effects between the two samples mean? Evans and Brandt suggest that there are two possible interpretations. The first is to accept that our study was simply too underpowered for a moderation analysis (N = 275, which provided less than 50% power to detect a medium-sized interaction). We feel that this is the most appropriate conclusion. Drawing conclusions on the basis of underpowered tests, especially when aimed at demonstrating the absence of an effect, can be highly misleading. Imagine a pharmaceutical company testing a new cancer drug in a large trial but failing to find an effect. Imagine that the company then combines that new drug data with a small data set of people receiving another but effective therapy (e.g., radiotherapy). All data combined show an overall effect (driven by the effective radiotherapy) but because of the limited, combined sample size, there is no interaction with the type of treatment. Should the company claim that their interventions are overall effective (main effect) and that their new drug is as effective as the radiotherapy (no interaction)? We caution against such an interpretation.

Evans and Brandt stress the importance of further research to resolve the issue, and we agree on this point. Ad interim they suggest using a Bayes factor to guide expectations going forward. They report that the Bayes factor in favor of a point null interaction is 3.47. This type of Bayes factor models the increase in posterior density compared with the prior density for effects that are exactly equal to zero. This is informative, but we suggest also calculating a Bayes factor that models the increase in posterior density for all effects larger than zero. This Bayes factor is 5.11, thereby indicating that our data show that a positive interaction effect (i.e., responses to classic dilemmas are more strongly related to the hypothetical than to the real-life dilemma) is about 1.5 times more credible than a null interaction effect. We want to emphasize that neither of the two Bayes factors provide sufficiently strong evidence to resolve this issue at this point. However, to the extent that Bayes factors should inform our conclusions in anticipation of further research, we believe that the most appropriate, tentative inference would be that hypothetical moral judgment does in fact relate more strongly to hypothetical behavior than it does to real-life behavior.

In sum, Evans and Brandt's additional analysis clarifies that our original study could not provide a conclusive answer in this particular regard. We acknowledge this. We consider our study an important, but far from conclusive, step in gaining insight into the disconnect between hypothetical moral judgment and real-life moral behavior. We genuinely hope that other researchers will join our future efforts to further advance this insight by using the promising real-life research paradigm.

Reply to Białek, Turpin, and Fugelsang (2019)

Białek and colleagues focus on a different subject. They do not contest our findings but assert that these findings do not invalidate hypothetical-trolley-dilemma research. They point out that only specific subsets of people are confronted with trolley-like situations in real life. Importantly, while only some of us have to face such real-life dilemmas, we do judge people who are confronted with such decisions. Hypothetical-dilemma research may still help us to understand people as moral agents themselves and how they judge other people as moral agents.

We would argue that trolley-like situations are more common than Białek et al. suggest. For instance, each white lie is a trolley-style decision pitting active harm against a greater good. Still, we agree with the crux of their argument: Even if not predictive of behavior, there is inherent value in knowing how people think about hypothetical dilemmas, not in the least because it informs how we perceive the moral character of others (Bostyn & Roets, 2017; Everett, Pizarro, & Crockett, 2016).

Yet obviously it is also important to investigate the match (and mismatch) between hypothetical judgments and real-life behavior in moral dilemmas, as this could help inform future theorizing. For example, associations between need for cognition and consequentialist responses to hypothetical dilemmas have been used to claim that consequentialist judgment is driven by deliberate reasoning (Conway & Gawronski, 2013). However, in our study, we did not find any association between need for cognition and participants' likelihood of making consequentialist decisions in the real-life case.

While our null results seem to have drawn most attention (both in Commentary form and online), we think it is worthwhile to reiterate that participants' responses to traditional hypothetical dilemmas still predicted their reaction times, their doubt, and their uncomfortableness in the real-life dilemma situation. Our study does not invalidate research on hypothetical trolley-style dilemmas, and we never intended to proclaim the death of trolley dilemmas. However, most of the research literature has neglected the possibility that there might be meaningful differences between hypothetical judgment and real-life moral behavior. Our

article shows that willful ignorance on this issue is not tenable and that our current models of human morality that are based on hypothetical research alone might need to be expanded.

Reply to Colman, Gold, and Pulford (2019)

Colman and colleagues focus their Commentary on how our work fits with their related work, noting both similarities and differences. First, they point out that we were not the first to study real-life trolley decisions. Indeed, in some of their studies, the commentators have also tried to bring trolley-inspired moral dilemmas to the "real world," albeit through a quite different design. When designing our real-life trolley dilemma, we had two core characteristics in mind: (a) The relevant behavior should concern inflicting harm (i.e., "positive punishment"), and (b) a real-life version should involve a confrontation with real victims in real time. We think that our version of the trolley dilemma, which involved participants being directly confronted with real-life creatures that were at risk of an electric shock is unique in this regard. The work cited by Colman et al. involved computer tasks about taking away a benefit that was introduced by the experiment, such as revoking preallocated charitable contributions or monetary gains (i.e., "negative punishment"). For example, the decisions made by participants in Gold, Colman, and Pulford's (2014) study impacted the distribution of additional meals, which would not have been donated without the experiment. Regardless of participants' choices, the experiment's net effect was added meals for the children. We believe that this net gain as well as the temporal distance (multiple weeks) and spatial distance (intercontinental) between decision and outcome makes their real-life translation of the trolley dilemma considerably different from ours. Nevertheless, we acknowledge that our criteria may be debatable, and we regret not recognizing these commentators' work in our original article. In our more recent work on monetary dilemmas, we do cite their work as particularly relevant (Bostyn, Sevenhant, & Roets, 2019).

Yet rather than debating who came first, we believe that the productive way forward would be to collaboratively study how differences in the core characteristics of real-life dilemmas can affect participants' responses, either in terms of actual behavior or in terms of appropriateness judgments. Indeed, differences in these core characteristics, such as the positive punishment versus negative punishment dimension (for lack of better terminology), may explain some of the diverging results, as previous research has demonstrated that people perceive negative outcomes or losses differently from positive outcomes or gains (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Tversky & Kahneman, 1991).

The second main comment by Colman et al. pertains to the use of deception in psychological research. They consider deception as ethically suspect because it might negatively impact future studies. Such concerns go well beyond our particular study and are probably better discussed in a different context. However, it should be noted that literature on this issue is more nuanced than the commentators imply and that good debriefing practices can counteract for the potential risks of deception (Boynton, Portnoy, & Johnson, 2013; Hertwig & Ortmann, 2008). This is why we took great care to debrief each participant individually immediately after the experiment. Some important psychological questions cannot be investigated empirically without using some form of deception. This is especially the case when studying reactions to harm or danger (e.g., experimental research on helping behavior in response to an apparent emergency). Psychologists of good will may differ in their opinions as to whether a particular research project is sufficiently important to warrant the use of deception. We agree that deception should be used sparingly, but we would not advocate abolishing it altogether.

Finally, Colman et al. imply that our use of deception may have caused participants to be suspicious of the experiment itself, which may have compromised our results. However, skepticism within a specific experiment is not determined by the use of deception in that experiment but by the believability of its design. Had we actually shocked the mice when the timer ran out, this would not have affected whether participants were skeptical at the moment they had to make their decision. Therefore, skepticism toward a specific study design is independent from the use of deception in that design. Crucially, however, research using deception typically includes explicit tests to determine whether skepticism may have affected the results, whereas studies that do not use deception but have designs that participants could consider far-fetched usually do not. Colman et al. imply that we assessed only the believability of our design through participants' uncomfortableness ratings, but this is inaccurate. We also asked participants to rate how skeptical they were, and we explicitly tested whether any of our results were moderated by skepticism or whether our results changed when we excluded our most skeptical participants. They did not. Nevertheless, even if deception does not adversely affect the study in which it is employed, we acknowledge that it might still lead to participants becoming more distrustful in future research. This and other potential costs should be compared with the benefits of each research project as part of the ethical review process.

Reply to Plunkett and Greene (2019)

The Commentary by Plunkett and Greene builds on our original article in two ways: First, they provide an alternative approach to one particular analysis, and this approach is based on an alternative measure. Second, they advance a more general critique of the practice of using responses to trolley dilemmas to explore individual differences.

In the trolley-dilemma battery in our study, we used the standard response format mentioned by the commentators (i.e., evaluate the consequentialist option) and additionally asked participants to evaluate the deontological option. In our regression models, we simultaneously incorporated both measures. Plunkett and Greene suggest combining both measures into a single difference measure, and they report a marginally significant association between this relative judgment measure and behavior in the real-life dilemma (significant with a one-tailed test).

We have several concerns with this alternative measure. According to the dual-process model, deontological and utilitarian processes are independent, and therefore, it is debatable whether combining measures of independent constructs into a single difference score is warranted from a theoretical perspective. Moreover, this approach is psychometrically muddled: It assumes that both measures are on the same ordinal scale, that they have equal variance, and that they have an equally strong influence on the outcome. None of these assumptions were tested by Plunkett and Greene, and at least one of them is verifiably false, that is, Levene's test of equal variance: $t(290) = -4.64$, $p < .001$. Furthermore, every difference measure reduces the richness of data and discards information. The difference measure does not differentiate participants who scored high on both moral preferences from those who scored low on both preferences, as these participants would all score around zero on the difference measure. Additionally, the difference-score approach is also mathematically superfluous and does not explain more variance on any of our outcomes. In fact, by including both measures in our analyses, we already accounted for an additive relationship between both predictors. Plunkett and Greene note that, similar to the results of their difference-score analysis, the deontological measure alone is also marginally significantly related to decisions in the real-life dilemma, and they suggest that this, too, adds nuance to our null conclusion. However, unlike the consequentialist-preference measure, this secondary measure failed to show a significant relationship with responses to the hypothetical dilemma or with responses to the mouse dilemmas in general when the data of both samples were combined (both ps > .196). We caution against interpreting marginally significant effects and even more so when they are based on flawed measures (i.e., the difference score) or when they do not show relationships with other relevant measures (i.e., the secondary deontological measure).

In sum, for all measure variations, responses to traditional dilemmas do not seem to be a meaningful predictor for real-life behavior. We can debate about whether there was truly no effect or rather a small nonsignificant effect in our study, but this hardly changes our conclusion that "hypothetical-dilemma research, while valuable for understanding moral cognition, has little predictive value for actual behavior" (Bostyn et al., 2018, p. 1084). Of course, further research is needed to corroborate the conclusions of our study, and high-powered, preregistered replication studies by other labs would be especially welcome.

As a second and more general critique, the commentators state that our article is based on "a widespread misunderstanding of what trolley-type dilemmas are supposed to do" (Plunkett & Greene, 2019, p. xxx). They argue that trolley research is aimed at exploring processes within people but not necessarily differences between people. For instance, multiple studies attempt to explain why people react differently to switch dilemmas compared with footbridge dilemmas.

The goal of our study was to investigate whether people respond differently to hypothetical rather than real-life dilemmas. This is similar both in approach and in underlying philosophy to contrasting responses to switch and footbridge dilemmas. Indeed, our first main analysis contrasted responses in the hypothetical mouse dilemma with those in the real-life mouse dilemma. Additionally, our experiment also investigated differences between people, but there is obvious value in this. Studying differences in moral cognition that exist between people is relevant and can also help us to understand the processes that explain differences within people. Plunkett and Greene conclude their Commentary by highlighting several studies that have shown meaningful individual-differences effects. Thus, it seems reasonable to conclude that trolley dilemmas can be valuable to the study of moral processes both within and between people, including the opportunity to cross-validate some effects. Rather than implying that one type of research is misguided, we think there is value in both approaches. We suggest letting empirical results guide us as to which contexts trolley dilemmas are informative in, but we agree that trolley dilemmas can be valuable even if they do not predict real-life behavior.

Conclusion

Trolley-style dilemmas have been instrumental in advancing research on moral decision making. At least

in our opinion, they will continue to be an important tool for studying moral cognition. However, our original study shows that there can be meaningful differences between hypothetical-dilemma judgment and real-life moral behavior. Rather than considering this a flaw of the paradigm, we see it as an opportunity for the domain to explore new questions: Do different processes affect how we think about moral dilemmas compared with how we act in such dilemmas? Is the disconnect between hypothetical judgment and real-life moral behavior dependent on whether the behavior involves inflicting harm versus taking away a benefit? Do we judge other people on the basis of an assessment of the morality of the behavior itself or, rather, on what we believe this tells us about their underlying moral character? Our article does not detract from previous work but encourages new avenues for research beyond the hypothetical.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

D. H. Bostyn drafted the manuscript. A. Roets provided critical revisions. Both authors approved the final version of the manuscript for submission.

ORCID iD


Dries H. Bostyn https://orcid.org/0000-0001-9994-4615

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323–370.

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, XXX–XXX.

Bostyn, D. H., & Roets, A. (2017). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General*, *146*(5), e1–e7.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science*, *29*, 1084–1093. doi:10.1177/0956797617752640

Bostyn, D. H., Sevenhant, S., & Roets, A. (2019). Beyond physical harm: How preference for consequentialism and primary psychopathy relate to decisions on a monetary trolley dilemma. *Thinking & Reasoning*, *25*, 192–206.

Boynton, M. H., Portnoy, D. B., & Johnson, B. T. (2013). Exploring the ethics and psychological impact of deception in psychological research. *IRB: Ethics & Human Research*, *35*(2), 7–13.

Colman, A. M., Gold, N., & Pulford, B. D. (2019). Comparing hypothetical and real-life trolley problems: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, XXX–XXX.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology*, *104*, 216–235.

Evans, A. M., & Brandt, M. J. (2019). Comparing the effects of hypothetical moral preferences on real-life versus hypothetical behavior: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, XXX–XXX.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*, 772–787.

Gold, N., Colman, A., & Pulford, B. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, *9*, 65–76.

Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior*, *18*, 59–92.

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science*, *30*, XXX–XXX.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*, 1039–1061.