# Order Effects in Moral Judgments

Forthcoming in *Philosophical Psychology*

Alex Wiegmann (Göttingen), Yasmina Okan (Granada), and Jonas Nagel (Göttingen)

July 19, 2011

# Abstract

Explaining moral intuitions is one of the hot topics of recent cognitive science. In the present article we focus on a factor that attracted surprisingly little attention so far, namely the temporal order in which moral scenarios are presented. We argue that previous research points to a systematic pattern of order effects that has been overlooked until now: Only judgments of actions that are normally regarded as morally acceptable are susceptible to be affected by the order of presentation, and this in turn is only the case if the dilemma is immediately preceded by a dilemma in which the proposed action was considered as not morally acceptable. We conducted an experiment that largely confirmed this pattern and allowed us to analyze by what individual level responses it was generated. We argue that investigating order effects is necessary for approaching a complete descriptive moral theory. Furthermore, we discuss the implications of these findings for moral philosophy.

Keywords: Order effects; Moral intuitions; Trolley dilemmas; Experimental philosophy

## 1. Introduction and Overview: Order Effects in the Context of Moral Intuitions

Moral dilemmas are ubiquitous in everyday life. In the medical, political or economical domain we frequently encounter situations where our values enter into conflict, but where we need to make a decision regarding which actions are acceptable and which are not. For example, should we allow ending the life of a person to relieve his or her intractable suffering? Is it acceptable to apply death penalty to criminals who have committed a very severe offence? Philosophers and psychologists have aimed to understand moral judgment in several types of dilemmas. In particular, during the past decades, trolley dilemmas have become a keystone for testing alternative normative (e.g., Kamm, 2007) and descriptive (e.g., Hauser, 2006) theories of moral judgment. In the present research we will focus on this kind of dilemma to study order effects in moral reasoning.

Since an out of control trolley was mentioned for the first time in the form of a thought experiment (Foot, 1967), a multitude of variants of this trolley case have been developed and are still used in philosophy as well as in psychology. In the standard description of the original trolley dilemma, a switch can be pulled to redirect a train that is out of control to a different track. If nothing is done, the train will run over five people who are standing on the train’s track. If the switch is pulled, the five people will be saved, but a person who is standing on the track onto which the train is then redirected will be killed. In an alternative and equally well known version of this scenario (Thomson, 1976), the only option available to save the five people is not to redirect the trolley, but instead to push a heavy person who happens to be standing on a footbridge above the tracks into the trolley’s path. This would stop the train but kill this person. A puzzling finding is that despite the fact that the number of people that can be killed vs. saved is the same in both scenarios, people tend to agree with the action proposed in the first scenario, but not in the second one (see, e.g., Hauser, 2006).

While in philosophy trolley cases have mainly been used in the form of thought experiments to argue for (or against) certain moral principles (e.g., Foot, 1967; Kamm, 2007, Otzuka, 2008; Thomson, 1976, 1986, 1990, 2008), psychologists have frequently collected responses of laypeople to trolley cases in questionnaire-based experiments to identify the psychological principles and factors underlying their moral judgments (e.g. Cushman, Young, & Hauser, 2006; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen, 2009; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000, 2007; Nichols & Mallon, 2006; Petrinovich & O'Neill, 1996; Waldmann & Dieterich, 2007).

Although the number of both studies analyzing moral intuitions and studies aiming to document factors influencing moral judgment is enormous, surprisingly little is known about the effects of the order in which several consecutive scenarios are presented. It is plausible to assume that consecutive scenarios will not be judged independently of each other. A principle or mechanism that becomes active when representing or evaluating a particular scenario might later be applied to a series of subsequent scenarios, making their evaluation dependent on the scenarios previously judged. Results obtained in studies that exceptionally manipulated the order of presentation of a series of dilemmas suggest that it is indeed the case that an initial moral judgment of a particular action (e.g., that the action is wrong) can in some cases affect how different subsequent actions will be judged (e.g., different subsequent actions are also judged to be wrong; Lanteri, Chelini, & Rizzello, 2008; Petrinovich & O'Neill, 1996).

We will address three main goals in the present work. First, we will review previous empirical research that found order effects in moral judgments, and we will argue that there is a systematic pattern of results that has been overlooked so far. Second, we will empirically test the existence, extent and direction of this kind of order effects. To this end, we will report results of an experiment designed and conducted specifically for this purpose. Finally, we will discuss the theoretical and practical implications of our findings, both for psychological and philosophical theories of moral reasoning.

### 1.1. Order Effects in Previous Research

When we speak of order effects in moral judgment, there are at least two possible understandings. First, it is possible that the order in which different pieces of information are presented within a particular case could influence how this information is weighted and integrated. This would directly affect the judgments made on the basis of this information. This category of order effects could be labelled “within-scenario order effect”. For instance, consider a scenario in which an action is described and the task is to judge the permissibility of this action. In one condition, the negative consequences of the action are described before the positive ones are mentioned. In the second condition, the order in which the consequences are described is reversed. Differences in the judgment between these conditions would be attributable to order effects that represent this first understanding. Some contexts where this kind of effect has been investigated include legal decision making (e.g. Constabile & Klein, 2005; Kerstholt & Jackson, 1998), or moral reasoning in children (Austin, Ruble, & Trabasso, 1977; Feldman, Klosson, Parsons, Rholes, & Ruble, 1976). In these cases judgments have been found to be directly affected by recency or primacy effects.

Second, a moral judgment regarding a potential action might be influenced by previous moral judgments. The judgment regarding an action in a particular scenario may be influenced by a judgment that had previously been made about a different action in another scenario. We might call this kind of order effect “between-scenarios order effect”. The present paper deals with this category only. We regard the isolated judgment of a given scenario (i.e., a judgment which has not been immediately preceded by another judgment of a different scenario) as a baseline value. This independent rating serves as a reference point against which we will evaluate whether a judgment of the same scenario *as part of a sequence of moral judgments* is affected by an order effect. Thus, whenever we refer to a judgment as being affected by an order effect, we denote cases in which one and the same scenario is judged differently depending on whether it is presented independently vs. immediately after a different scenario has been judged. We are focussing on the influence of the moral judgment *immediately* preceding the present judgment, although we do not want to rule out that, potentially, a given judgment might be influenced by *any* preceding judgment.

To the best of our knowledge, very few studies have addressed this issue in reasoning about moral dilemmas. One of the few exceptions is a series of studies conducted by Petrinovich and O'Neill (1996), where order effects were analyzed by presenting a set of moral dilemmas in two different orders, with the aim of establishing how this manipulation affected participants' level of agreement or disagreement with the action proposed in each case. In one condition (standard order), the first dilemma presented was the one predicted to lead to the highest level of agreement with the potential action, and the last one was the one expected to lead to the lowest level of agreement (intermediate ones were also ordered according to this prediction). In the second condition (reversed order), the same set of dilemmas was presented, but the order was reversed.

In the first experiment exploring order effects (Study 2, Forms 1 and 1R), the first dilemma presented in the standard order condition was a standard switch-trolley dilemma (Trolley). The second dilemma in the set (Scan) described a situation where five dying persons could be saved by scanning the brain of a healthy individual, who in this case would die as a result. In the third dilemma (Dissection), five dying persons could be saved by performing surgery on a healthy person in order to extract five of his organs to distribute amongst the patients, thus causing the death of this individual. This order of presentation of dilemmas was reversed in the second experimental condition.

The authors analyzed the existence of order effects by exploring the three-way interaction between: (1) the dilemma to be judged (2) the order in which the dilemma was presented (3) whether the question involved conducting an action or not conducting it. This interaction was found not to be significant in this case. However, an alternative inspection of their data is warranted to test whether or not a "between-scenarios order effects" as defined above is present. Following a comparison of the ratings for agreement with the potential action in the two conditions, it can be observed that the judgment of the switch-trolley dilemma was indeed affected by an order effect, whereas the judgments of the other two dilemmas were not. When the switch-trolley scenario was presented first, the average rating for agreement with the action was 2.9, whereas it was only 1.5 when it was preceded by the other two dilemmas (agreement or disagreement with action was rated on a scale ranging from +5, “strongly agree”, to -5, “strongly disagree”). A reanalysis of Petrinovich and O’Neill’s (1996) data proves this difference to be statistically significant ($t_{57}$=2.11, $p$<.05, two-tailed, $\eta^2$=.07). In contrast, the ratings for the other two dilemmas remained almost unchanged between both experimental conditions.

In contrast to the first experiment, the content of the three dilemmas did not change substantially in the second experiment (Study 2, Forms 2 and 2R). In this case, three variations of the trolley dilemma were included: (1) the switch-trolley dilemma (Trolley), (2) a similar dilemma where the proposed action was to push a button that would cause the trolley to jump onto a bridge where a person was standing, running over this person but saving the five (Button), and (3) the footbridge dilemma described above, where the potential action was to push a large person standing on a bridge who would fall onto the track and die, stopping the train and saving the five (Person). Again, the order of presentation was reversed in one of the experimental conditions, and in this case a reanalysis of the data revealed that judgments of Trolley ($t_{57}$=2.93, $p$<.01, two-tailed, $\eta^2$=.13) and Button ($t_{57}$=2.58, $p$<.05, two-tailed, $\eta^2$=.10) were affected, but not the judgment of the dilemma that involved pushing a person.

Finally, the last experiment (Study 2, Forms 3 and 3R) included a series of five dilemmas where, as in the first experiment, the context changed substantially. As in previous experiments, the different dilemmas were presented in two different orders according to the predicted level of agreement with the proposed action in each case. Three of the scenarios elicited positive average ratings (i.e., in this specific study ratings indicating agreement with the claim that the action should be done).A reanalysis of the original results showed that the judgment of one of those dilemmas (Shield, presented in the third position in both conditions) was affected by the order of presentation: The average rating was lower when it had been directly preceded by a dilemma that received lower ratings, than when directly preceded by a dilemma that received higher ratings ($t_{68}$=2.88, $p$<.01, two-tailed, $\eta^2$=.11). Taken together, these results seem to show that only judgments of dilemmas in which the action is rated positively are in some cases affected by the order of presentation, while judgments of dilemmas that receive a negative rating remain unaffected.

Another study investigating order effects was conducted by Lanteri et al. (2008). They asked participants to judge the potential action in the standard switch-trolley dilemma and in the footbridge dilemma described above. In one condition the former was presented first, followed by the latter (standard treatment), while in the other condition this order was reversed (reversed treatment). The results showed that the percentage of people who regarded the proposed action in the switch-trolley scenario as morally acceptable was lower in the reversed treatment than in the standard treatment (78% vs. 94%, respectively). However, the percentage of people who judged pushing the stranger in the footbridge dilemma as acceptable was equally low in both conditions (46% vs. 48%, respectively). It is interesting to note how these results mirror those obtained by Petrinovich and O’Neill (1996, Study 2, Forms 2 and 2R).

Similar order effects have also been found unexpectedly by other researchers. For example, in a recent study designed to investigate the role of explicit moral commitments in the judgment of trolley problems, Lombrozo (2009) reported that her subjects judged acting in the switch case to be less permissible if it had been preceded by the push scenario than if it had been presented before the push scenario. Due to the different focus of her work, she did not discuss this result any further. Similarly, Nichols and Mallon (2006) unexpectedly found that in a switch trolley equivalent case acting was marginally more likely judged to be breaking a rule when this case had been preceded by a footbridge equivalent case than when it had been presented in the first position. No analogous effects of a preceding switch-trolley-like case on judgments in the footbridge-like case were reported in either study.

Finally, Alistair Norcross (2008) describes an interesting order effect which was not observed in an experimental setting but that is nevertheless relevant for the present research, since it follows the pattern of the studies reviewed so far. Namely, he points out that when he asks his students to evaluate the standard switch-trolley dilemma in the first position, the majority judge diverting the trolley to be permissible. However when this dilemma is preceded by the transplant scenario (described above), the proportion of students that judge diverting the trolley to be permissible is considerably lowered.

### 1.2. A Systematic Pattern

We claim that a closer look at the reported findings reveals a systematic pattern: First, only dilemmas which receive a positive rating (in the sense that the proposed action is on average rated as morally right, acceptable, permitted, required, etc. - depending on the specific question asked) when judged in isolation are susceptible to be affected by an order effect. Dilemmas that receive a negative rating (i.e., the proposed action is on average rated as morally wrong, unacceptable, forbidden, etc.) seem to be unaffected. Second, dilemmas that are rated positively are only affected by an order effect if directly preceded by a dilemma that was rated negatively. In this case, the ratings for the dilemmas that are usually rated positively seem to be lowered or, in those cases in which the response format is dichotomous instead of a quantitative scale, the proportion of people that judge the action to be acceptable is lowered.

To illustrate the pattern, let us again consider the first experiment by Petrinovich and O’Neill (1996, Study 2, Forms 1 and 1R). The ratings for action in both Dissection and Scan are negative, implying that on average the proposed actions are regarded as morally wrong. The order of presentation did not affect the judgment of actions proposed in these dilemmas. The first part of our proposed pattern captures these findings by claiming that only judgments of dilemmas that are rated positively are susceptible to be affected. The switch-trolley dilemma constitutes one of such cases (see Hauser et al., 2007, where this is confirmed in a large, multicultural sample). In line with our proposed pattern, a strong order effect was found for this dilemma. Recall that ratings were considerably lower for this scenario in the condition where it had been preceded by Scan (a dilemma that was rated negatively) than in the condition where it had been presented first. This phenomenon is captured by the second part of the proposed pattern, which states that judgments of scenarios that are normally rated positively are affected if directly preceded by a dilemma that was rated negatively.

From the data reported by Petrinovich and O'Neill (1996) and by Lanteri et al. (2008) it is not evident which ratings at the level of individual participants' responses generated the asymmetric results at the level of group means. We propose the following hypothesis about the individual-level responses leading to the described asymmetric order effects: A given subject is most likely to "transfer" his or her initial judgment about a given scenario onto a subsequent scenario if (i) his or her initial judgment of the first scenario is negative, (ii) his or her isolated judgment of the subsequent scenario is expected to be positive, and (iii) both scenarios are perceived as sufficiently similar in morally relevant respects. In what follows, we will clarify each part of this hypothesis and report the results of an experiment designed to test it.

We define the "transfer" of an initial judgment to a subsequent scenario as follows: A judgment is transferred if a participant's rating of the subsequent scenario deviates from the expected rating in the direction of the initial rating, where the expected rating is the average isolated rating of the subsequent scenario by an independent sample of subjects. If, for example, a subject rates the initial scenario with -4 on the Petrinovich and O'Neill (1996) scale and the subsequent scenario with -2, while the average isolated rating of the subsequent scenario by an independent sample of subjects is +2, we regard this as an instance of judgment transfer. If the likelihood of judgment transfer is sufficiently high, the judgments of subsequent scenarios will (on the group level, i.e. the mean rating) be affected by an order effect as defined above.

Our hypothesis is concerned with factors influencing the likelihood of judgment transfer. As reviewed above, previous research suggests that order effects are most likely to occur if a scenario which is rated strongly negatively precedes another scenario which is rated less negatively in isolation, giving rise to the asymmetric pattern we have outlined. This suggests that, at the level of individual responses, judgments are most likely to be transferred if (i) the initial rating is strongly negative, and (ii) the subsequent rating is expected to be positive. Given the results of previous studies we suspect (i) and (ii) to be jointly necessary conditions for order effects to occur. They also account for the asymmetric nature of the order effects shown at the group level.

Part (iii) of our hypothesis, in contrast, is not concerned with the asymmetry of the effect, but rather with its strength. We expect that, given that (i) and (ii) are present, the likelihood of judgment transfer will increase with increasing perceived morally relevant similarity between the subsequent scenarios. It is intuitively plausible that judgments are more readily transferred to similar as opposed to dissimilar scenarios, although determining precisely in what sense two scenarios need to be similar is an intricate matter. As a working definition for the present purpose, we regard two scenarios as similar in morally relevant regards if they differ only on few of the dimensions that have previously been shown to influence people's isolated moral judgment. For instance, the classic bystander and footbridge cases as explored by Lanteri and colleagues (2008) differ on a number of these dimensions, including physical contact (Cushman et al., 2006), personal force (Greene et al., 2009), locus of intervention (Waldmann & Dieterich, 2007) and whether the victim constitutes a means or a side-effect in rescuing the five workers (e.g., Hauser et al., 2007). Thus, according to the working definition provided above, both scenarios can be considered to be relatively dissimilar, since they differ on many of the dimensions which have been shown to affect moral judgment. Part (iii) of our hypothesis therefore predicts that the judgment of the standard trolley scenario will be more strongly affected by order if it is not *immediately* preceded by the footbridge case (i.e., where all of the relevant structural differences are introduced concurrently in the same scenario), but if instead several different scenarios are interposed between them in which these differences are introduced one by one, thus maximizing the perceived similarity within each pair of consecutive scenarios.

We will now describe an experiment designed to test the hypothesis outlined above. We will use several variations of the trolley dilemma, including the two most popular ones, which from now onwards we will refer to as Standard (pressing a switch to redirect a trolley away from a number of people onto a sidetrack where only one person will die) and Push (pushing a person from a bridge in order to save the larger number of people). Three further variations will be inserted between Push and Standard in order to decrease the differences within each pair of consecutive scenarios to a minimum. For all scenarios we will measure the extent to which people agree that the action proposed in each case should be conducted. Following our hypothesis, we predict (i) that order effects will only occur if Push is the first scenario in the sequence, but not if Standard is the first scenario and (ii) that judgments of those scenarios which are most positively rated in isolation (i.e., that receive higher ratings of agreement with the proposed action) will be most strongly affected. Furthermore, due to the gradually introduced differences between consecutive scenarios, we expect (iii) that the judgment of Standard will be more strongly affected by order than in the studies reviewed above. As yet there is no evidence for a change of people’s judgments at a qualitative level. That is, in previous studies investigating order effects, average judgments varied only in degree of acceptability or unacceptability, but judgments were never affected strongly enough to switch from one of these categories to the other. We aim to show that, if the similarity between consecutive scenarios is maximized, the resulting order effects can be strong enough to lead people to regard an action which is considered to be acceptable in independent judgments as inacceptable instead.

## 2. Experiment

### 2.1. Subjects

Fifty participants (35 women) were recruited using the lab in the Psychology department in the University of Göttingen. They were randomly assigned to one of the two experimental conditions. The average age was 23 years (*SD*=2.83).

### 2.2. Materials

We presented participants with a series of five moral dilemmas. Each dilemma or scenario consisted of a brief description of a situation and of an action that could potentially be conducted in each case, accompanied by a diagram depicting the situation schematically (see Figure 1). The initial description of the situational set-up was identical for all scenarios and reads as follows (translated from German):

*On the test ground of a modern railroad property an unmanned speed-train that normally can be remote-controlled got out of control due to a technical defect. This speed-train is heading towards three railroad workers that are maintaining the tracks. Since these workers are wearing a novel hearing protection, they would not notice the speed-train on time and hence would be run over by it. Karl, an employee of the rail track control center, recognizes the upcoming accident. However, it is not possible to stop the train on time any more.*

This introduction was followed by a description of a specific action that Karl could conduct in order to save the three workers on the track. This action was different for each of the five scenarios, but in all cases it involved the death of one innocent stranger (for brief summaries of the action descriptions, see Figure 1; for complete wording of the instructions and the action descriptions, see Appendix). Instructions were included to ensure that participants assumed that the proposed action was the only one available in each case, and that, if taken, it would always lead to the described outcome. We also explicitly stated that participants' task was not to tell us what they would actually do in the described situations, but what Karl should do in terms of morality. Furthermore, participants were instructed to assume that no matter which decision Karl made, he would not be prosecuted. The number of potential victims (3 vs. 1) was kept constant across scenarios.

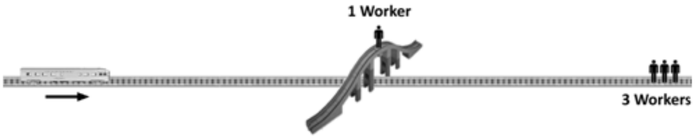
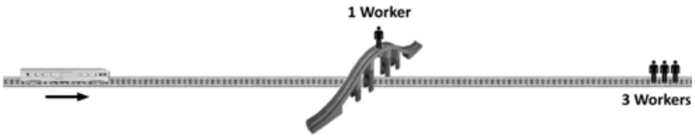
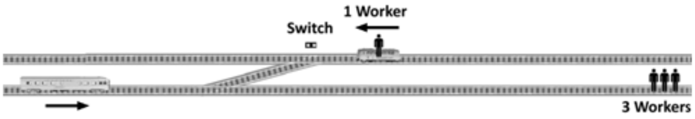
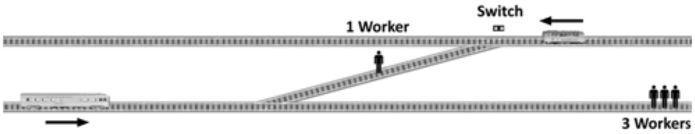
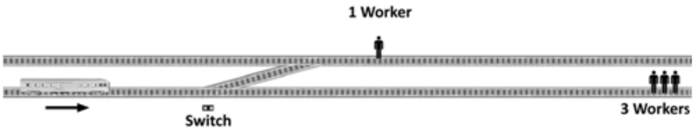
| Scenario | Proposed action | Illustration |
|---|---|---|
| Push | Push the large person from the bridge to stop the train |  |
| Trap | Push a button that will open a trap door that will let the person on top of the bridge fall and stop the train |  |
| Redirect | Redirect a train with a person inside that is on a parallel track onto the main track to stop the train |  |
| Run Over | Redirect an empty train that is on a parallel track onto the main track to stop the train, running over a person that is on the connecting track |  |
| Standard | Press a switch that will redirect the train that is out of control to a parallel track where one person will be run over |  |

**Figure 1** Summaries of the actions proposed in the five dilemmas with corresponding schematic illustrations.

In order to establish a baseline of agreement with the proposed action in each of the five different scenarios we conducted a pilot study using an independent sample which consisted of 100 students of the University of Göttingen. The participants were individually approached on campus and presented with only one scenario each (between-subjects design with five scenario conditions, each *n*=20). In all cases the question was articulated as follows: "Should Karl perform the proposed action?" Participants were asked to indicate their answer on a scale ranging from 1 to 6, where 1 was "not at all" and 6 was "absolutely" (cf. Kahane & Shackel, 2010, for a discussion about which question wordings are suited for asking moral questions). Table 1 shows the average ratings assigned to each scenario.

**Table 1** Mean ratings (standard deviations) of agreement and percentage of subjects disagreeing with the proposed action in the five scenarios when evaluated independently.

|  | Scenario (each *n*=20) |  |  |  |  |
|---|---|---|---|---|---|
| Measure | Push | Trap | Redirect | Run Over | Standard |
| Mean Rating (SD) | 1.95 (1.76) | 3.4 (1.76) | 4.15 (1.42) | 4.4 (1.14) | 4.45 (1.15) |
| % Disagreement | 80 | 40 | 30 | 10 | 15 |

*Note.* % Disagreement is the percentage of subjects who gave a rating ≤ 3 on a scale ranging from 1 to 6.

Based on these results, which are largely in line with expectations derived from previous research, we ordered the five scenarios as follows: Push, Trap, Redirect, Run Over, Standard. In this specific order, the structural differences within each pair of consecutive scenarios are minimized and the level of agreement with the proposed action in each case is steadily increasing.

### 2.3. Procedure

The experiment was run individually on computers. First, the instructions were presented on the screen, and subsequently the five different scenarios were presented in a sequence. After each scenario participants were requested to rate, on a scale from 1 to 6, whether Karl should act in the proposed way or not, using the same scale and question wording as in the pilot study described above. Half of the participants were presented with the sequence of dilemmas in increasing order of agreeability (Least Agreeable First condition, beginning with Push and ending with Standard), and the other half were presented with the sequence of dilemmas in the reverse order (Most Agreeable First condition, beginning with Standard and ending with Push). The task was computerized to ensure that participants evaluated one dilemma at a time and to avoid changes in ratings previously given. This format guaranteed that each new dilemma would be judged before the following one was presented, and avoided the possibility for participants to withhold their judgment until the end of the sequence.

**2.4. Results**

First, we tested the existence of an asymmetrical order effect at the group level. Specifically, to test whether the pattern of ratings of the dilemmas differed between the two orders of presentation, a 2 × 5 mixed analysis of variance (ANOVA) was conducted, where the first factor was the Order of presentation of dilemmas (Least Agreeable First vs. Most Agreeable First, between-subjects) and the second factor was the Scenario judged (within-subjects). The results are shown in Table 2 and Figure 2. The ANOVA revealed a main effect of Order, where the average ratings were significantly lower in Least Agreeable First than in Most Agreeable First ($F_{1,48}$=8.03, $p$<.01, $\eta^2$=.14) and a main effect of Scenario ($F_{4,192}$=23.44, $p$<.001, $\eta^2$=.33), indicating that, as expected, the scenarios received different average agreeability ratings when averaged across the orders of presentation. Crucially, the ANOVA also revealed a significant Order × Scenario interaction ($F_{4,192}$=8.2, $p$<.001, $\eta^2$=.15), suggesting the presence of a strong asymmetrical order effect, in line with our predictions.

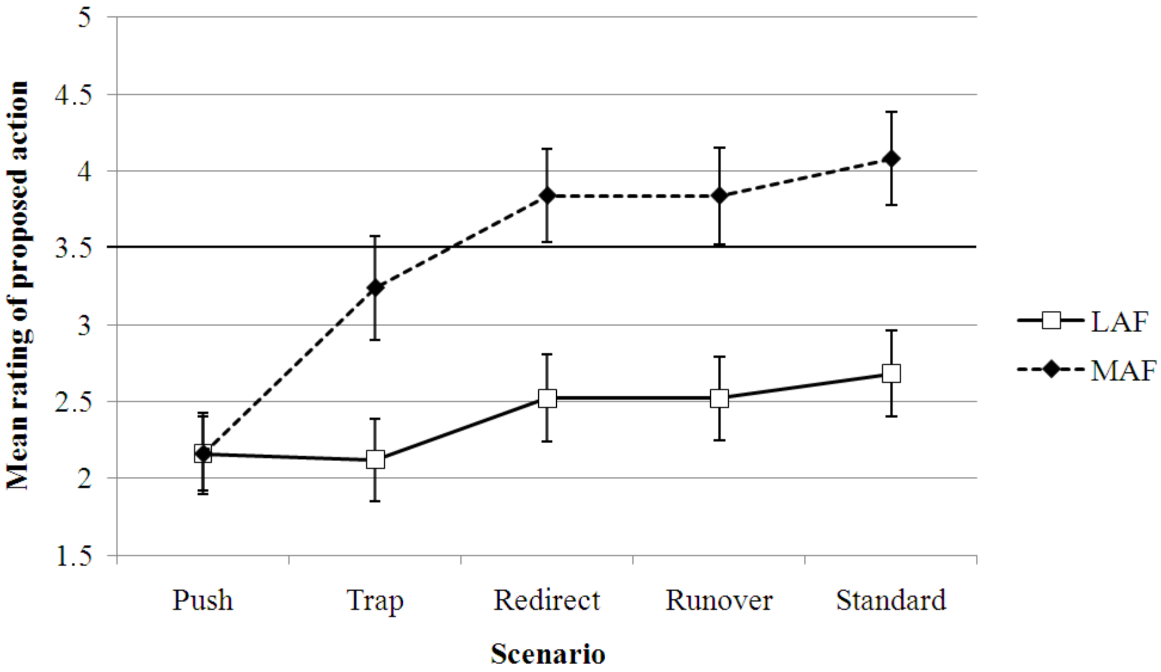


**Figure 2** Mean ratings of agreement with the proposed action in the five scenarios when evaluated sequentially, as a function of the order of presentation. Error bars indicate SEM. The bold line at *y*=3.5 indicates the line between average agreement and disagreement with the proposed action. MAF = Most Agreeable First; LAF = Least Agreeable First.

**Table 2** Mean ratings (standard deviations) of agreement and percentage of subjects disagreeing with the proposed action in the five scenarios evaluated sequentially, as a function of the order of presentation.

| Order | Scenario | | | | |
|---|---|---|---|---|---|
| Condition | Push | Trap | Redirect | Run Over | Standard |
| | | | Mean ratings (SD) | | |
| MAF (*n*=25) | 2.16 (1.21) | 3.24 (1.69) | 3.84 (1.52) | 3.84 (1.57) | 4.08 (1.53) |
| LAF (*n*=25) | 2.16 (1.31) | 2.12 (1.33) | 2.52 (1.42) | 2.52 (1.36) | 2.68 (1.41) |
| | | | % Disagreement | | |
| MAF (*n*=25) | 76 | 52 | 40 | 32 | 32 |
| LAF (*n*=25) | 80 | 80 | 72 | 72 | 68 |

*Note*. % Disagreement is the percentage of subjects who gave a rating ≤ 3 on a scale ranging from 1 to 6. MAF = Most Agreeable First. LAF = Least Agreeable First.

In order to test our prediction more specifically, we conducted planned contrasts involving Standard and Push as prototypical examples of scenarios typically eliciting high vs. low agreeability ratings, respectively. Focusing on Standard it can be observed that the average rating for this scenario varied considerably depending on the position in which it appeared. When Standard had been evaluated first (Most Agreeable First condition), the average rating was 4.08 (*SD*=1.53), while the average was only 2.68 (*SD*=1.41) when it appeared at the end of the sequence (Least Agreeable First condition). This difference was significant, $t_{48}$=3.37, *p*<.01, $\eta^2$=.19. In contrast, the average rating for the Push scenario was the same in both orders (*M*=2.16, *SD*=1.21 in Most Agreeable First; *M*=2.16, *SD*=1.31 in Least Agreeable First). Additionally, we computed the within-subjects differences between the ratings for the Standard and the Push scenarios by subtracting the rating for Push from the rating for Standard for each subject. The average of these within-subjects differences was significantly larger in the Most Agreeable First condition (*M*=1.92, *SD*=1.47) than in the Least Agreeable First condition (*M*=.52, *SD*=1.08; $t_{48}$=3.83, *p*<.001, $\eta^2$=.23). Taken together, these results strongly support our prediction of an asymmetrical order effect.

It is worth noting that the difference between the ratings for the Standard scenario in the two order conditions is relevant not only in quantitative but also in qualitative terms: Treating ratings below 3.5 as “rather disagree” and above 3.5 as “rather agree”, the majority of participants’ ratings in Least Agreeable First would fall into the first category (18 out of 25; 72%) whereas the majority of participants’ ratings in Most Agreeable First would fall into the second (18 out of 25; 72%; see also section 3.3.2. for further discussion of this point). This difference was significant ($\chi^2_{1, N=50}$=9.68, *p*<0.01). The same is true for Run Over ($\chi^2_{1, N=50}$=8.01, *p*<0.01), Redirect ($\chi^2_{1, N=50}$=5.20, *p*<0.05), and Trap ($\chi^2_{1, N= 50}$=4.37, *p*<0.05), but not for Push ($\chi^2_{1, N=50}$=0.12, *p*=0.73).

We went on to take a closer look at the results at the level of individual participants’ responses. In particular, we explored the data treating each individual’s ratings as a set of binary choices (i.e., treating ratings ≤ 3 as indication of disagreement with the proposed action and ratings ≥ 4 as indication of agreement) and observed that if an agent tended to disagree with an action, this judgment was in most cases “transferred” onto the judgment of the action in the next scenario. That is, an action that would have likely been rated positively if it had been judged independently (as indicated by the independent ratings of our pilot study) now received a negative rating by this particular participant who had rated the proposed action in the preceding scenario negatively. However, positive ratings did not affect the ratings of the next action (by changing them into positive ones) if this action was normally rated as morally wrong in isolation. For instance, in the Least Agreeable First condition, 20 participants disagreed with the proposed action in Push. Out of these 20 “initial disagree-ers” only three switched to agreement over the whole sequence, resulting in 17 participants who disagreed with the proposed action in Standard. In contrast, when participants started with Standard — where 17 participants agreed with the proposed action—eleven of these “initial agree-ers” switched to disagreement on the way to Push, resulting in only six positive ratings for the action proposed in this case. These findings are in line with part (ii) of our hypothesis, as they suggest that order effects occur if the number of participants disagreeing with the proposed action in the preceding scenario is relatively high, while the number of participants that would disagree with the action in an isolated rating of the subsequent scenario is relatively low. This “excess of disagreements” is transferred to the subsequent scenario, leading to an order effect.

### 2.5. Discussion

In sum, the data were largely in line with the pattern we outlined above: First, Push but not Standard was potent to affect subsequent ratings. Recall that part (i) of our hypothesis stated that judgments would be most likely transferred if the initial rating was strongly negative. Accordingly, we found order effects exclusively in the Least Agreeable First condition, while the ratings in the Most Agreeable First condition neatly matched the ratings for the same scenarios when presented independently. Second, the more positively an action was rated when judged independently, the larger was its susceptibility to be affected by order. Ratings for actions that clearly received a positive rating when judged independently (i.e., Standard, Run Over, Redirect) differed significantly between the two order conditions. In contrast, ratings for the action in Push (which was clearly rated negatively when judged independently) did not differ in the two conditions. The fact that the judgment of Trap (which was rated slightly negatively in isolation) was also affected by order seems to suggest that a scenario's *relative* agreeability compared to the preceding scenario is crucial for whether or not its judgments are susceptible to be affected by order, rather than whether it is judged positively or negatively in isolation. In accordance to that, part (ii) of our hypothesis should be slightly modified to reflect this finding: It seems that the expected isolated judgment of the subsequent scenario does not have to be *positive* in order to be susceptible to be affected by order, but merely *less negative* than the preceding judgment.

Finally, we found evidence consistent with part (iii) of our hypothesis, which stated that order effects would be especially pronounced when subsequent scenarios are perceived as similar in morally relevant respects. As mentioned before, previous studies found the acceptability of acting in standard trolley cases to be significantly decreased by a preceding push case, but never before was the mean rating tilted to a value in the negative direction. In contrast, our results indicate that our participants approved of acting in Standard if it was presented first, whereas they clearly disapproved of acting if it was presented in the last position. We believe that the reason for this exceptionally strong effect lies in the limited change of content and morally relevant structure between each pair of consecutive scenarios. The large similarity between each scenario and its predecessor might have made it difficult for participants to perceive any scenario as fundamentally different from the (intuitively unacceptable) initial Push scenario. In contrast, if the Standard case had followed immediately on the Push case (as was the case in most previous studies), the existing structural and content differences between both scenarios might have led participants to perceive them as fundamentally different. As a consequence, judgments regarding the action proposed in each case might have been different, and order effects at the group level would have been less pronounced. Future work should test this claim directly by experimentally manipulating the similarity between scenarios in one and the same study.

While the exceptional *strength* of the present order effect might be related to a notion of perceived similarity, the question of which psychological mechanisms account for the *asymmetry* of the effect has not yet been treated. This question has not been addressed by our present experiment, which aimed to document boundary conditions for order effects exclusively in terms of observable reactions to stimuli but was not designed to identify cognitive processes leading to those reactions. However, since we believe this is an important question, we will advance some speculations about potential psychological mediators of order effects in the following section. Finally we will outline implications of the present findings for both descriptive and normative theories of morality.

## 3. General Discussion

### 3.1. Psychological Mechanisms

It is not possible to determine from our data which psychological mechanisms best account for the described asymmetry. Petrinovich and O'Neill (1996) concluded that "the initial 'set' regarding strength of agreement or disagreement influenced the strength of succeeding responses" (p. 160). However, the assumption that order effects are merely grounded in the *intensity* of the initial judgment, regardless of its valence, cannot account for the observed asymmetry.

A possible explanation of our results is the existence of a difference in the urge to justify prohibitions and permissions. For instance, when we prohibit a child to play with knives we automatically think of a justification for this prohibition. Prohibitions seem to call for a justification. In contrast, we do not think about a justification regarding most things we permit. We do not feel an urge to explain or justify to someone why he or she is allowed to breathe, for example. Normally, we only justify or explain permissions when a prohibition is the default case. For instance, we might explain to a child that in case of emergency an ambulance is permitted to drive over red lights although this is usually prohibited. Applying this line of reasoning to the asymmetric pattern obtained in our results, one could argue that those participants who prohibited the proposed action in Push were, may it be consciously or unconsciously, thinking about a justification for their prohibition. If they reached a rough justification like "You must not kill an innocent person", they might have kept this principle in mind and applied it to the remaining scenarios. Since in all scenarios an innocent person had to be killed in order to rescue the three persons, the application of this principle might have led them to judge all proposed actions as prohibited. In contrast, when they started with a scenario in which most participants judged the proposed action as permissible, no or only little effort might have been invested in justifying this judgment and, therefore, no such justification would have been applied to the remaining scenarios.

Another potential explanatory mechanism involves the differential emotional engagement elicited by different scenarios. As Greene and his collaborators have shown, footbridge-like dilemmas are much more likely to activate brain regions associated with emotional processing than switch-like cases (Greene et al., 2001, 2004). Lanteri and colleagues (2008) attempted to explain their results following Green and colleagues' (2001, 2004; Greene & Haidt, 2002) notion that scenarios containing “personal” moral features such as Push trigger hard-wired moral emotions to a greater extent than “impersonal” scenarios like Standard. Instead, responses to scenarios like Standard would be driven to a greater extent by reflective moral reasoning. Lanteri and colleagues suggested that judgments concerning scenarios that activate moral emotions should be less variable than judgments that do not activate such emotions. The rationale is that such “instinctive” emotions would exert a strong power on moral judgment for the majority of people. They further claim that once such emotions are activated, they can affect reflective reasoning, while the reverse (i.e., emotions being affected by reflective reasoning) would not hold. It should be noted, however, that the authors propose this explanation as the most plausible one to account for their findings, while admitting that further studies gathering neuroscientific data should be conducted to test it directly.

Both of the mechanisms outlined above (i.e., the existence of a difference in the urge to justify prohibitions and permissions, and a differential emotional engagement) could also be jointly responsible for the order effects observed in our study. Maybe, in order to be effectively applied to subsequent scenarios, a justifying principle needs to be backed up by consonant moral emotions (see also Nichols, 2002). The lack of order effects in the Most Agreeable First condition would then be due to an initial lack of emotional engagement to interfere with subsequent responses. At the same time, the potency of a scenario to elicit strong emotional reactions might make this scenario immune against the impact of principles transferred from previous cases. The strong, emotionally backed intuition that acting is wrong in such cases might serve as a stable criterion against which the permissibility of the action can be judged. Such an emotional anchor might be missing in switch-like scenarios, leading judgments regarding such scenarios to be much more malleable by incidental factors such as the order of presentation.

Unfortunately, the present evidence is not suited to determine the role of different potential explanatory mechanisms. Further research is needed in order to yield both theoretical insights and practical implications for the design and conduction of investigations on moral psychology.

### 3.2. Implications for Descriptive Moral Theories in Psychology and Practical Implications

An important goal of descriptive moral theories is to provide a comprehensive explanation of an average person’s moral judgments. A potential source of variance in moral judgments which has received considerable attention is the structural set-up of the situations in question (e.g., the existence of physical contact with the victim, Cushman et al., 2006; whether she constitutes a means or a side-effect in rescuing the larger number of workers, Hauser et al., 2007, etc.). However, often the effects generated by the manipulation of these factors are fairly small, accounting only for a very limited amount of the total variance in moral judgments and thus leaving a good portion of interindividual differences unexplained. It therefore seems necessary to take into account psychological mechanisms that influence how a given situational set-up is apprehended, represented, and evaluated. In our experiment, for example, previously judged scenarios sometimes served as a reference point which informed the judgment of subsequent scenarios. This reference is exogenous to the subsequent scenarios, yet indispensable to predict and explain the reactions towards them.

According to our results an initial judgment about a given moral dilemma can be transferred to the judgment of a subsequent dilemma under certain conditions. Specifically, such order effects can be expected to be especially large under conditions which strongly suggest the adequacy of transferring a certain judgment from one scenario to the next, such as a within-subjects design where several scenarios similar in structure or content need to be judged sequentially. As our results suggest, extreme caution is required if responses generated under such conditions are to be attributed exclusively to properties of the scenarios themselves. As we have demonstrated, order effects can occur in a systematic and predictable fashion and thus should not be dismissed as odd but meaningless incidents.

Finally, we believe that between-scenario order effects might also play a role outside the laboratory. We acknowledge that the nature of the specific dilemmas used in the present study might raise concerns regarding the generalizability of our findings to situations in which laymen normally reason about moral issues. In trolley dilemmas the parameters are highly specified (e.g., a full certainty exists about the set of possible actions and their consequences; Gigerenzer 2008), which is an uncommon feature in moral dilemmas we might encounter in daily life. Additionally, we have acknowledged that the similarity between the dilemmas used in our study might have particularly encouraged the transference of standards of evaluation between scenarios. Nevertheless, we believe that the delimitation of the conditions under which order effects occur in experimental settings can contribute to understand how such effects may shape moral judgments in ecological settings. For instance, the specification of conditions which may favour order effects could guide the design of public opinion polls or surveys that consecutively gauge responses to several (moral) issues. This would expand previous research in other contexts showing that such instruments can be highly sensitive to effects of question positioning (e.g., Benton & Daly, 1991; DeMoranville & Bienstock, 2003). The extent and boundaries of this influence should be investigated in future research by using more ecologically valid dilemmas which are more resembling of those that one might encounter in everyday life discussions.

### 3.3. Implications for Normative Moral Theories in Philosophy

Do our findings support the claim that moral judgment can be unstable and subject to biases (Sinnott-Armstrong, 2008; Swain, Alexander, & Weinberg, 2008)? Some philosophers (e.g., Weinberg, Nichols, & Stich, 2001) argue that results of empirical studies that surveyed people’s intuitions about various subject matters in philosophy show that the traditional method of using intuition as evidence is mistaken. The reason is that in these studies the participants’ intuitions vary according to seemingly irrelevant factors such as cultural or educational background. In this final section, we will discuss whether the results of our study support the idea that our moral intuitions are flawed in this sense. In order to answer this question we will first discuss what kind of studies and results would be necessary to make such a strong claim reasonable. In our opinion, it is not sufficient to present statistically significant results of a study investigating moral judgments when a seemingly irrelevant factor was manipulated. Instead, we will propose three criteria that have to be met in order for this claim to be made, and will discuss whether our study fulfills these criteria.

### 3.3.1. The irrelevance of the manipulated factor

Our first criterion postulates that the manipulations that yield different ratings have to address morally irrelevant factors. This may sound trivial at a first glance. However, sometimes a factor is hastily assumed to be morally irrelevant although acknowledged philosophers argue quite differently (see, for example, Kamm's, 2007, discussion about the moral significance of distance). We consider a factor to be irrelevant with regards to moral judgment if virtually no one within the philosophical community believes this factor to be morally relevant. Framing effects like different wording or context (e.g., ordering) are typical examples of effects that are caused by factors that are commonly believed to be morally irrelevant (e.g., Kern & Chugh, 2009; Nadelhoffer & Feltz, 2008; Petrinovich & O'Neill, 1996; Tversky & Kahneman, 1981). Thus, we believe that the first criterion is met in the present study.[1]

### 3.3.2 Significant differences at the nominal scale level

As discussed at the beginning of this paper, several studies have found a number of factors that influence moral judgment. However, generally these factors have only been found to affect judgment in quantitative terms, that is, the means of a moral judgment might differ significantly between two conditions but the individual ratings (when treated as binary choices) show the same or a similar pattern. Since moral questions often take the form "Is it permissible to do X?" or "Ought X to be done?" we are mainly interested in a Yes/No-answer. The question of the degree to which we believe in our answer is subordinated. Therefore, we believe it would be overhasty to claim that mere quantitative differences within one qualitative category show that our intuitions are seriously flawed. In order to support this claim, consider the following case: Two groups of participants are presented with the Standard Trolley scenario with one group receiving the normal version and the other group receiving a version in which the death of the innocent bystander is described in a very brutal way (if you consider this to be a morally relevant factor, then think instead of a factor you consider to be irrelevant). Let us assume that participants have to give their answer on the same scale as participants in our study. For the sake of the argument, let us also assume that all participants in the group that was presented with the original scenario answered with a six (meaning they absolutely agree with the proposed action) while the participants in the "brutal" group all answered with a four (rather agree with the proposed action). At an interval scale level the difference between the two groups would be highly significant. However, all participants decided in favour of the proposed action. In such a case, should we claim that participants' intuitions were flawed (assuming the way of describing the death of the innocent bystander is morally irrelevant)? We do not think so (but see Sinnott-Armstrong, 2008, who takes the quantitative differences found in the study of Petrinovich & O'Neill, 1996, as sufficient evidence for the claim that our moral intuitions are subject to morally irrelevant factors). Instead, we suggest that the differences found ought to be significant when the ratings are treated as nominal data (that is, the number of participants who decide against or in favour of the action differ significantly between the two conditions).

Let us now consider the ratings of the Standard dilemma in this study. Since we used a six point scale without a neutral point it is plausible to interpret ratings equal or below three as "rather disagree" with the proposed action and ratings equal or above four as "rather agree" with the proposed action. For the Most Agreeable First condition the average rating for Standard was 4.08 and the majority of participants (18 out of 25) agreed with the potential action of redirecting the trolley (i.e., reported a rating of 4, 5, or 6). In contrast, the mean rating in Least Agreeable First was 2.68 and the majority disagreed with redirecting the trolley in Standard (18 out of 25). These ratings differ in both quantitative and qualitative terms. The only exceptions were two subjects in this condition who showed the same pattern as found in the independent ratings, and whose judgments were, accordingly, not affected by the order of presentation (the remaining five subjects decided in a consequentialist manner throughout).

The same is true for Run Over, Redirect, and Trap. The $\chi^2$-tests (testing nominal data) for these scenarios were significant. However, Push seemed to be especially robust since the average rating remained exactly the same in both order conditions. In summary, in four out of five scenarios moral intuitions were significantly influenced by a morally irrelevant factor (order).

### 3.3.3. Generalizability

With generalizability we refer to the question of whether the effects found in an experiment are limited to a controlled setting in a lab or also likely to occur in a natural setting as well. If the setting under which such effects occur is very artificial the findings might not be relevant for the moral reasoning of laypeople or professionals. If, for example, an experiment shows that inflicting strong pain on participants can change their moral judgments then these results would not be really relevant to the question of whether our moral intuitions are generally flawed because such extraordinary circumstances are usually not present when we make moral judgments.

In this context, with "natural setting" we refer to settings in which people who are interested in moral questions, especially philosophers, normally consider these moral questions. The setting of our experiment is partly natural, partly artificial in this regard. It is natural insofar as it is often the case that one judges two or more scenarios after each other. Especially philosophers often discuss many (artificial) scenarios in a row (for example, Unger, 1996; Kamm, 2007). Hence, with regards to this feature our setting was not artificial.

However, two features of the study are rather artificial: First, participants had to judge each proposed action right after having read the description and before being presented with the next scenario. Instead, when philosophers judge such scenarios there is not such a constraint and one can consider several scenarios in a row without being forced to judge each of them right away. We must acknowledge that the ratings in our study might have been different without the constraint that was imposed. This question is open to further research.

Second, we tested only a limited range of scenario arrangements. As noted above, the level of agreeableness was steadily increasing in one condition and decreasing in the other one. This feature of the experiment might have also influenced the process of moral reasoning during the experiment and makes it potentially different from moral reasoning under normal circumstances.

As noted above, another feature that might seem artificial at first glance is that all scenarios used in this study are considerably similar to each other. However, we would argue that this is not really an artificial feature in a philosophical context: when a philosopher is searching for or evaluating moral principles, he or she often considers several cases which are constructed in a way that all possible pairs of scenarios only differ with respect to a specific factor such as, for example, the way in which the potential victim would die. This contributes to enhance the level of control and to constrain the variables under study. Hence, we can consider that with regards to this feature our experiment was not artificial.

Finally, it should also be noted that participants in our study were exclusively laymen. No professional philosophers were included in the sample. An objection against the significance of judgments collected from laymen is called the expertise defence (Weinberg, Gonnerman, Buckner & Alexander, 2010). Applied to our study, a supporter of this view could raise the following objection: The response pattern of professional philosophers would look different from the one of laymen because laymen have no training in the use of intuitions. Due to their lack of training, it is likely that the responses given by laymen are more erroneous than those of professional philosophers and, furthermore, might not even qualify as intuitions in the sense usually attributed to this concept by philosophers (Hales, 2006; Kauppinen, 2007; Ludwig, 2007; Sosa, 2006; Williamson, 2005). According to this view, philosophers might not be affected by the order of presentation of dilemmas, and, therefore, it would not be justified to generalize from our findings. However, we agree with Weinberg and colleagues (2010), who suggested that, absent empirical evidence to the contrary, there are good reasons to assume that philosophers and folk will in most cases display the same pattern of intuitions.

In summary, it can be stated that in 4 out of 5 scenarios participants' intuitions were seriously flawed. In these scenarios participants' judgments regarding whether the proposed action should be conducted or not were strongly influenced by a morally irrelevant factor, namely the order in which the scenarios were presented. Although we have argued that these results indicate that moral intuitions can strongly be influenced by morally irrelevant features, it would be overhasty to claim that our moral intuitions are generally flawed. As noted above, this is partly due to some limitations of our experiment, such as the fact that some features of the experiment were rather artificial. Furthermore, even if all criteria had been met there might still be room to avoid the conclusion that intuitions should not be used as evidence (see Burkhard & Gertken, submitted)

### 3.4. Summary

In this article, we have demonstrated that order effects can have a substantial influence on judgments of actions in moral dilemmas. First, we have outlined an asymmetrical pattern underlying the data reported in previous research. A central aspect of this pattern is that only judgments of actions that are normally regarded as morally acceptable are susceptible to be affected by the order of presentation, and this is only the case if the dilemma is immediately preceded by a dilemma in which the proposed action was judged negatively. Second, we have conducted an empirical study which has largely confirmed this pattern and also showed that order effects can be strong enough to lead to qualitative judgment reversals if the consecutive scenarios are very similar.

In sum, our study provides valid evidence that people's moral intuitions can sometimes be influenced by morally irrelevant factors in a relevant sense. Additionally, it contributes to highlight the relevance of considering order effects as important sources of variance in descriptive theories of moral judgment. Due to the important theoretical and practical implications of these findings, future research should aim to outline more precisely the determinants, explanatory mechanisms, and boundary conditions of order effects in moral judgment.

## Acknowledgements

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre ‘Evolution of Social Behaviour’, University of Göttingen (funded by the German Initiative of Excellence). We thank Michael Waldmann and two anonymous reviewers for helpful comments, Johanna Kirchhoff for data collection, and Gunda Johannes for proofreading.

## Notes

[1]

We thank an anonymous reviewer for pointing out the possibility that theories giving emotions a central role in moral judgments (e.g., Prinz, 2008) could suggest that order of presentation is not a morally irrelevant factor. Reading a normally negatively judged scenario first may generate a negative feeling that carries over to subsequent scenarios. Such feeling is not generated by reading a normally positively judged scenario first. If emotion is seen as a defining constituent of moral judgment and order of presentation affects emotional reactions, then order of presentation might be argued to be morally relevant. However, as we have pointed out, we believe that it is unlikely that a member of the philosophical community would actually argue in this way

**References**

Austin, V. D., Ruble, D. N., & Trabasso, T. (1977). Recall and order effects as factors in children's moral judgments. *Child Development*, *48*(2), 470–474.

Benton, J. E. & Daly, J. L. (1991). A question order effect in a local government survey. *Public Opinion Quartely*, *55*, 640-642.

Burkhard, A. & Gertken, J. (submitted). Empirically based objections to the reliability of moral intuitions.

Constabile, K. A., & Klein, S. B. (2005). Finishing strong: Recency effects in juror judgments. *Basic and Applied Social Psychology*, *27*, 47–58.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment. *Psychological Science*, *17*, 1082–1089.

DeMoranville, C. W. & Bienstock, C. C. (2003). Question order effects in measuring service quality. *International Journal of Research in Marketing*, *20*, 217-231.

Feldman, N. S., Klosson, E. C., Parsons, J. E., Rholes, W. S., & Ruble, D. N. (1976). Order of information presentation and children's moral judgments. *Child Development*, *47*, 556–559.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 5-15.

Gigerenzer, G. (2008). Moral intuition = Fast and frugal heuristics?. In W. Sinnot-Armstrong (Ed.). *Moral psychology. Volume 2: The cognitive science of morality: Intuition and diversity* (pp. 1-26). Cambridge: MIT Press.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*, 517-523.

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., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144-1154.

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*, 389-400.

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.

Hales, S. D. (2006). *Relativism and the foundations of philosophy*. Cambridge: MIT Press.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Hauser, M., Cushman, F., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*, 1-21.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgment. *Mind and Language, 25*, 561-582.

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm.* Oxford: Oxford University Press.

Kern, M. C. & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science*, *20*, 378-384.

Kerstholt, J. H., & Jackson, J. L. (1998). Judicial decision making: Order of evidence presentation and availability of background information. *Applied Cognitive Psychology*, *12*, 445–454.

Kauppinen, A. (2007). The rise and fall of experimental philosophy. *Philosophical Explorations*, *10*, 95-118.

Lanteri, A., Chelini, C. & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics, 83,* 789-804.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*, 273-286.

Ludwig, K. (2007). The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31*, 128.

Mikhail, J. (2000). *Rawls' linguistic analogy: A study of the „generative grammar" model of moral theory described by John Rawls in "A theory of justice."* Doctoral dissertation, Cornell University.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*, 143-152.

Nadelhoffer, T., & Feltz, A. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics*, *1*, 133-144.

Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*, 221-236.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*, 530-542.

Norcross, A. (2008). Off her trolley? Frances Kamm and the metaphysics of morality. *Utilitas, 20*, 65-80.

Otsuka, M. (2008). Double effect, triple effect and the trolley problem: Squaring the circle in looping cases. *Utilitas, 20*, 92-110.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*, 145-171.

Prinz, J. J. (2008). *The emotional construction of morals*. New York: Oxford University Press.

Sinnott-Armstrong, W. (2008). Framing moral intuitions. In W. Sinnot-Armstrong (Ed.). *Moral psychology. Volume 2: The cognitive science of morality: Intuition and diversity* (pp. 47-76). Cambridge: MIT Press.

Sosa, E. (2006). Experimental philosophy and philosophical intuition. *Philosophical Studies, 132*, 99-107.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on Truetemp. *Philosophy and Phenomenological Research*, *76*, 138–155.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*, 204-217.

Thomson, J. J. (1986). The trolley problem. In J. J. Thomson, *Rights, restitution, and risk. Essays in moral theory* (pp. 94-116). Cambridge: Harvard University Press.

Thomson, J. J. (1990). *The realm of rights*. Cambridge: Harvard University Press.

Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs*, *36,* 359-374.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *30,* 453-458.

Unger, P. K. (1996). *Living high and letting die: Our illusion of innocence*. New York: Oxford University Press.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science, 18,* 247-253.

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, *23*, 331.

Weinberg, J., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, 29, 429–460.

Williamson, T. (2005). Armchair philosophy, metaphysical modality and counterfactual thinking. *Proceedings of the Aristotelian Society, 105,* 1-23.

## Appendix

Instruction:

Dear participant,

Thank you for taking part in this study!

On the following pages we will present you with several situations in which a person has to make a decision.

Please try to understand the situations as well as possible. Your task will be to judge what the person in this situation should do (in terms of morality).

Please assume that in the situation everything happens as described (even if you think that the scenario - or parts of it - are completely unrealistic). That implies that the mentioned options are the only ones available.

You can and should also assume that there are no general rules for those kinds of situations and that the person is not going to be prosecuted for his decision.

After having understood the situation, you will be asked about your moral judgment.

We guarantee you that the collected data will be made anonymous and only be used in the study at hand.

Scenario Descriptions:

Initial Situation (*same for all scenarios*):

On the test ground of a modern railroad property an unmanned speed-train that normally can be remote-controlled got out of control due to a technical defect. This speed-train is heading towards three railroad workers that are maintaining the tracks. Since these workers are wearing a novel hearing protection, they would not notice the speed-train on time and hence would be run over by it. Carl, an employee of the rail track control center, recognizes the imminent accident. However, it is not possible to stop the train on time any more.

Push:

There is only one possibility to avoid the death of the three workers: Carl could run to a nearby bridge on which a heavy worker is standing and push this worker from the bridge. Thereby this worker would fall on the tracks and collide with the speed-train. Due to the collision with the heavy worker (Carl would not be heavy enough to stop the train) the speed-train would stop before it reaches the three workers. The heavy worker would lose his life due to the collision.

Trap:

There is only one possibility to avoid the death of the three workers: Carl could push a button that would open a trap door and thereby causing a heavy worker on top of the bridge to fall on the tracks.

The speed-train would collide with the heavy worker and be stopped before it reaches the three workers. The heavy worker would lose his life due to the collision.

Redirect:

There is only one possibility to avoid the death of the three workers: Carl could throw the switch and thereby redirect a train carrying one worker from a parallel track onto the main track.

The speed-train would collide with this train and be stopped before it reaches the three workers. The one worker on the train would lose his life due to the collision.

Run Over:

There is only one possibility to avoid the death of the three workers: Carl could throw the switch and thereby redirect an empty train from a parallel track onto the main track.

The speed-train would collide with this train and be stopped before it reaches the three workers. On its way to the main track the empty train would run over one worker (also wearing the novel hearing protection). The one worker would lose his life due to the collision.

Standard:

There is only one possibility to avoid the death of the three workers: Carl could throw the switch and thereby redirect the speed-train from the main track onto a parallel track before it reaches the three workers.

On the parallel track the speed train would run over one worker (also wearing the novel hearing protection). The one worker would lose his life due to the collision.