

**Overlooked Evidence and a Misunderstanding of What Trolley Dilemmas Do Best:
Commentary on Bostyn, Sevenhant, & Roets (2018)**

Dillon Plunkett

Joshua D. Greene

Department of Psychology, Center for Brain Science

Harvard University

Department of Psychology, 33 Kirkland St., Cambridge, MA 02138.

jgreene@wjh.harvard.edu

Words excluding references: 999

Number of references: 20

Abstract

Bostyn, Sevenhant, & Roets (2018) had participants consider hypothetical trolley-type dilemmas and then make a real decision about harming one mouse to prevent harm to five mice. They report that participants' judgments about the hypothetical dilemmas do not predict their mouse-shocking decisions. However, a reanalysis of their data using a more appropriate method for predicting real decisions, along with other results not emphasized in their paper, casts doubt on the null hypothesis asserted by Bostyn and colleagues. More generally, this research—like other recent trolley criticism—reflects a misunderstanding of what is most valuable about these dilemmas. These dilemmas are best understood as high-contrast cognitive probes (akin to flashing checkerboards), useful for illuminating cognitive structure *within* people. They can also be used as individual-difference measures, aimed at explaining variance *between* people, but their original and most reliable function is not as moral personality tests or predictors of individual behavior.

Recently, Bostyn, Sevenhant, & Roets (2018) assessed the real-world predictive power of hypothetical trolley-type dilemmas. Participants responded to such dilemmas and then made a real decision about harming one mouse versus five mice. The authors report that the trolley-type judgments don't predict the real decisions. We regard their research as valuable and endorse their most general conclusion: Studying hypothetical judgments cannot replace studying real decisions. However, a closer look at their data casts doubt on their central claim. Moreover, their research strategy reflects a common misunderstanding of what makes trolley dilemmas most useful.

Bostyn et al.'s hypothetical dilemmas employed a non-standard response format. In nearly all research using trolley-type dilemmas, participants evaluate only the proposed utilitarian action (e.g., pushing the man off the footbridge to save five lives) and do not separately assess the deontological alternative (e.g., not pushing). Because participants give only a single judgment, their responses are inherently *comparative*, accounting for both "horns" of the dilemma. Bostyn et al., however, had subjects separately evaluate the utilitarian and deontological options. Having done this, the most natural approach would have been to calculate a difference score to model the *relative appeal* of the two options according to each participant. This reflects the logic of their experiment, which was aimed at predicting a real choice between a utilitarian option and a deontological option. A logistic regression using difference scores reveals marginally significant evidence that hypothetical judgments predict real judgments ($OR:1.56, z=1.77, p=.077$; a significant effect with a one-tailed test ($p=.038$) based on a clear directional prediction; with age and gender controls as in original: $OR:1.62, z=1.86, p=.063, p=.031$ one-tailed).

Bostyn et al. take a different approach. They include both measures separately in their regression and report that evaluations of the hypothetical utilitarian options do

not significantly predict the mouse-shocking decisions ($p=.41$). However, they mention only in an endnote that participants' evaluations of the hypothetical deontological options are marginally significant predictors of mouse-shocking ($z=-1.75$, $p=.081$, $p=.040$, one-tailed). Participants' evaluations of the hypothetical utilitarian and deontological options are equally relevant predictors in asking whether hypothetical judgments predict real judgments. The results described above (marginal or not) are inconsistent with claiming strong evidence for the null hypothesis. Repeating Bostyn et al.'s Bayesian analysis with the difference score (scaled, as per Gelman et al., 2008) yields $BF_{10}=0.95$ with a 95% credible interval for the regression coefficient of 0.00-1.89, indicating no evidence in favor of the null. Thus, while their data provide no strong evidence that hypothetical trolley judgments predict real mouse-shocking decisions, their data also provide no evidence for the null hypothesis asserted by Bostyn et al.

Our broader concern, however, is with a widespread misunderstanding of what trolley-type dilemmas are supposed to do. What's most interesting about trolley dilemmas is the *contrast* between cases (Thomson, 1985): In the *switch* case, people reliably approve of hitting a switch that will turn a trolley away from five and onto one. In the *footbridge* case, people reliably disapprove of pushing one person off of a footbridge in order to save five people. Why such different answers? And what does this say about our moral thinking?

The dual-process theory (Greene, 2001; Shenhav & Greene, 2014) provides an answer: In response to both cases, people engage in simple, cost-benefit reasoning favoring action. But in the *footbridge* case, the harmful action is more emotionally salient, generating a competing response that makes most people disapprove (or approve reluctantly; cf. Bostyn et al.'s significant results for "doubt" and response times). This theory has received strong support from studies using manipulations targeting specific

processes (e.g., Crockett et al., 2010; Shenhav & Greene, 2014; Tremoliere, De Neys, & Bonnefon, 2012), and studies of clinical populations with process-specific deficits, including patients with VMPFC and hippocampal lesions (Ciaramelli et al., 2007; Koenigs et al., 2007; McCormick et al., 2016), psychopathy (Koenigs et al., 2012), and FTD (Mendez, Anderson, & Shapira, 2005).

Critically, these studies focus on dissociating *processes* that exist *within* healthy people. This explains why people are so puzzled when they first confront the *switch* and *footbridge* cases together. Recently, however, some researchers have assumed that trolley-type dilemmas, in order to be useful, must make reliable predictions about differences *between people*, either as moral personality tests (Bartels & Pizarro, 2011; Kahane, 2015; Kahane et al., 2015; Kahane et al. 2017) or as laboratory surrogates for real-world decisions (Bauman et al., 2014; Kahane et al., 2015). This reflects a misunderstanding of what trolley dilemmas do best and what the dual-process theory is trying to explain—akin to criticizing the Müller-Lyer illusion for failing to predict people’s visual acuity.

But shouldn’t trolley dilemmas tell us *something* about real-world behavior? They should, and they do—indirectly. Psychopaths and various lesion patients have *real-world* moral deficits, and they respond to trolley dilemmas in ways that are precisely predicted by the dual-process theory, with affective deficits leading to more utilitarian judgment in *footbridge*-like cases (Bartels & Pizarro, 2011; Ciaramelli et al., 2007; Koenigs et al., 2007, 2012; Mendez et al., 2005). Likewise, a recent lesion-based network analysis credits the dual-process theory with explaining patterns in damage leading to criminal behavior (Darby et al. 2017). And contra Kahane et al. (2015), utilitarian judgments also reflect pro-social motivations in healthy people. See Conway, Goldstein-Greenwood, Polacek, & Greene (2018).

Trolley-type dilemmas are best understood as high-contrast cognitive probes (like flashing checkerboards) that can dissociate processes *within* people, not as moral personality tests or surrogates for real-world emergencies. They can serve as individual-difference measures, especially with process dissociation (Conway & Gawronski, 2013; Conway et al., 2018), and there is some evidence (in addition to that presented above) that trolley dilemmas can predict real individual behavior (Dickinson & Masclet, 2018). But even if there were no individual variation to explain—for example if everyone said “yes” to *switch*-type cases and “no” to *footbridge*-type cases—trolley dilemmas would retain their original interest and purpose. Their greatest value lies not in their ability to explain our moral differences, but in their ability to reveal the fault lines running through our shared capacity for moral cognition.

References

- Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154-161.
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, *8*, 536-554.
- 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.
- Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*, 84-92.
- 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.
- 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.
- Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, *107*, 17433-17438.
- Dickinson, D. L. & Maslet, D. (2018). Using ethical dilemmas to predict antisocial choices with real payoff consequences: An experimental study. IZA Discussion Paper No. 11592. Available at SSRN: <https://ssrn.com/abstract=3205879>

- Darby, R. R., Horn, A., Cushman, F., & Fox, M. D. (2017). Lesion network localization of criminal behavior. *Proceedings of the National Academy of Sciences*, *115*, 601-606.
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*, 1360–1383.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108.
- Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*, 551-560.
- Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., et al. (2017). Beyond sacrificial harm: A Two-dimensional model of utilitarian psychology. *Psychological Review*, *152*, 131-164.
- Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu, J. (2015). ‘Utilitarian’ judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209.
- Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*, 708-714.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908–911. <http://dx.doi.org/10.1038/nature05631>.
- McCormick, C., Rosenthal, C. R., Miller, T. D., & Maguire, E. A. (2016). Hippocampal damage increases deontological responses during moral decision making. *Journal of Neuroscience*, *36*, 12157-12167.

- Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, 18, 193–197.
- Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience*, 34, 4741-4749.
- Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, 94, 1395-1415.
- Trémolière, B., Neys, W. D., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, 124, 379–384.
<https://doi.org/10.1016/j.cognition.2012.05.011>.