Research Article

Psychological Science
2018, Vol. 29(7) 1084–1093

DOI: 10.1177/0956797617752640
www.psychologicalscience.org/PS

Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas

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

Abstract

Scholars have been using hypothetical dilemmas to investigate moral decision making for decades. However, whether people's responses to these dilemmas truly reflect the decisions they would make in real life is unclear. In the current study, participants had to make the real-life decision to administer an electroshock (that they did not know was bogus) to a single mouse or allow five other mice to receive the shock. Our results indicate that responses to hypothetical dilemmas are not predictive of real-life dilemma behavior, but they are predictive of affective and cognitive aspects of the real-life decision. Furthermore, participants were twice as likely to refrain from shocking the single mouse when confronted with a hypothetical versus the real version of the dilemma. We argue that hypothetical-dilemma research, while valuable for understanding moral cognition, has little predictive value for actual behavior and that future studies should investigate actual moral behavior along with the hypothetical scenarios dominating the field.

Keywords

morality, utilitarianism, trolley, consequentialism, open data, open materials

Received 5/5/17; Revision accepted 12/11/17

Famed skeptic David Hume once bemoaned that "Nothing is more dangerous to reason than the flights of the imagination" (Hume, 1739/2003, p. 191). Hume warned his fellow philosophers of the ease with which our mental faculties are led astray by our propensity for fantasy, and accordingly, he cautioned against the use of thought experiments to strategically probe our minds. Originally intended for philosophers, Hume's admonition is appreciated by many modern-day psychologists as well. Nevertheless, entire research paradigms within psychology are built on the use of imaginative hypothetical scenarios as the primary technique to distill psychological fact from carefully controlled reveries (Cushman & Greene, 2012).

In the field of moral psychology, sacrificial "trolley-style" dilemmas have become the de facto means of scientific enquiry. These dilemmas are aimed at investigating the tension between consequentialist (utilitarian) and deontological normative ethics. In their archetypal formulation, these dilemmas require participants to imagine a runaway trolley train on a deadly collision course with a group of unsuspecting victims. Participants are asked whether they would consider it morally appropriate to save the group but sacrifice a single innocent bystander by pulling a lever to divert the trolley to another track, where it would kill only the single bystander. Consequentialists argue that one should focus on the two outcomes in this scenario and that the action that results in the least amount of harm is preferable (Rosen, 2005). Deontologists, on the other hand, argue that it is immoral to enact harm because of situational happenstance (Kant, 1785/2002). Many researchers assume that subjects' responses to these philosophical dilemmas are reflective of their ethical commitments. Some even suggest that it corresponds to a genuine fault line within our moral minds that is associated with two different modes of information processing that drive moral behavior (Greene, 2007; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

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

Though this paradigm has spawned an impressive amount of research, it is still an open research question whether subjects' hypothetical moral judgments are predictive of the actual behavior they would display in a dilemma-like situation in real life. Some researchers have tried to deal with this issue by actively measuring to what extent subjects were able to suspend their disbelief (Greene et al., 2009). Others have attempted to bridge this explanatory gap by using virtual reality paradigms that allow for a more vivid enactment of the trolley-style scenarios, assuming participants' judgments in virtual reality would more closely mirror their real-life behavior because of the increased contextual salience of the dilemma presentation (McDonald, Defever, & Navarrete, 2017; Moretto, Làdavas, Mattioli, & di Pellegrino, 2010; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014). These studies have yielded some interesting findings, including the observation that subjects may be more consequentialist when they are responding to the virtual trolley-style dilemmas compared with the hypothetical versions (Francis et al., 2016). Yet no touchstone research is available to corroborate whether this judgment–behavior discrepancy is indeed caused because virtual reality research is more lifelike.

Until recently, this judgment–behavior discrepancy has been an academic concern plaguing only moral psychologists. However, trolley-dilemma-like situations are becoming increasingly relevant to model the moral decisions of artificial intelligence, such as self-driving autonomous vehicles (Bonnefon, Shariff, & Rahwan, 2016). Accordingly, whether or not hypothetical moral judgment is related to real-life behavior is prone to become a matter of public interest. We are aware of one study that has directly compared hypothetical moral judgment with real-life behavior: FeldmanHall et al. (2012) found that people are more willing to harm others for monetary profit in a real-life scenario than they are in a hypothetical version of the same scenario, thus confirming that real-life behavior can differ dramatically from hypothetical judgment. The current research was a first attempt to study this difference in the trolley-dilemma context through the admission of a "real-life" dilemma that required participants to make a trolley-dilemma-like decision between either allowing a very painful electroshock to be administered to five mice or choosing to deliver the entire shock to a single mouse.

Method

Participants and procedure

We collected two samples from the same student population. Participants completed the experiment in return for course credit. A first group of students completed the real-life version of the mouse dilemma (described below), whereas a second group completed a hypothetical version of the same dilemma to serve as a reference. The total size of our subject pool was limited to approximately 300 students. Given that our primary focus was on the behavior in the real-life dilemma, we wanted to ensure that we had sufficient power to detect small effects on the rate of consequentialist versus deontological judgment in this group, and we tuned our sample size accordingly. We calculated that a sample of 200 participants would have enough power to detect a small effect, $OR = 1.68$ (equivalent to a Cohen's d of 0.20; Chen, Cohen, & Chen, 2010), assuming that the distribution of the consequentialist versus deontological decisions in our real-life dilemma would not be extremely unbalanced. In particular, a sample of 200 would have 75% to 95% power to detect small effects at incidence ratios from 50:50 (equal distribution of the alternatives) to 85:15 (strongly unequal distribution). Accordingly, we aimed for a sample of about 200 participants to complete the real-life version of a trolley-style dilemma and planned for the remaining students to participate in the reference group. Whereas this approach entailed having a different sample size for each group in our experiment, we considered an adequate sample size especially crucial for the real-life group.

In the real-life dilemma, a total of 208 students (we exceeded our goal because we anticipated some dropout) completed an online questionnaire containing a moral-dilemma battery to measure their preference for consequentialist moral reasoning on hypothetical dilemmas and, as a secondary measure, their preference for deontological moral reasoning on hypothetical dilemmas. Next, they completed measures for various individual-differences variables known to be positively associated with an increased likelihood of consequentialist decision making on hypothetical dilemmas (i.e., need for cognition and primary psychopathy), measures that are typically associated with a decreased likelihood of consequentialist decision making (i.e., empathic concern, perspective taking, and moral identity), and a measure for animal empathy as a control (see Conway & Gawronski, 2013; Kahane, Everett, Earp, Farias, & Savulescu, 2015). One to two weeks after completing the online questionnaire, each participant was invited to the lab for an individual session in which the real-life dilemma was administered. Though all 208 participants were invited for this individual session, 15 participants failed to follow through. Additionally, 1 participant opted out once the experimental setup was revealed. This student still received full course credit and was debriefed. Therefore, a total of 192 participants completed the real-life dilemma. The remaining 83 students in the subject pool completed the moral battery and,

subsequently, a hypothetical version of the same real-life dilemma. These participants did not complete the individual-differences measures.

Ethical approval

Given that a key aspect of trolley-dilemma-like dilemmas is the idea of serious harm inflicted on innocent, involuntary "victims," a real-life experimental version of such dilemmas cannot involve human victims. Indeed, if humans were to act as victims in a trolley-style dilemma, informed consent would obviously be required, and such use of "voluntary victims" would distort a basic premise of the trolley dilemma. Therefore we used animals (mice) for the present study. As a first step, an ethical approval procedure was started at the Ghent University Research Ethics Board for Animal Testing. During the initial contact with the board, it became apparent that the proposed experiment did not, in fact, qualify as an animal test per European Union or Belgian regulations; therefore, no formal approval was required under animal-testing regulations. In particular, animals were merely present during the experiment but were otherwise left alone and not harmed in any way (as we will clarify later, they did not actually receive any electroshocks). Although no formal approval for animal testing was required, we did follow all ethical guidelines for animal care that apply at our university.

Secondly, because the current experiment used misdirection and would most likely be stressful for our participants, who had to make this real-life moral decision, a second ethical approval application was submitted to the research ethics board at the psychology department. The current study was approved on November 7, 2016 (REB approval: 2016/86/Dries Bostyn).

Measures

Moral judgment. To measure participants' preference for consequentialist moral judgment, we required participants to respond to a moral-dilemma battery consisting of 10 hypothetical trolley-style dilemmas (Bostyn & Roets, 2017, adapted from Greene et al., 2001).[1] On each dilemma, participants were required to indicate to what extent they judged the consequentialist option to be morally appropriate using a 5-point scale ranging from 1 (*absolutely inappropriate*) to 5 (*absolutely appropriate*). As a secondary measure, participants were also asked to indicate to what extent they found the deontological option to be morally appropriate, using the same 5-point scale.

The most prominent theoretical model for moral cognition in the context of trolley-style dilemmas posits that preference for consequentialist and deontological reasoning are independent constructs and that each is determined by a different mental process (Greene, 2007). However, it is worth emphasizing that most research in the field has focused on studying preference for consequentialist reasoning and that much less is known about the drivers of deontological preference (for an exception, see the literature on process dissociation in the context of moral cognition, e.g., Conway & Gawronski, 2013). Therefore our main analyses focused on the consequentialism measure, and the study included a measure for deontological reasoning mainly for exploratory purposes. All moral dilemmas used in the current study are available at the Open Science Framework (https://osf.io/kvb99/).

Need for cognition. Need for cognition refers to participants' desire for effortful cognitive activity and was measured through the Need for Cognition questionnaire (Petty, Cacioppo, & Kao, 1984). Participants were asked to respond to 18 items and rate to what extent each item was characteristic of them on a 5-point scale ranging from 1 (*extremely uncharacteristic*) to 5 (*extremely characteristic*). An example item read, "I prefer complex to simple problems."

Empathic concern. Participants' concern for other people was measured using the Empathic Concern subscale of the Interpersonal Reactivity Index (Davis, 1983). Participants were asked to rate seven items on a 5-point scale ranging from 1 (*does not describe me well*) to 5 (*describes me very well*). An example item read, "I often have tender, concerned feelings for people less fortunate than me."

Perspective taking. Participants' ability and desire to know what other people are feeling was measured by the Perspective Taking subscale of the Interpersonal Reactivity Index (Davis, 1983). Perspective taking was measured using seven items rated on the same 5-point scale as for empathic concern. An example item read, "I try to look at everybody's side of a disagreement before I make a decision."

Primary psychopathy. As a measure of participants' antisocial tendencies, the Primary Psychopathy scale (Levenson, Kiehl, & Fitzpatrick, 1995) was administered. This measure consists of 16 statements, and participants rated each statement on a 4-point scale ranging from 1 (*disagree strongly*) to 4 (*agree strongly*). An example item read, "Success is based on survival of the fittest; I am not concerned about the losers."

Moral identity. The extent to which participants find moral ideals, traits, and actions important was measured through the Moral Identity Scale (Aquino & Reed, 2002),

which consists of an internalization and a symbolization subscale. Participants were presented with nine moral terms (e.g., generous, helpful, honest) and asked to visualize the kind of person who has these traits. Participants then indicated how well each of five internalization statements (e.g., "It would make me feel good to be a person who has these characteristics") and five symbolization statements (e.g., "The types of things I do in my spare time e.g., hobbies clearly identify me as having these characteristics") described themselves. Participants rated these statements on a 7-point scale ranging from 1 (*not true of me*) to 7 (*completely true of me*).

Animal empathy. Participants' empathy toward animals was measured by the Animal Empathy Scale (Paul, 2000). Participants were asked to indicate to what extent they agreed with 22 statements on a 5-point scale ranging from 1 (*completely disagree*) to 5 (*completely agree*). A reverse-scored example item read, "So long as they're warm and well fed, I don't think zoo animals mind being kept in cages." Additionally, we added three items to measure participants' empathy toward mice specifically, for example, "I don't like to see mice getting hurt."

Mouse dilemma.

The real-life version. All participants were invited to the lab in individual sessions. Before participants entered the lab, an experimenter read a briefing about the general nature of the experiment to the participant. Each participant was informed that he or she would be required to make a real-life ethical decision. Because electroshocks were used and we assumed that most participants would be familiar with the Milgram studies, the briefing also included a one-sentence statement that the experiment was not about obedience and that they should feel free to make whichever decision they felt was most appropriate. We further told all participants that they could quit the study at any point and would still receive full credit for their participation.

Once a participant entered the lab, we explained the full setup of the experiment to him or her. Inside, the participant saw an electroshock machine (DS5, Digitimer, Hertfordshire, England) that was hooked up to two cages with metal netting on the bottom and the sides. One cage contained a single mouse; the other contained five mice. A laptop displaying a 20-s timer was connected to the electroshock machine. The experimenter explained to the participant that a very painful but nonlethal electroshock would be applied to the cage containing the five mice when the timer reached 0 s and that he or she could choose to intervene by redirecting the electrical current to the cage containing the single mouse by pressing the button in front of him or her. The experimenter started the timer immediately after explaining the setup and remained present in the background during the run of the experiment but did not interact any further with the participant. If a participant decided to press the button, a response time was recorded. A visual depiction of the experimental setup and a translated version of the text read to the participants is available on the Open Science Framework.

When the timer reached 0, no electroshocks were administered. Instead, the experiment ended, and the participant was ushered into a different room for an immediate debriefing with another experimenter, during which the aim of the study was explained and the participant was reassured that no shocks had been administered. During this debriefing, all participants were asked to explain the motivation behind their decision, to rate to what extent they had doubted their decision and to what extent they had felt uncomfortable making this decision (separately, but on the same 7-point scale), and finally, how sure they had been that no shocks would be administered, on a scale from 0% (*absolute certainty that shocks would be administered*) to 100% (*absolute certainty that no shocks would be administered*).

The hypothetical version. The participants from the reference sample were presented with a hypothetical version of the mouse dilemma after completing the initial moral-dilemma battery. This dilemma read as follows:

> Imagine the following situation. You are participating in an experiment as part of a course in Social Psychology. Previously, you were asked to respond to several moral dilemmas, much like the ones you have answered. You are guided to the lab, the door opens and you see two cages with mice: one cage containing a single mouse, one cage containing five mice. An electroshock is hooked up to both cages. The experimenter tells you that after a 20 second timer, an electrical shock will be administered to the cage with the five mice but that you can push a button to redirect this shock to the cage containing the single mouse. The shocks are very painful but nonlethal. Would you press the button?

Results

The scripts needed to replicate all analyses, along with the data, are available on the Open Science Framework.

Preliminary analysis

As an initial exploration of our data, we conducted a reliability and correlational analysis of the moral-preference and individual-differences measures for the real-life sample. As Table 1 demonstrates, reliabilities

Table 1. Reliabilities and Correlation Matrix for All Self-Report Measures Included in the Real-Life-Dilemma Sample

		Correlations								
Variable	Cronbach's α	1	2	3	4	5	6	7	8	9
1. Preference for consequentialism	.86	—								
2. Preference for deontology	.81	−.12† [−.25, .01]	—							
3. Need for cognition	.85	.23*** [.09, .35]	−.04 [−.18, .09]	—						
4. Empathic concern	.84	−.14* [−.27, −.00]	−.12† [−.25, .02]	.06 [−.08, .19]	—					
5. Perspective taking	.80	.03 [−.10, .17]	−.08 [−.21, .06]	.29*** [.16, .41]	.50*** [.39, .60]	—				
6. Primary psychopathy	.83	.24*** [.11, .36]	.03 [−.10, .17]	−.07 [−.20, .07]	−.57*** [−.66, −.47]	−.33*** [−.45, −.21]	—			
7. Internalization	.80	−.13† [−.26, .01]	.04 [−.10, .17]	.10 [−.04, .23]	.52*** [.42, .62]	.32*** [.20, .44]	−.47*** [−.57, −.36]	—		
8. Symbolization	.73	.06 [−.08, .19]	.09 [−.04, .23]	.08 [−.06, .21]	.13† [−.00, .26]	.04 [−.10, .17]	−.08 [−.22, .05]	.17* [.04, .30]	—	
9. Animal empathy	.85	−.03 [−.16, .11]	−.07 [−.21, .06]	.09 [−.05, .22]	.41*** [.29, .52]	.27*** [.14, .39]	−.40*** [−.51, −.28]	.35*** [.22, .46]	.11 [−.03, .24]	—
10. Empathy for mice	.71	−.14* [−.27, −.00]	−.05 [−.18, .09]	.04 [−.10, .17]	.37*** [.24, .48]	.21** [.08, .34]	−.31*** [−.43, −.18]	.28*** [.15, .40]	.02 [−.12, .15]	.55*** [.44, .63]

Note: Correlations are for the real-life sample only ($n = 208$). Cronbach's αs for preference for consequentialism and preference for deontology were calculated for the combined samples ($N = 292$); all other αs are for the real-life sample only. Values in brackets are 95% confidence intervals.
†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

were good, and most correlations were as expected: For instance, empathic concern was strongly positively related to perspective taking, $r = .50$, $p < .001$, and strongly negatively related to primary psychopathy, $r = -.57$, $p < .001$.

Preference for consequentialist reasoning displayed medium-sized positive associations with need for cognition, $r = .23$, and primary psychopathy, $r = .24$, and a small negative association with empathic concern ($r = -.14$), corroborating findings from earlier studies (Conway & Gawronski, 2013; Kahane et al., 2015). Our secondary measure, preference for deontological reasoning, was not significantly related to any of the individual-differences measures.

We also wanted to check whether our participants were not overly skeptical of our study design. Crucially, only 12 out of 198 participants indicated during the debriefing that they were 100% certain that no shocks would be delivered ($M = 55.1\%$ certain). Yet of these skeptics, only 4 claimed that they did not feel uncomfortable making a decision on the mouse dilemma, suggesting that even the skeptical participants could not comfortably assume that no shocks would be given. In fact, most participants felt very uncomfortable ($M = 5.34$ on a 7-point scale), with 59 participants giving the maximum rating. We argue that this suggests that most, if not all, participants found our setup convincing. Importantly, analyzing the data with or without these skeptical participants yielded the same qualitative results, and including participants' levels of skepticism as a moderator did not moderate any of our results. The results we report in the current article are based on the full sample and do not include skepticism as a moderator, though these additional analyses can be found in the Supplemental Material available online.

Main analyses

All reported analyses were controlled for participants' age and gender. Analyzing the data without these controls yielded the same results. We first compared the proportion of deontological versus consequentialist decisions on the hypothetical mouse-dilemma with those on the real-life version. Accordingly, we fitted a logistic regression model with the type of choice (consequentialist or deontological) as the dependent variable and type of dilemma (hypothetical or real life) as a predictor variable. This analysis demonstrated that participants were more than twice as likely to make a deontological decision (vs. a consequentialist one) when faced with the hypothetical dilemma (34% of decisions) than they were when faced with the real-life version (16% of decisions), a difference that was statistically significant, $z = 2.39$, $p = .017$.

We then tested whether participants' moral preference, as measured by the traditional hypothetical-moral-dilemma battery, predicted their response on the mouse dilemma. A logistic regression demonstrated that the likelihood of consequentialist judgment on the hypothetical mouse dilemma was significantly predicted by participants' preference for consequentialist reasoning (real: $M = 2.99$, $SD = 0.70$; hypothetical: $M = 2.97$, $SD = 0.78$), $OR = 2.14$, $z = 2.17$, $p = .030$. However, no effect was found on the real-life version of the dilemma, $OR = 1.35$, $z = 0.83$, $p = .406$.[2] We quantified the strength of the evidence in favor of this null effect by calculating a Bayes factor (BF) using the Savage-Dickey density ratio (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) with the *brms* package in R (Bürkner, 2016). A weakly informative Student's t distribution ($\nu = 3$, $\mu = 0$, $s = 2.5$) was used as a prior for the regression coefficients. Convergence of Markov chain Monte Carlo permutations was checked through visual inspection of the trace plots. This analysis suggested a BF_{H0} of 5.61, indicating no association between participants' preference for consequentialist judgment and their decision on the real-life version of the mouse dilemma. More details of the Bayesian analyses, along with a prior sensitivity analysis, are available in the Supplemental Material.

Additionally, a logistic regression model using the individual-differences measures and animal empathy to predict the likelihood of a consequentialist decision on the real-life mouse dilemma demonstrated that none of these measures significantly predicted participants' decisions (all $|z| < .150$, all ps $> .134$, BF_{H0}s = 2.12–13.13).

Furthermore, we wanted to test whether participants' moral preferences exhibited in the dilemma battery were related to their reaction times, the extent to which they doubted their decision, and the extent to which they felt uncomfortable when making a decision on the real-life dilemma. A linear regression analysis on the subsample of participants that made a consequentialist decision (as no reaction times were available for the participants who made a deontological decision) demonstrated that participants' preference for consequentialist reasoning was significantly related to a speedier consequentialist judgment, $\hat{b} = -1.40$, $SD = 0.57$, $t(161) = -2.47$, $p = .015$, whereas participants' preference for deontological reasoning was unrelated to their reaction time, $t(161) = 0.96$, $p = .341$. A second linear regression demonstrated that participants' preference for consequentialist reasoning was negatively related to their self-reported doubt about the decision on the mouse dilemma, $\hat{b} = -0.60$, $SD = 0.18$, $t(187) = -3.30$, $p = .001$, whereas their preference for deontological reasoning was marginally positively related to self-doubt, $\hat{b} = 0.36$, $SD = 0.22$, $t(187) = 1.66$, $p = .098$. A third linear regression

analysis with participants' discomfort as the dependent variable demonstrated that participants with a high preference for consequentialist reasoning felt less uncomfortable about their decision, $\hat{b} = -0.70$, $SD = 0.15$, $t(187) = -4.67$, $p < .001$, and participants with a high preference for deontological reasoning felt marginally more uncomfortable, $\hat{b} = 0.35$, $SD = 0.18$, $t(187) = 1.97$, $p = .051$.

As a final control, we also wanted to test whether any of the reported results were dependent on participants' empathy for animals or their empathy for mice specifically. Both empathy for animals and empathy for mice were strongly related to participants' feelings of discomfort, $r = .28$, 95% confidence interval (CI) = [.14, .41], $p < .001$, and $r = .25$, 95% CI = [.11, .38], $p < .001$, respectively. However, both these measures were unrelated to participants' levels of doubt, their reaction times, or the type of decision they made on the real-life dilemma (all ps > .258). Furthermore, neither animal empathy nor empathy for mice moderated any of the effects of either consequentialist or deontological moral reasoning on participants' levels of self-doubt, their feelings of discomfort, their reaction times or, most crucially, their decision (all ps > .088).

Discussion

The current results paint an intriguing picture. Although preference for consequentialist decisions, as measured with a traditional hypothetical-moral-dilemma battery, was predictive of participants' decision on the hypothetical mouse dilemma, this preference was not predictive of how participants behaved on the real-life moral dilemma (despite the higher power of the latter test). We did find that participants' preference for consequentialist reasoning in hypothetical scenarios predicted their reaction times in the real-life dilemma, and more importantly, it predicted the extent to which they doubted their decision and how uncomfortable they felt while making it. Hence, while participants' judgment in hypothetical scenarios was not predictive of their real-life behavior, it was associated with a cognitive and affective measure surrounding that behavior: their (lack of) self-reported doubt and degree of discomfort. Therefore, these results do not imply that the traditional moral-dilemma paradigm completely fails to measure the relevant drivers of real-life behavior, but they do suggest that important aspects of the real-life decision-making process are not captured through the standard trolley paradigm.

In line with this conclusion, we found that participants who were confronted with the real-life dilemma were less than half as likely to make a deontological decision than those who were confronted with a hypothetical version of the same dilemma. This further corroborates the results from studies that have used virtual reality paradigms to study moral cognition, as these also uncovered higher rates of consequentialist responding when confronting participants with dilemmas that are more lifelike (Francis et al., 2016; Patil et al., 2014). Notably these results seem to run counter to the main theoretical model in the field: a dual-process model that equates consequentialist moral reasoning with a cognitive mental process, and deontological moral reasoning with an intuitive, affective mental process. In particular, this model would predict that the increased affective arousal of being confronted with a real-life moral dilemma should actually lead to an increased proportion of deontological responses (Greene, 2007). Moreover, we found that none of the individual-differences measures predicted participants' decisions in the real-life version of the dilemma, despite the robust associations that have been reported in the literature between these measures and moral decision making, and despite the associations we found between these measures and participants' judgments on the battery of traditional trolley-style dilemmas. This, too, is crucial, as the relationship with these measures has been used in previous work to support the aforementioned model. The current study therefore suggests that the theoretical importance of these associations might be overrated and that some of these associative patterns are potentially a side effect of the hypothetical nature of traditional research but are not related to consequentialist moral reasoning per se. Participants typically have not had any real-life experience with trolley-dilemma-like situations. Individual differences in how participants translate an abstract, textual dilemma into a sufficiently salient mental simulation that allows them to fill in this experiential gap might explain the aforementioned associative patterns.

In this regard, the lack of an association between primary psychopathy and participants' actual decision is perhaps illustrative. Previous research has uncovered that measures of antisocial personality are associated with consequentialist reasoning in hypothetical dilemmas. This has been a matter of some controversy, as it seems to suggest that consequentialism is not motivated by the moral concern to minimize harm but is instead driven by a lack of empathy toward harming innocent other people (Bartels & Pizarro, 2011; Koenigs, Kruepke, Zeier, & Newman, 2012). While we did find the expected association between primary psychopathy and preference for consequentialist reasoning as measured with the traditional moral-dilemma battery, primary psychopathy had no meaningful relationship whatsoever with participants' behavior on the real-life dilemma. Additionally, it is worth reiterating that a larger number

of participants favored the consequentialist alternative on the real-life version of the dilemma. It seems unlikely that participants were more empathic toward a hypothetical mouse than toward a real mouse; therefore, at least some participants must change their preference when confronted with the reality of the situation. To us, this indicates that when participants made an actual consequentialist decision, this was not driven by antisocial tendencies or by a lack of empathy but by a genuine concern for the greater good.

One obvious, potential limitation of the current study is that, unlike traditional trolley-style dilemmas, our real-life dilemma did not pertain to moral choices involving humans. From a sacred or protected value perspective, one could argue that people interact in a fundamentally different way with animals than they do with other humans (Tetlock, 2003). However, recent research suggests that there is a symmetry between how people tend to treat animals and other humans (Amiot & Bastian, 2015). In particular, it is probably more appropriate to liken people's treatment of animals to how they treat human out-groups than to assume people treat animals as an entirely different moral category (Dhont, Hodson, & Leite, 2016). Accordingly, we argue that the core mechanisms driving participants' choices on the current dilemma should be the same as those behind traditional, "human" trolley-style dilemmas. After all, the moral conflict that structures the dilemma is the same, regardless of whether it involves humans or animals. Crucially, participants' empathy for animals (or mice) did not moderate any of the effects we have reported in the current article.

A second potential limitation is that our mouse dilemma is an impersonal dilemma (modeled after the archetypical switch-trolley dilemma) because the consequentialist choice entailed redirecting an existing threat, whereas most of the hypothetical dilemmas in our moral-dilemma battery were of a more personal nature (Greene et al., 2001). Skeptical readers might contend that the results we report may be caused by a mismatch along this dimension. While we acknowledge that this is a potential limitation, we advance several arguments as to why this critique ultimately fails. First, our two most crucial results—(a) participants become more consequentialist when confronted with a real-life dilemma, and (b) responses to hypothetical dilemmas are not predictive of decisions on a real-life dilemma but do predict decisions on a hypothetical version of that same dilemma—cannot be explained by referring to the impersonal nature of the mouse dilemma. Secondly, a more fine-grained analysis (available in the Supplemental Material) that does incorporate the personal–impersonal distinction demonstrated that the impersonal hypothetical dilemma included in our moral-dilemma battery did not predict responses to either mouse dilemma better than did the personal dilemmas. Finally, even though some previous research does differentiate between these two types of dilemmas, it is still assumed that participants' responses to impersonal dilemmas are systematically related to their responses on personal dilemmas and that responses to both are generally driven by the same processes. Accordingly, some recent methods of measuring participants' moral preferences have abandoned this distinction altogether (e.g., process dissociation; Conway & Gawronski, 2013). Therefore, we believe that this distinction is mostly irrelevant in our design and does not undermine, or account for, the difference in effects we found for the real-life versus the hypothetical mouse dilemma.

The current study uncovered some important discrepancies between hypothetical judgment and real-life behavior, much like previous work by FeldmanHall et al. (2012). However, when it comes to explaining these discrepancies, the current research is only a first step. We found that this divergence cannot be accounted for by potential differences in empathic concern or cognitive deliberation in real-life versus hypothetical cases. For now, we can only speculate about possible alternative explanations for the discrepancy. A first possibility is that lack of experience with actual trolley-dilemma-like situations causes participants to misjudge hypothetical situations, distorting their own inclinations toward consequentialism or deontology. Another possibility is that answers to hypothetical situations may be determined to a greater extent by virtue signaling than actual behavior is (Everett, Pizarro, & Crockett, 2016). Future research will have to investigate these and other possibilities. In any case, we advance the argument that we will be able to bridge the gap between moral judgment and moral behavior only by exploring new research paradigms that bring moral decision making into the real world.

Action Editor

Marc J. Buehner served as action editor for this article.

Author Contributions

D. H. Bostyn and S. Sevenhant developed the study concept. All authors contributed to the study design. Data were collected and analyzed by D. H. Bostyn and S. Sevenhant. D. H. Bostyn drafted the manuscript, and S. Sevenhant and A. Roets provided critical feedback. All authors contributed to revisions of the manuscript. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Alexandra Melania Pavliuc for proofreading the finished manuscript.

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.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617752640

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/kvb99/. The design and analysis plans for this study were not preregistered. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797617752640. This article has received badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Notes

1. Our moral-dilemma battery consisted of a mix of personal and impersonal dilemmas. This distinction did not influence any of the reported results, and thus we do not make this distinction in the current article. However, interested readers can find a more fine-grained analysis in the Supplemental Material available online.

2. Our exploratory measure for participants' preference for deontological reasoning (real: $M = 2.47$, $SD = 0.56$; hypothetical: $M = 2.38$, $SD = 0.57$) was not associated with participants' decision on either mouse dilemma (both $|z|s < 1.75$, $ps > .081$).

References

Amiot, C. E., & Bastian, B. (2015). Toward a psychology of human-animal relations. *Psychological Bulletin*, *141*, 6–47.

Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*, 1423–1440.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154–161. doi:10.1016/j.cognition.2011.05.010

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573–1576. doi:10.1126/science.aaf2654

Bostyn, D. H., & Roets, A. (2017). An asymmetric moral conformity effect: Subjects conform to deontological but not consequentialist majorities. *Social Psychological & Personality Science*, *8*, 323–330.

Bürkner, P.-C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). doi:10.18637/jss.v080.i01

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics: Simulation and Computation*, *39*, 860–864.

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. doi:10.1037/a0031021

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*, 269–279. doi:10.1080/17470919.2011.614000

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*, 113–126.

Dhont, K., Hodson, G., & Leite, A. C. (2016). Common ideological roots of speciesism and generalized ethnic prejudice: The Social Dominance Human–Animal Relations Model (SD-HARM). *European Journal of Personality*, *30*, 507–522.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*, 772–787. doi:10.1037/xge0000165

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*, 434–441.

Francis, K. B., Howard, C., Howard, I. S., Gummerum, M., Ganis, G., Anderson, G., & Terbeck, S. (2016). Virtual morality: Transitioning from moral judgment to moral action? *PLOS ONE*, *11*(10), Article e0164374. doi:10.1371/journal.pone.0164374

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: Vol. 3. The neuroscience of morality: Emotion, disease, and development* (pp. 35–80). Cambridge, MA: MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364–371.

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. doi:10.1126/science.1062872

Hume, D. (2003). *A treatise of human nature*. North Chelmsford, MA: Courier Corp. (Original work published 1739)

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.

Kant, I. (2002). *Groundwork for the metaphysics of morals* (A. W. Wood, Ed. & Trans.). New Haven, CT: Yale University Press. (Original work published 1785)

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*, 708–714. doi:10.1093/scan/nsr048

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology*, *68*, 151–158.

McDonald, M. M., Defever, A. M., & Navarrete, C. D. (2017). Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas. *Evolution & Human Behavior*, *38*, 770–778.

Moretto, G., Làdavas, E., Mattioli, F., & di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience*, *22*, 1888–1899. doi:10.1162/jocn.2009.21367

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience*, *9*, 94–107. doi:10.1080/17470919.2013.870091

Paul, E. S. (2000). Empathy with animals and with humans: Are they linked? *Anthrozoös*, *13*, 194–202.

Petty, R. E., Cacioppo, J. T., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

Rosen, F. (2005). *Classical utilitarianism from Hume to Mill*. New York, NY: Routledge.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*, 320–324.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.