FINAL AUTHOR VERSION

Sequential decision-making impacts moral judgment: how iterative dilemmas can expand our perspective on sacrificial harm.

D.H. Bostyn & A. Roets

Ghent University; Department of Developmental, Personality, and Social Psychology; Henri Dunantlaan 2, B-9000, Ghent, Belgium.

dries.bostyn@ugent.be, ORCid: 0000-0001-9994-4615

arne.roets@ugent.be, ORCid: 0000-0001-5814-1189

WORD COUNT: 8103

Declarations of interest: None.

Authors' Note:

This research was supported by a postdoctoral research grant from the Research Foundation - Flanders (FWO.3E001619) awarded to the first author.

Correspondence concerning this article should be addressed to Dries H. Bostyn, Department of Developmental, Personality, and Social Psychology, Henri-Dunantlaan 2, B-9000, Ghent, Belgium. E-mail: Dries.Bostyn@Ugent.be.

Abstract

When are sacrificial harms morally appropriate? Traditionally, research within moral psychology has investigated this issue by asking participants to render moral judgments on batteries of single-shot, sacrificial dilemmas. Each of these dilemmas has its own set of targets and describes a situation independent from those described in the other dilemmas. Every decision that participants are asked to make thus takes place within its own, separate moral universe. As a result, people's moral judgments can only be influenced by what happens within that specific dilemma situation. This research methodology ignores that moral judgments are interdependent and that people might try to balance multiple moral concerns across multiple decisions. In the present series of studies we present participants with *iterative* versions of sacrificial dilemmas that involve the same set of targets across multiple iterations. Using this novel approach, and across five preregistered studies (total n = 1890), we provide clear evidence that a) responding to dilemmas in a sequential, iterative manner impacts the type of moral judgments that participants favor and b) that participants' moral judgments are not only motivated by the desire to refrain from harming others (usually labelled as deontological judgment), or a desire to minimize harms (utilitarian judgment), but also by a desire to spread out harm across all possible targets.

Keywords: Moral Cognition, Iterative Decision Making, Utilitarianism, Deontology, Trolley Dilemmas

Highlights

Research on sacrificial harm usually asks participants to judge single-shot dilemmas.

We investigate sacrificial moral dilemma judgment in an iterative context.

Sequential decision making impacts moral preferences.

Many participants express a non-utilitarian concern for the overall spread of harm.

Introduction

Is it morally appropriate to actively harm someone if doing so can ensure an overall better outcome? Psychologists have studied how people respond to this problem by probing them with "sacrificial" moral dilemmas (Cushman et al., 2010; Greene, 2008; Greene et al., 2001, 2004). Sacrificial dilemmas are named as such, because these dilemmas necessitate some form of sacrifice to accomplish a greater good. The quintessential example of these dilemmas, the trolley dilemma, describes a situation where an unfortunate collision between a trolley-tram and five unsuspecting victims can only be averted by diverting the trolley, causing it to collide with another unsuspecting victim instead. The life of the single victim is the sacrifice needed to save the group of five.

Within psychological research, the two possible choices associated with sacrificial dilemmas are typically interpreted as reflecting two opposing moral desires: the decision to save the five is interpreted as stemming from a desire to minimize overall harm, whereas the decision not to interfere is understood to come from a desire to refrain from actively harming others. With a nod to philosophical ethics, moral psychologists commonly label these as "utilitarian" and "deontological" judgments respectively, although it should be noted that the aptness of these labels is actively debated (see: (Conway et al., 2018; Everett & Kahane, 2020; Kahane et al., 2015).[1]

While this research paradigm has been very successful, it has also been critiqued. Some of these critiques have focused on the ecological or predictive validity of sacrificial dilemma research (Bauman et al., 2014; Bostyn et al., 2018; Hester & Gray, 2020). Others have contended this paradigm leads to an impoverished conceptualization of utilitarianism (Kahane et al., 2018) or confounds deontological inclinations with inaction tendencies

[1] Throughout this paper, when referring to the two types of judgments in a sacrificial dilemma as "utilitarian" or "deontological", we use these terms as a pragmatic short-hand for "minimizing harm" and "refraining from sacrificial harm" respectively, not to denote that participants making such judgments necessarily adhere to utilitarian or deontological philosophical tenets.

(Gawronski et al., 2017). In the current manuscript, we focus on a different limitation: traditional trolley dilemma research confronts participants with isolated moral situations and asks them to render a single-shot judgment. As a result, research on this topic has neglected the possibility that people might strive to balance different moral concerns across multiple decisions.

Nearly all research on sacrificial harm uses dilemmas that require participants to make life-or-death decisions (for some exceptions see, Bostyn et al., 2019; Gold et al., 2013; Millar et al., 2016; Millar et al., 2014). As a consequence, each dilemma has to involve a 'fresh set of targets' and every dilemma decision takes place within a different, separate moral universe. It is as though the decisions that participants are required to make, are taken in a moral vacuum, independent from any prior or future decisions they might make. This is not just unrealistic, it also ignores that moral decisions are embedded within a web of other moral decisions and sometimes cannot be interpreted without this wider context. Many of the moral decisions we make, impact the same people over and over. So when we are confronted with the question of whether we have sufficient justification to (sacrificially) harm a specific person, we will likely also consider how we have treated that person previously and consider our prior moral choices as relevant to the new moral issue at hand. We argue that the consequences of this 'embeddedness' have been underappreciated.

Consider the following thought experiment: An electroshock machine is hooked up to 6 people. Five of those people will receive a painful shock but this harmful outcome can be averted by redirecting the shock to the sixth person instead. Although no mortal harm is involved, this is a dilemma in the mold of those that are used throughout the literature. Or at least, it would be if only a single shock were to be delivered. But what if, after an initial decision has been made, the dilemma is repeated and another shock is to be delivered to the

same targets. Would repeating a dilemma, *the exact same dilemma*, lead participants to making the same decision over and over again?

A traditional perspective on sacrificial harm would predict they would. As the structure and content of the dilemma remain unchanged across iterations, participants' responses should remain the same as well. However, we argue that this is unlikely because by repeating the dilemma, new choice options and new moral concerns appear. When a dilemma is repeated, we are not just limited to either refraining from harm or minimizing overall harm, we also have the option to spread harm across all targets. Prior research has already demonstrated that fairness considerations can sometimes trump utilitarian concerns (Roets et al., 2020). By constraining the study of sacrificial harm to research paradigms involving one-shot judgments of mortal harm dilemmas, moral psychologists have ignored the contextualized nature of moral decision making, and have removed this additional moral concern from the equation. As a result, the traditional approach provides an impoverished understanding of peoples' moral concerns within the context of sacrificial harm. We hypothesize that when confronted with a repeated sacrificial dilemma, many people will display a moral preference that is not related to either a desire to refrain from harming or a desire to minimize harm, but instead signals a third moral concern: the overall spread of harm.

The current manuscript describes the results of five preregistered studies that were designed to allow moral concerns regarding the overall distribution of harm to manifest themselves. In each of these studies, we use typical sacrificial dilemmas of the type used all throughout the literature. However, we use *iterative* versions of such dilemmas (without mortal harm). This novel approach allows us to test a) if responding to moral dilemmas in an iterative manner impacts the type of moral judgment that participants prefer, and b) to what extent the moral judgment of sacrificial harms is driven, not merely by concerns to minimize harm or refrain from doing harm, but also by a concern for the overall spread of harm.

Open Science Statement

Preregistrations for studies 1 to 5 are available at https://osf.io/n9k2u/?view_only=a3276b944ba94a449ad8bf295b0644af, https://osf.io/xfs97/?view_only=43079108fbb9471e8e9d1f1b122ab914, https://osf.io/8s24u/?view_only=211483b8345946199221841d852081b4, https://osf.io/d8wm2/?view_only=86f8077a75de4c8cb059de2f5b835993, https://osf.io/d5qzm/?view_only=45190acf87234c7c99f8256d7de7929c, respectively.[2] Detailed information regarding these preregistrations is available through the "files" section of each preregistration. All data, materials, and analysis scripts are available through the OSF project associated with this manuscript at https://osf.io/c5yqk/?view_only=3fe2a534a5984e8994608b420fac1ac2. All analyses were conducted in R (R Core Team, 2012). Studies 1,2, 4 and 5 were conducted in March and May, 2020. Study 3 was conducted in September 2021, in response to the suggestion of an anonymous reviewer. For all studies, we report how we determined our sample size, and all data exclusions, all manipulations, and all measures that were administered. Finally, all studies were conducted in accordance with the ethical protocol of the host institution's Ethical Committee. Figure 1 gives an overview of the type of iterative dilemmas studied in each of the five studies we conducted.

[2] The overarching preregistration page also lists the preregistrations of studies not included in the current manuscript but which are part of the broader moral decision-making research project.

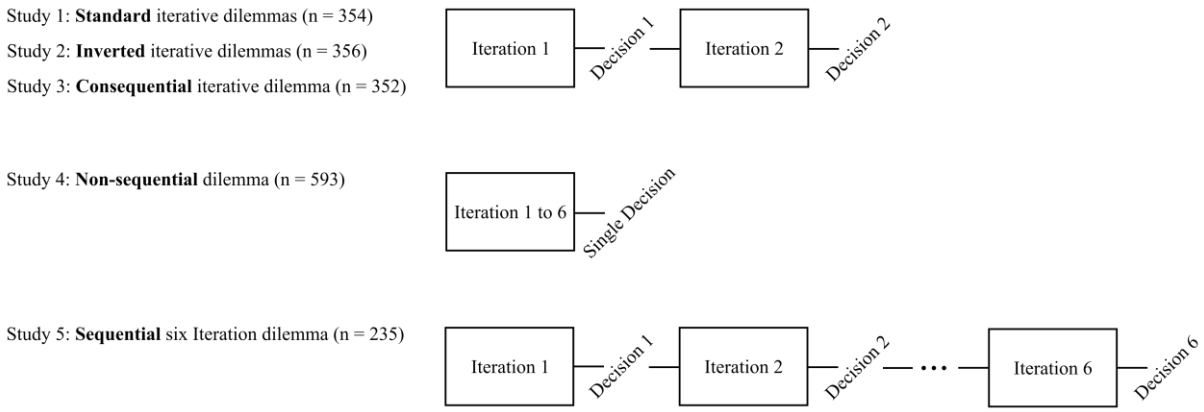

Figure 1. An overview of the type of iterative dilemmas used throughout all studies.

Study 1

Method

Participants and procedure

We recruited 354 North American participants from Amazon Mechanical Turk, each of whom was paid US$0.75. Sample size was determined a priori based on a power calculation demonstrating that a sample of 153 participants per condition was required to obtain 80% power for the least powerful of two planned tests: a one-sided correlational test within each condition. This power analysis was conducted assuming a population level effect size of $r >$ .20. Participants were asked to complete basic demographic information (age & gender) and were subsequently confronted with an iterative dilemma task. On each dilemma, participants had to decide which option they would pursue to do and were subsequently asked to briefly motivate their moral judgment by responding to an open-ended question. Responses to these open-ended questions served as a data-quality check to eliminate participants that were clearly inattentive or had a poor grasp of English.

A total of 70 participants failed the data quality checks and were subsequently eliminated from the sample. Results without any exclusions are available in the online supplementary materials and are qualitatively similar to those reported in the main text.

Accordingly, our final sample consisted of a total of 284 participants of which 121 identified as female and 163 identified as male. Participants had an average age of 36.56 (*SD* = 10.47).

Materials: iterative dilemma task

Participants were confronted with two iterative sacrificial dilemmas. A first dilemma described a situation in which an infrastructure project required expropriating land from five people. This outcome could be averted by expropriating land from a single person instead. The second dilemma described a situation in which a gas leak in a hospital threatened to lengthen the recovery time of five injured patients with an additional month. Again, this outcome could be averted only by redirecting the harm to a single injured patient instead. Accordingly, each dilemma contrasted a response option to minimize overall harm with a response option to refrain from personally doing harm.

Importantly, each participant was confronted with two iterations of each dilemma. The second iteration was framed to reflect the repeated nature of the dilemma but the overall structure and the moral conflict underlying the dilemmas was identical in each iteration. The order in which the expropriation and gas leak dilemmas were presented was randomized across participants.

Halve of all participants were randomly allocated to a control condition in which the second iteration of both dilemmas involved a *new* set of targets. In contrast, the other halve was presented with the experimental condition in which the second iteration of each dilemma explicitly involved the *same* targets as the first iteration. On each iteration of each dilemma, participants were asked to respond whether or not they would commit the sacrificial harm (i.e. either refrain from intervening, or act and thereby commit the sacrificial harm).

Results

If participants' moral judgments are not influenced by their prior moral decisions, then repeated iterations of a dilemma should have no impact on participants' moral preferences. In

the control condition which involved *different* targets for each iteration, this appeared to be the case: 77.4% and 81.8% of participants consistently favored sacrificing the single person to the benefit of the group (the utilitarian response), 19% and 10.9% consistently chose not to interfere in the dilemma (the deontological choice), and only 3.6% and 7.3% of participants switched their response between iterations of the same dilemma (for the infrastructure and gas-leak dilemma respectively). However, in the experimental condition, when repetitions involved the *same* targets, 57.1% and 70.1% favored the utilitarian option, 12.2% and 12.2% favored the deontological option, and 30.6% and 17.7% of participants switched their response. These results are summarized in Figure 2.

We compared the proportion of 'switchers' between both conditions. A one-sided proportion test (as per our preregistration) confirmed the expected increase in the proportion of switchers in the experimental condition compared to the control condition, $\chi^2(1) = 33.70$, $p < .001$ and $\chi^2(1) = 6.00$, $p = .007$, for the expropriation and gas leak dilemma respectively. A power sensitivity test demonstrates this analysis had > 80% power to find effects larger than Cohen's h = .15, assuming a one-tailed test. Moreover, we found that switching behavior on the expropriation dilemma was associated with switching behavior on the gas dilemma in the experimental condition, $r(145) = .43$, $p < .001$, but not in the control condition, $r(135) = .10$, $p = .269$. An unplanned follow-up Fisher z-test confirmed that this difference between the correlations was itself statistically significant, $z = 3.01$, $p = .003$.

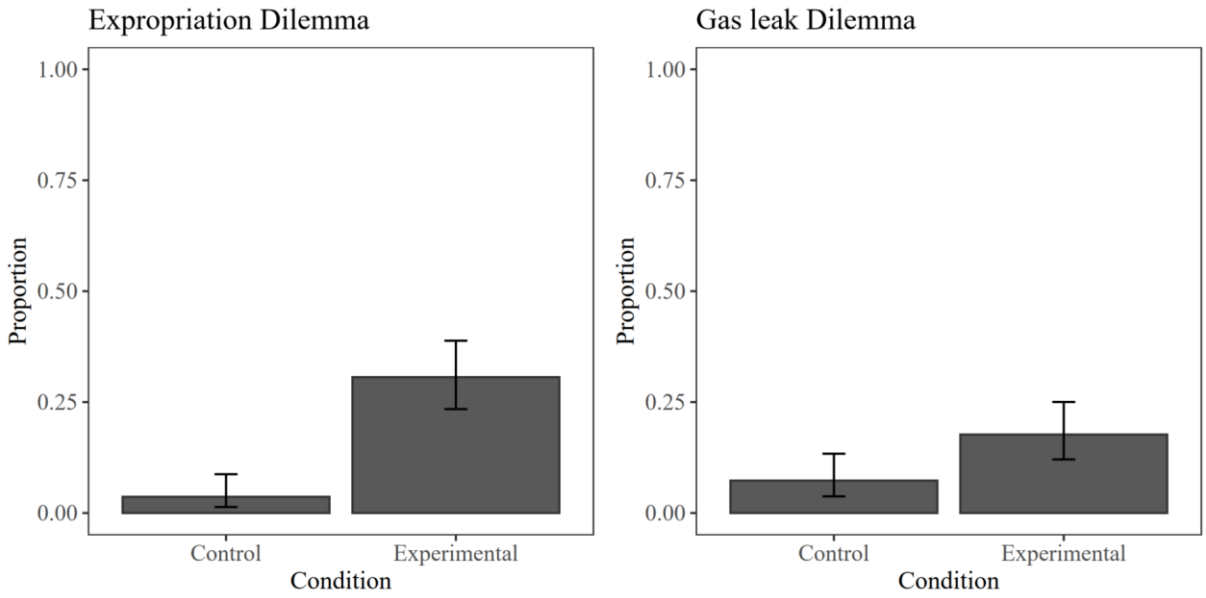

Figure 2. Proportion of "switch responses" for each condition and dilemma. Error bars denote the 95% confidence interval around each estimate.

Discussion

The results of our first study confirm that an increased proportion of participants switch their moral response when confronted with an iterative dilemma if a second iteration involves the same targets. Because the structure of our iterative dilemmas remained the same across iterations, these findings are difficult to fit into the traditional narrative on how people approach problems of sacrificial harm. One possibility is that differential responding across dilemmas is caused by a form of moral ambivalence: perhaps some participants feel drawn (equally) to both options of the dilemma and a "switch" between judgements merely serves to signal this dual allegiance. However, such an explanation ignores the substantially higher occurrence of switch decisions when both iterations involved the same set of targets compared to when it involved different targets.

Accordingly, the more likely interpretation of our results is that repeated victimization of the same targets has an impact on people's moral calculus, and that prior moral decisions influence subsequent moral decisions. In this specific instance, our use of iterative dilemmas

appears to have triggered a third moral concern in people: a concern related to the overall distribution of harms across potential targets. The results of this study demonstrate that, even with minimal opportunity (only 2 iterations), a considerable number of participants take prior harm into account when deciding to what extent new harm is appropriate.

Study 2

While the first study demonstrated that people exhibit a concern related to the overall spread of harm, it does not fully establish that this concern is distinct from the other two moral concerns. To strengthen our case, we conducted a second study in which concerns' related to the distribution of harms are in opposition to both the concern to minimize harm and the concern to refrain from harming. To accomplish this, we constructed a new type of sacrificial dilemmas we labelled as "inverted" sacrificial dilemmas. As the label suggests, these dilemmas invert the traditional structure of a sacrificial dilemma: instead of asking participants whether it is appropriate to harm an individual to save a group, these dilemmas ask whether it is appropriate to harm a group to save an individual. On such an inverted dilemma, both people that prefer not to harm others and those that want to minimize overall harm have an easy choice: non-interference accomplishes either goal. Yet, in an iterative version of such a dilemma, deciding to pursue the sacrificial harm on one of the iterations allows spreading the overall harm across more individuals. Would people still prefer to spread harm across all possible targets when doing so is both anti-deontological and anti-utilitarian?

Method

Participants and procedure

We recruited a total of 356 North-American participants from Amazon Mechanical Turk who were each paid US$0.60. Sample size was determined through a series of power

calculations in which we tested a range of possible effect and sample sizes. We decided upon a sample size of 300 participants as a balance between cost of data-acquisition and power. A full description of these power analyses is available in the preregistration of this study.

Participants completed basic demographic information (age & gender) and were subsequently confronted with an iterative dilemma task. After responding to the dilemma task, participants were asked to guess the hypothesis of the study through an open-ended question. Responses to these open-ended questions served as a data-quality check to eliminate participants with a poor grasp of English or who were inattentive. A total of 69 participants were eliminated from the sample reducing the final sample size to 287 participants. This final sample included 187 participants who identified as male and 100 as female, with an overall average age of 37.26 (SD = 12.09). As was the case for Study 1, results without any exclusions are available in the online supplementary materials and are qualitatively similar to those reported in the main text.

Materials: inverted iterative dilemma task

Participants were confronted with two *inverted* sacrificial dilemmas presented in a random order. Whereas on a typical sacrificial dilemma participants are asked whether it is appropriate to sacrificially harm an individual to the benefit of a larger group, these inverted dilemmas asked participants whether it is appropriate to sacrificially harm the larger group to spare the individual. In traditional dilemmas, sacrificial harm might be deemed acceptable due to the overall utilitarian benefit that can be accomplished through such harm. Hence, these dilemmas contrast the deontological concern to refrain from doing harm with the utilitarian concern of minimizing harm. Importantly, the sacrificial harm on an inverted dilemma always leads to an increase in overall harm and thus the sacrificial harm is no longer sensible for utilitarian reasons. Consequently, on such inverted dilemmas, the deontological concern not to harm is in line with the utilitarian concern to minimize harm: Both concerns dictate refraining

from action. If responses to the problem of sacrificial harm are driven by deontological and utilitarian concerns alone, then no (moral) participant should ever condone sacrificial harm in inverted dilemmas. However, if participants have a third, independent concern related to the overall distribution of harms, then they may approve of the sacrificial harm when they are confronted with an iterative version of such an inverted dilemma.

A first inverted dilemma described a hypothetical situation in which 6 monkeys were hooked up to an electroshock machine. Participants were told that a single monkey would be shocked but that this could be averted by shocking the five other monkeys instead. A second inverted dilemma described a situation in which the noise of construction workers would disturb the sleep of two families. However, this noise could be redirected towards five other families instead. All participants were confronted with two iterations of each dilemma and each iteration involved the same set of targets. As participants have no reason to condone the sacrificial harm on the first iteration, we included no control condition in this experiment. On each iteration of each dilemma, participants were asked whether or not they would commit to the sacrificial harm (i.e., refrain from intervening, or commit the sacrificial harm).

Results

As sacrificial harm in inverted dilemmas runs counter to both the deontological concern to refrain from harming *and* the utilitarian concern to minimize harm, a traditional perspective on sacrificial harm would suggest that such harm would not be condoned. Indeed, only a small proportion of participants decided in favor of sacrificial harm on the first iteration of both the electroshock dilemma (12%) and on the first iteration of the construction dilemma (9.8%). However, when confronted with the second iteration of these dilemmas, an increased proportion of participants condoned the sacrificial harm: 40% of participants favored sacrificial harm on the second iteration of the electroshock dilemma and 36% of participants on the second iteration of the construction dilemma (See Figure 3). A McNemar

test for paired proportion confirmed this difference was statistically significant in both cases: χ^2 (1) = 58.88, $p < .001$; χ^2 (1) = 56.25, $p < .001$, for the shock and construction dilemma respectively. A power sensitivity test demonstrates this analysis had > 80% power if participants were at least 2.6 times more likely to switch from a decision not to condone the sacrificial harm to a decision to condone harm, than they were to switch their decisions in the opposite direction (assuming at least 4% of participants switch in the latter way).

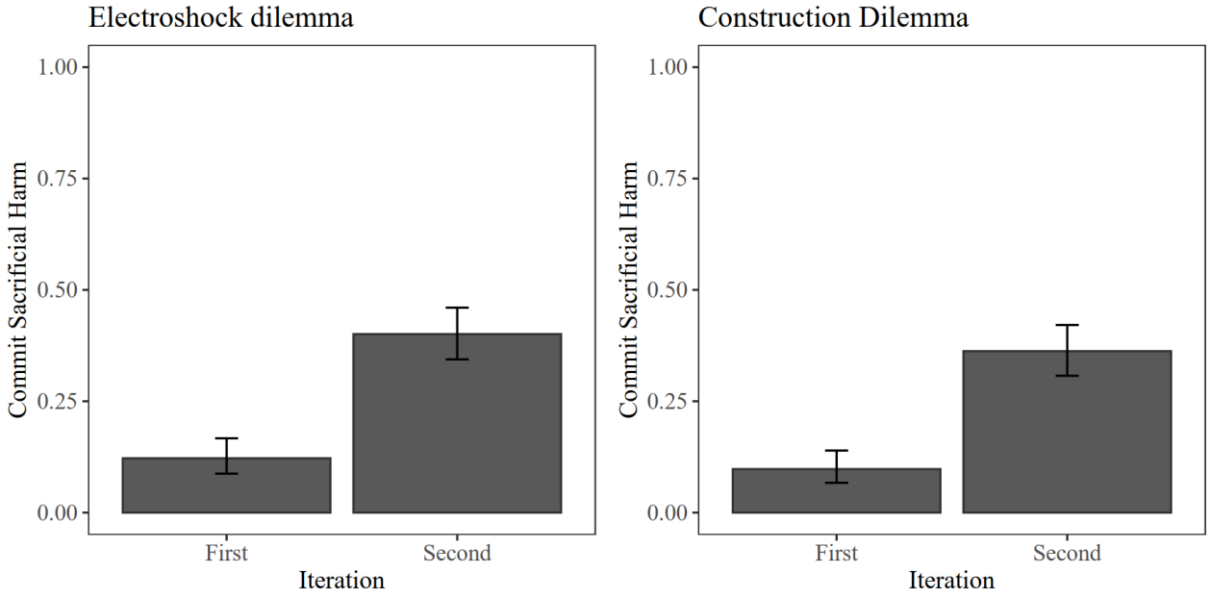

Figure 3. Proportion of participants that approved of the sacrificial harm for each condition and dilemma. Error bars denote 95% confidence interval around each estimate.

Discussion

The results of our second study further corroborate our findings from Study 1. When people are confronted with iterative dilemmas involving the same targets, they take a prior moral judgment into account when making a subsequent moral judgment. This study also confirms that people have a moral concern relating to the overall distribution of harms that is distinct from their concern to avoid doing harm and from their concern to minimize overall harms. Even when sacrificial harm increases the total amount of harm done, many participants still favor committing such harm if doing so spreads harm across all targets.

Study 3

While our first two studies established that participants care about the overall spread of harm, these studies asked participants to make judgments on hypothetical moral dilemmas. Decisions on such dilemmas have no real repercussions. This may matter, because the study of sacrificial dilemmas aims to investigate to what extent participants weigh the consequences of their decisions when making moral judgments. Accordingly, and per the suggestion of an anonymous reviewer, we decided to confront participants with an iterative dilemma where their decisions would have real-life repercussions.

Method

Participants and procedure

A total of 352 North-American participants were recruited from Amazon Mechanical Turk and paid US$0.60. In terms of statistical power, the design of this study was similar to that of Study 2 and thus we decided to use the same sample size.

Participants completed basic demographic information (age & gender) and were subsequently confronted with an iterative dilemma task. After responding to the dilemma task, participants were asked to guess the hypothesis of the study through an open-ended question as a data-quality check. We ran this study through Cloud Research and opted to use their list of Approved participants (Litman et al., 2017). As a result, we had to eliminate fewer participants: only 7 participants were eliminated from the sample based on the same criteria as in our prior studies. Our sample thus included 345 participants, 189 of which identified as female, 153 identified as male and 3 identified as “other”. Participants had an average age of 39.08 (*SD* = 12.08).

Materials: consequential iterative dilemma task

The goal of the current study was to confront participants with a consequential, iterative dilemma, i.e. a dilemma in which the choices they make have real-life repercussions. Inspired by a consequential, charitable donation sacrificial dilemma from Gold et al. (2014), we selected four funding projects from GoFundMe. Each of these funding projects was related to a specific individual raising money to pay for a potentially life-saving medical procedure. We anonymized each person's information, giving them an alias and describing their situation in a one-line summary. Before confronting participants with the moral dilemma, we first informed them that we would ask them to make a decision on a charitable donation, to be paid by the researchers, involving these four people. We also informed them that each of these people was a real person, that we would donate to these people as per the decisions they made, and that our study did not involve any misdirection. We did not tell participants about the iterative nature of the dilemma.

Subsequently, we confronted participants with the first iteration of the dilemma. We first introduced the four people (through their alias and a one-line summary) and we told each participant that we would donate $0.25 to one of the four people. However, if they so desired, they could change this and have us donate $0.25 to each of the three other people instead. Across participants, we randomized which individual was selected to be the single individual and which individuals were part of the group of three. After making a decision on this first iteration, we asked participants to make a second decision on a second donation. To ensure that participants understood the iterative nature of the task, we explicitly told them either that their prior decision had already resulted in the single individual receiving $0.25, or that their prior decision had already resulted in the group of people receiving $0.25 each (based on what they had selected). We then informed each participant that we would make a second $0.25 donation, which was again set to go to the single person. Once again, they could decide to have us donate $0.25 to each of the other three people instead. Once the experiment was

completed, we donated to each of the four GoFundMe projects as per participants' decisions. As a result, we donated a total of US$374.

Results

In line with our expectations, on the first iteration of the dilemma, the majority of participants (76.8%) decided to intervene and have us donate to the group rather than the single individual (the utilitarian option that maximized donations). However, on the second iteration, only 40% of all participants favored that same option. Looking across both iterations, 24.6% of participants decided to make the utilitarian decision to donate to the group both times, and only 7.8% of participants decided in favor of the deontological decision not to interfere in the donation to the individual on both occasions. An overwhelming majority of 67.5% of participants decided to switch across iterations, thereby demonstrating a concern for the overall distribution of harms (or arguably, in this specific case, the overall distribution of benefits). A preregistered McNemar test for paired proportions demonstrated that participants switched in an asymmetric manner, thereby confirming that switch decisions are not caused by random responding. Participants that had initially donated to the group were more likely to switch their preference on the second donation, compared to participants that had initially favored donating to the individual were, $\chi^2(1) = 68.14$, $p < .001$. A power sensitivity test demonstrates this analysis had > 80% power if participants were at least 2.4 times more likely to switch from a decision not to condone the sacrificial harm to a decision to condone harm, than they were to switch their decisions in the opposite direction (assuming at least 4% of participants switch in the latter way).

Table 2 shows a cross-table of participants' decisions across both iterations. Results without exclusions are available in the online supplementary materials and are qualitatively similar to those reported in the main text.

Table 2

Cross-table of participants' choices on the consequential iterative dilemma.

	Iteration 2: donate to individual	Iteration 2: donate to group
Iteration 1: donate to individual	27	53
Iteration 1: donate to group	180	85

Discussion

The results of this third study show that participants take their prior moral decisions into account, not just when they are confronted with hypothetical dilemmas, but also when confronted with a consequential dilemma. Notably, an overwhelming majority of participants (67.5%) switched their moral judgment across iterations in this consequential dilemma, which is substantially higher than in the two previous studies. It is likely that the extent to which participants favor minimizing harm, refraining from harm or spreading harm will depend on a number of different factors. However, one possible explanation for the increased proportion of switch decisions uncovered in the current study might be that participants were asked to make sacrificial moral judgments that were framed in terms of benefits rather than harms. Prior research by Roets et al. (2020) has established that participants tend to be more fairness-minded when they are asked to distribute benefits, and more utilitarian when asked to make decisions on distribution tasks involving harms.

Study 4

Up until this point, we have investigated iterative sacrificial dilemmas with only two iterations. Even in this minimal iteration context, we found that presenting participants with an iterative dilemma in a sequential manner impacts the type of judgments that they make: Participants seem to take their prior judgment into account and they appear to be concerned

with the overall distribution of harms (or benefits) in a way that is distinct from a concern to simply minimize harms (or maximize benefits).

One may wonder whether concerns for the overall distribution of harms arise only when participants are presented with iterative dilemmas in a sequential manner. Therefore, in our fourth study we decided to present participants with a non-sequential version of an iterative sacrificial dilemma that still allows participants to spread out the harm if they so desire. Through such a design, we can determine whether concerns relating to the overall distribution of harms arise only because of the sequential nature of the dilemmas we have studied so far, or whether this concern emerges in non-sequential versions of such dilemmas as well.

Method

Participants and procedure

A total of 593 North-American participants were recruited from Amazon Mechanical Turk and paid US$0.40. Sample size was determined a priori through an accuracy of estimation approach and based on the funding available for the project. We aimed for 100 participants per condition which would ensure that the 95% CI around each proportion would have a width of .20 (at most).

Participants completed basic demographic information (age & gender) and were subsequently confronted with an iterative dilemma task. After responding to the dilemma task, participants were asked to guess the hypothesis of the study through an open-ended question to check their proficiency with English. A total of 50 participants were eliminated from the sample based on the same criteria as in the prior studies. Our final sample included 543 participants, of which 277 who identified as male and 266 who identified as female. Participants had an average age of 37.88 (SD = 13.36).

Materials: non-sequential iterative dilemma task

In the non-sequential version of the iterative sacrificial dilemma, participants were asked to imagine a scenario in which an electroshock machine was hooked up to 6 monkeys and about to give *multiple* shocks to five of the monkeys. As per the structure of a typical sacrificial dilemma, each of those shocks could be averted to the sixth monkey by pressing a button. Pressing one button would redirect one shock, pressing two buttons would redirect two shocks, etc. We randomly varied the total amount of shocks that would be delivered (and thus also the total amount of hypothetical buttons) between participants from 2 to 6 shocks in total. Participants were asked how many (if any) of the buttons they would press.

Results

For each total number of shocks that could be delivered, we calculated the proportion of participants that redistributed none of the shocks (which we designated as the deontological choice), redistributed all of the shocks (which we labelled as the utilitarian choice) or redistributed some, but not all of the shocks to the single monkey (thus distributing some of the harm to all possible targets). Results are summarized in Figure 4. Two key findings emerged. First, about 38% of all participants favored neither the deontological nor the utilitarian solution to the dilemma and thus displayed a concern for the overall distribution of harms. Secondly, this proportion, as well as the proportion of utilitarian and deontological choices, remained stable, even when the total amount of shocks that could be given was increased (see Figure 4). A test for a difference in proportions demonstrated that the proportion of participants favoring ‘distributing’ the shocks in the dilemma when a total of 2 shocks were to be delivered, was statistically similar to the proportion of participants favoring spreading out the shocks when a total of 6 shocks could be delivered, , χ^2 (1) = 0.48, p = .488.

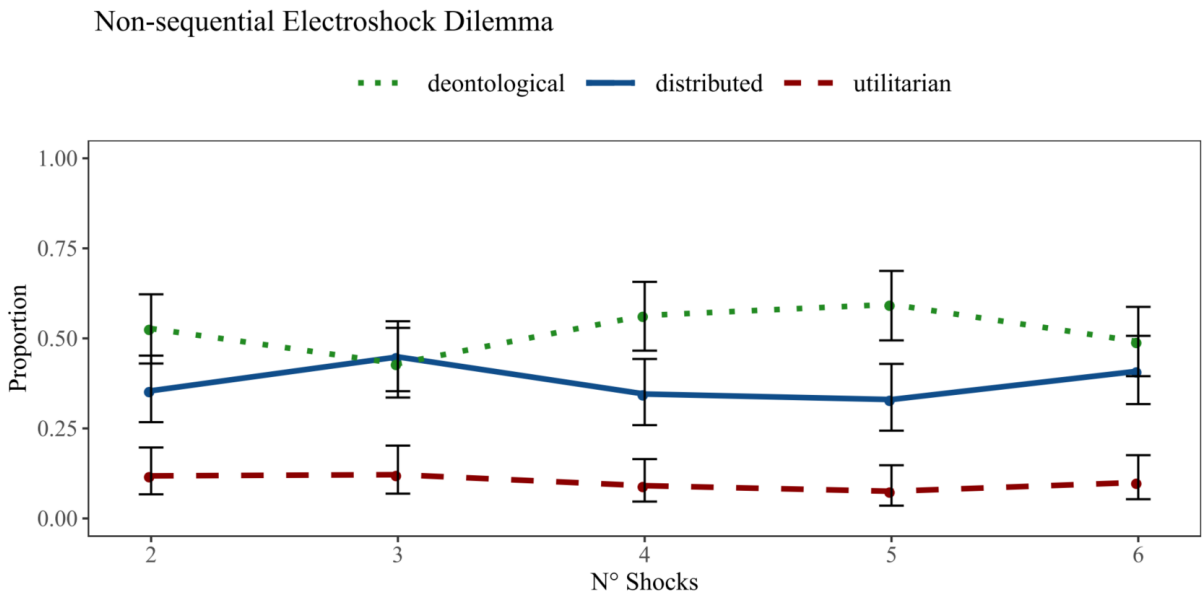

Figure 4. Proportion of utilitarian, deontological and distributed responses. Error bars denote 95% confidence interval around each estimate.

Discussion

Our fourth study demonstrated that even when an iterative dilemma is not presented in a sequential manner and participants are asked to render a single judgment, a concern for the distribution of harm is present. Moral concern for the overall distribution of harms does not only arise when dilemmas are presented sequentially, but seems to arise as soon as it is possible for this concern to play a role. Interestingly, the proportion of participants that is concerned with the overall distribution of harms did not increase monotonically with increases in the total amount of shocks that could be given.

Accordingly, a question arises to what extent responding to dilemmas in a sequential manner impacts participants' moral concerns beyond allowing participants to express a concern for the overall spread of harm in non-sequential contexts. Does responding in a sequential manner cause participants to pay more attention to the spread of harm?

Study 5

In a fifth study, we explored the impact of adding additional sequential iterations to iterative dilemma tasks, to test whether sequential decision making impacts the type of moral decisions that participants favor.

Method

Participants and procedure

A total of 235 North-American participants were recruited from Amazon Mechanical Turk and paid US$0.80. Sample size was determined a priori through an accuracy of estimation approach. As the current study involved a within-subject design, we could afford to increase the precision with which we estimated each proportion. We aimed for a total of 200 participants which ensured that the 95% CI around each proportion estimate would have a width of .13 (at most).

Participants first completed basic demographic information (age & gender) and were subsequently confronted with an iterative dilemma task. After completing the dilemma task, participants were asked to guess the hypothesis of the study as a data-quality check. A total of 34 participants failed this check and were eliminated from the sample. Our final sample consisted of 201 participants with an average age of 37.11 (SD = 11.58), of which 80 self-identified as female and 121 self-identified as male.

Materials: Sequential iterative dilemma task

Participants were confronted with a single sequential iterative sacrificial dilemma. As was the case in the previous study, participants were asked to imagine a scenario in which an electroshock machine was hooked up to 6 monkeys and about to give a painful electroshock to five of the monkeys. Participants were asked if they would press a button that would divert this shock to the sixth monkey instead. After responding, participants were confronted with an

additional iteration of the same dilemma and were asked the same question. Participants were asked to respond to six, sequential iterations of this dilemma, each of which involved the same targets. Participants were not informed about how many iterations of this dilemma they would be exposed to.

Results

For each iteration, we calculated the proportion of participants that had (up until that point) redirected a) all of the shocks (utilitarian response), b) none of the shocks (deontological response) or c) at least one, but not all of the shocks (thus demonstrating concern for the overall distribution of harms). Results are summarized in Figure 5.

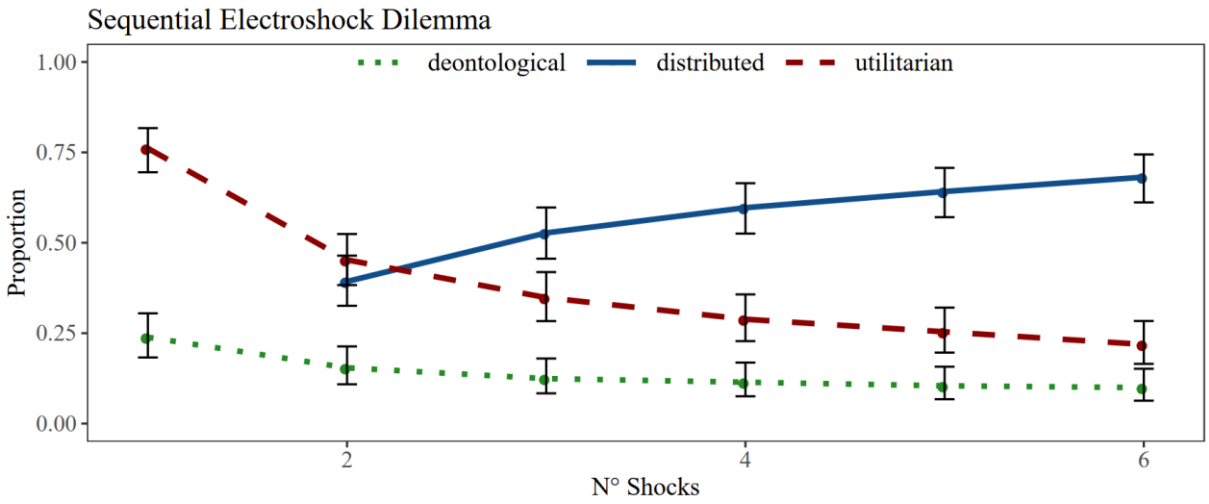

Figure 5. Proportion of utilitarian, deontological and distributed responses. Error bars denote 95% confidence interval around each estimate.

The current study demonstrates that sequentiality impacts the type of moral judgments that people make. In contrast to the results of the single-decision, non-sequential design in the previous study, in the present sequential design, the proportion of participants that decided to "spread out the harm" increased with each additional iteration. After the second iteration,

about 39%, $CI_{95\%}$ [33% to 46%], of all participants had decided to spread out some of the harm. At the sixth and final iteration, this proportion had risen to 68%, $CI_{95\%}$ [61% to 74%].

These results suggest that responding in a sequential fashion causes participants' concern for the overall distribution of harm to accumulate and impacts their moral preferences. In fact, this conclusion can be bolstered by comparing the results of the current study with those of the previous study. First, the extent to which participants preferred the deontological and utilitarian options is impacted by asking them to respond to this dilemma in a sequential manner. In Study 4, when participants were asked how to distribute two shocks simultaneously, only 11.8%, $CI_{95\%}$ [6% to 20%] of them favored the utilitarian option. In Study 5, when they were confronted with the exact same problem in a sequential manner, 45%, $CI_{95\%}$ [38% to 52%] favored the utilitarian option. Secondly, looking at the distribution of how exactly participants decided to spread out the harm is informative as well. Figure 6 displays a density plot of how participants decided to distribute harm in Study 5 and compares this to how the subset of participants that were confronted with 6 shocks decided to spread out these shocks in Study 4. Notably, whereas few participants in Study 4 favored a balanced spread in which each possible target receives an equal amount of shocks, a plurality of participants favored such a spread in Study 5.

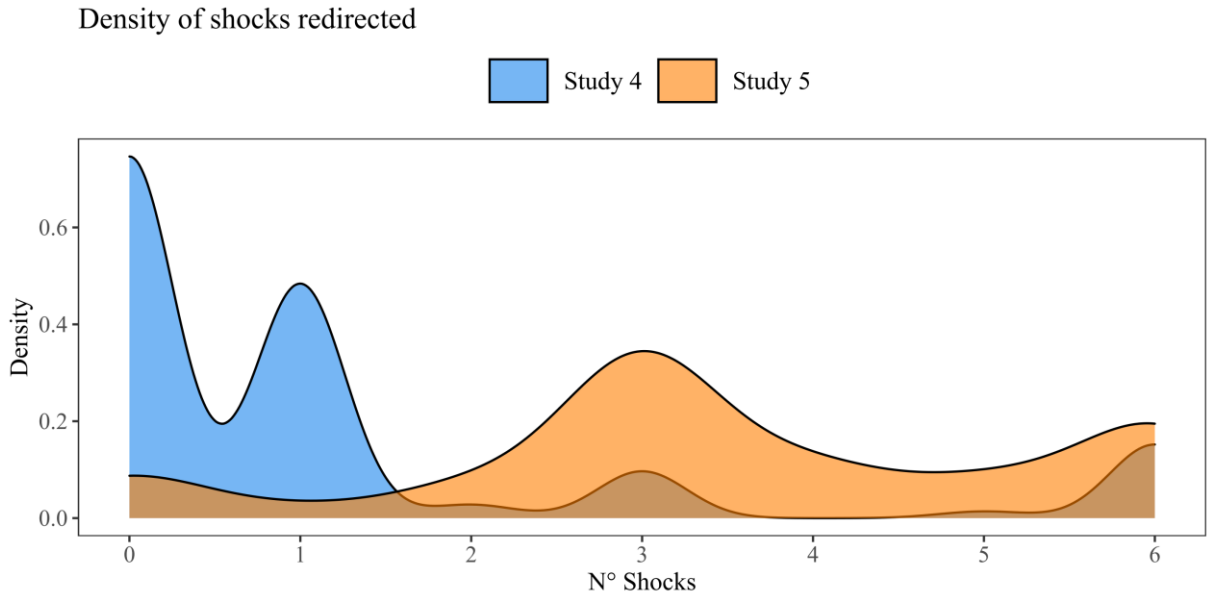

Figure 6. Density plot of the number of shocks redirected (out of a total of six) for the subset of participants confronted with 6 shocks in Study 4 and Study 5.

Discussion

The results of our fifth study demonstrate that asking participants to respond to an iterative moral dilemma in a sequential manner impacts the moral preferences they display. Participants' moral concern for the overall spread of harm increases with additional sequential iterations. With each additional iteration, fewer participants favored resolutions that purely minimized overall harm or purely refrained from doing harm, and more participants favored spreading out the harm. Furthermore, the results of this study demonstrate that asking participants to respond in a sequential fashion causes them to favor a spread of harm that is arguably more balanced. Whereas only a minority of participants favored a fully balanced spread in study 4, a plurality of participants favored such a spread in Study 5.

General Discussion

The current manuscript describes five studies that investigate how people respond to iterative sacrificial dilemmas. Two key results emerged from this work. First, we uncovered that participants are not only concerned with refraining from harm (which moral psychology researchers typically label as a deontological concern) or with minimizing overall harm (typically labeled as a utilitarian concern), but that they also appear to have a third, distinct type of moral concern pertaining to the overall distribution of harms. This result is perhaps best illustrated by our second study where we presented participants with inverted versions of iterative sacrificial dilemmas. On such inverted dilemmas, harm to one or a few can only be averted if participants decide in favor of harm to many instead. In other words, on such inverted dilemmas, sacrificial harm increases the total amount of harm done. Accordingly, on

a dilemma like this, sacrificial harm is contraindicated both if one desires to refrain from harming and when one wishes to minimize harm. Nevertheless, a substantial proportion of participants favored such sacrificial harm when confronted with the second iteration of an inverted dilemma.

Secondly, our studies demonstrate that responding to sacrificial dilemmas in a sequential manner impacts the type of moral judgments that participants prefer.[3] In our fourth and fifth study, we asked participants how many out of a series of shocks they would divert from a group to an individual. When participants were asked to decide about all shocks at once, most favored non-interference. However, when they were asked about each shock individually in a sequential manner, a plurality of participants favored a fully balanced spread.

Moral deliberation in iterative contexts

The iterative lens we have adopted prompts some intriguing questions about the nature of moral deliberation in the context of sacrificial harm. Existing theoretical models on sacrificial harm can be described as 'competition models' (for instance, Conway & Gawronski, 2013; Gawronski et al., 2017; Greene et al., 2001, 2004; Hennig & Hütter, 2020). These models argue that opposing psychological processes compete to deliver a specific moral judgment and that the process that wins out, will determine the nature of that moral judgment. As such, these models presume that the goal of moral deliberation is about deciding whether to refrain from harm or minimize harm in a mutually exclusive manner. Even if participants are tempted by both options, eventually, their judgment settles wholly on one or

[3] It could be noted that many of the dilemmas we studied involved making decisions about harming animals. Prior research (Caviola et al., 2021) has established that people tend to be more utilitarian when making decisions involving animals. While some of our results might be biased in this regard, the conclusions we reach are not related to the specific level of utilitarian preference people display. Rather, our conclusions are based in changes to this preference when confronted with iterative dilemmas. Furthermore, multiple of our studies asked participants about moral decisions involving humans as well. The findings from these dilemmas corroborate the findings from dilemmas that involved animals.

the other. This is sensible in the context of non-iterative dilemmas in which outcomes hinge on a single decision but is it equally sensible in iterative contexts?

Consider the results of Study 4. In this study, we asked (a subset of) participants how many shocks they would divert out of a total six shocks. Interestingly, 32% of these participants decided to divert a single shock out of the six (See Figure 6), thus shocking the individual once, and the group five times. How should such a decision be interpreted? These participants did not fully refrain from harming others, nor did they fully minimize harm, nor did they spread harm in the most balanced of ways. Responses like this seem to straddle different moral concerns. While future research will need to corroborate these findings, we suggest that responses like this, i.e. responses that seem to straddle multiple moral concerns, cannot be explained by competition models but necessitate theoretical models that explicitly take into account that participants might strive to strike a (idiosyncratic) pluralistic balance between multiple moral concerns.

Are moral judgments independent?

Additionally, our studies prompt methodological questions. The current findings demonstrate that moral decisions can be embedded within a stream of other moral decisions and that some moral decisions only make sense when interpreted from the perspective of other moral decisions. While this might appear to be a straightforward observation, we would argue that the consequences of this embeddedness are underappreciated in the literature. We advance that future research on sacrificial harm would do well to integrate dimensions of repeated interactions more often within research designs. If prior moral judgments contextualize future moral judgments then the single-shot approach to the study of moral cognition that is currently the standard within the literature, will invariably miss out on important aspects of our moral minds. Perhaps, trying to understand moral cognition by

focusing on how people respond to single, isolated instances of sacrificial harm is akin to trying to understand why people like music by analyzing single, isolated fragments of music, each in separation from one another. Although such a study of music is not without value, our appreciation for a specific fragment of music is not wholly determined by what happens in that specific fragment. How it fits within the wider piece of music is arguably just as important.

Moral psychologists have been focused on explaining why people prefer specific responses to specific moral dilemmas. The present series of studies shows that this focus is limiting, as it prevents meaningful moral concerns, such as those related to the overall distribution of harms, from materializing in their decisions. From the perspective of entangled moral decision making, the answers we give to individual dilemmas may matter less than the moral preferences we display across sets of moral decisions. A more holistic approach to moral psychological research, one that explicitly embraces how moral judgments of harm relate to one and other, and explicitly accounts for the embedded nature of our moral judgements might lead to more nuanced theories of how people deal with moral complexities.

References

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*(9), 536–554. https://doi.org/10.1111/spc3.12131

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*(7), 1084–1093. https://doi.org/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*(2), 192–206. https://doi.org/10.1080/13546783.2018.1497536

Caviola, L., Kahane, G., Everett, J. A. C., Teperman, E., Savulescu, J., & Faber, N. S. (2021). Utilitarianism for animals, Kantianism for people? Harming animals and humans for the greater good. *Journal of Experimental Psychology: General*, *150*(5), 1008–1039. https://doi.org/10.1037/xge0000988

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), 216–235. https://doi.org/10.1037/a0031021

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. https://doi.org/10.1016/j.cognition.2018.04.018

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford Handbook of Moral Psychology*, 47–71.

Everett, J. A. C., & Kahane, G. (2020). Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134. https://doi.org/10.1016/j.tics.2019.11.012

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, Norms, and Generalized Inaction in Moral Dilemmas: The CNI Model of Moral Decision-Making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making*, *9*(1), 65–76.

Gold, N., Pulford, B. D., & Colman, A. M. (2013). Your Money Or Your Life: Comparing Judgements In Trolley Problems Involving Economic And Emotional Harms, Injury And Death. *Economics and Philosophy*, *29*(2), 213–233.

Greene, J. D. (2008). The secret joke of Kant's soul. In *Moral psychology, Vol 3: The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). MIT Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*(2), 389–400. https://doi.org/10.1016/j.neuron.2004.09.027

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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Hennig, M., & Hütter, M. (2020). Revisiting the divide between deontology and utilitarianism in moral dilemma judgment: A multinomial modeling approach. *Journal of Personality and Social Psychology*, *118*(1), 22–56. https://doi.org/10.1037/pspa0000173

Hester, N., & Gray, K. (2020). The Moral Psychology of Raceless, Genderless Strangers. *Perspectives on Psychological Science*, *15*(2), 216–230. https://doi.org/10.1177/1745691619885840

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. https://doi.org/10.1016/j.cognition.2014.10.005

Kahane, G., Everett, J. A., Earp, B., Caviola, L., Faber, N., Crockett, M., & Savulescu, J. (2018). Beyond Sacrificial Harm: A Two-Dimensional Model of Utilitarian Psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Millar, J. C., Starmans, C., Fugelsang, J., & Friedman, O. (2016). It's personal: The effect of personal value on utilitarian moral judgments. *Judgment and Decision Making*, *11*(4), 326–331.

Millar, J. C., Turri, J., & Friedman, O. (2014). For the greater goods? Ownership rights and utilitarian moral judgment. *Cognition*, *133*(1), 79–84. https://doi.org/10.1016/j.cognition.2014.05.018

R Core Team. (2012). R: A language and environment for statistical computing. *Http://Www.R-Project.Org*. https://ci.nii.ac.jp/naid/20001689445/

Roets, A., Bostyn, D. H., De Keersmaecker, J., Haesevoets, T., Van Assche, J., & Van Hiel, A. (2020). Utilitarianism in minimal-group decision making is less common than

equality-based morality, mostly harm-oriented, and rarely impartial. *Scientific Reports*, *10*(1), 13373. https://doi.org/10.1038/s41598-020-70199-4