Order Effects in Moral Judgment
Searching for an Explanation

Alex Wiegmann (alex.wiegmann@psych.uni-goettingen.de)
Department of Psychology, Gosslerstr. 14
University of Göttingen, Germany

Yasmina Okan (yokan@ugr.es)
Department of Experimental Psychology, University of Granada
Campus Universitario de la Cartuja s/n, 18071, Granada, Spain

Abstract

Research on moral judgment has shown that the order in which dilemmas are presented to subjects often has a strong influence on their judgment. However, the psychological mechanisms underlying order effects are still opaque. In this paper we aimed to isolate the features that a scenario must exhibit in order to influence judgment of subsequent scenarios. For this enterprise, we identified several features from a scenario known to cause order effects, and tested which of these features are necessary to influence subsequent scenarios. Although we still do not have a full understanding of what causes order effects, we made significant progress towards this aim. In five experiments we ruled out some promising explanations such as order effects being driven by an emotional activation linked to the first scenario. Instead, we found order effects to depend on whether the scenario being influenced and its preceding scenario share rather subtle structural similarities.

Keywords: Order effects; moral judgment; trolley dilemma.

Introduction

Imagine one group of subjects is presented with two moral dilemmas, A and B, one after the other. For each of these scenarios subjects have to make a judgment concerning which of two different hypothetical actions should be taken by the agent in each case. In both dilemmas, the life of people is at stake. Imagine a second group of subjects is presented with the same task, the only difference being the order in which the two scenarios are presented. From a normative perspective it seems clear that the order of presentation should not influence subjects' judgments. However, a number of studies have shown that the order of presentation actually influences judgments. Moreover, the impact of the order of presentation is often stronger than that of factors that are generally considered to influence moral judgments (e.g., the existence of physical contact with the potential victim in the scenarios; Wiegmann, Okan, & Nagel, 2012) Interestingly, not only lay people are susceptible to order effects but also professional philosophers (Schwitzgebel & Cushman, 2012).

In the paper at hand we present five experiments aiming to identify the factors causing order effects. The question guiding the experiments that will be reported is: Which are the features of a scenario known to cause order effects that enable it to influence other scenarios?

Wiegmann et al. (2012) claimed that research on moral judgment pointed to a systematic pattern of order effects that had been previously overlooked: Only judgments of actions that are normally (i.e., if judged in isolation) regarded as morally acceptable are 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 not considered as morally acceptable. If there is such a constellation, judgments of actions normally regarded as morally acceptable can approach judgments of previous actions (i.e., they can be deemed as less acceptable).

In order to test this claim Wiegmann and colleagues presented two groups of subjects with five trolley dilemmas, one after another (see Table 1). In all cases a train out of control was heading towards three railroad workers. An action was described that could be conducted by an agent in the situation to save the workers. This action varied in each of the five scenarios. The ordering of the scenarios was based on the level of agreement with the proposed action in each case, according to independent judgments provided in an independent pilot study (see Table 1). Level of agreement was measured on a scale of 1 to 6, where 1 was "not at all", and 6 was "absolutely". While in one group the level of agreement with the proposed action steadily increased, it decreased in the other group.

Table 1: Summaries of the actions proposed in the five dilemmas.

Scenario	Proposed action
Push	Push a large person from a bridge to stop the train
Trap	Push a button that will open a trap door that will let a 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

Table 2: 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	1.95	3.4	4.15	4.4	4.45
(SD)	(1.76)	(1.76)	(1.42)	(1.14)	(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.

According to the pattern of order effects outlined above, subjects' ratings for actions in the condition where the level of agreement was steadily decreasing (i.e., from Standard to Push; in the following called Least Aversive First, LAF) should not differ from ratings for the same actions when presented separately. The reason is that in such a constellation it is never the case that a judgment of an action normally (i.e., if judged in isolation) regarded as morally acceptable is preceded by an action that is normally regarded as morally unacceptable. In contrast, ratings for the actions in the last three scenarios in the Most Aversive First (MAF) condition (i.e., Redirect, Run Over and Standard) should decrease to the level of agreement of the preceding scenario (i.e., Trap), according to the pattern outlined above. That is, the low rating of Trap is assumed to reduce the level of agreement in Redirect, that in turn is assumed to decrease the rating for Run Over, that eventually decreases Standard's rating. This prediction was confirmed (see Figure 1). Unexpectedly, the ratings of the action in Trap were also affected by the ratings of the action in Push, although Trap is normally judged as unacceptable by a slim majority of subjects.

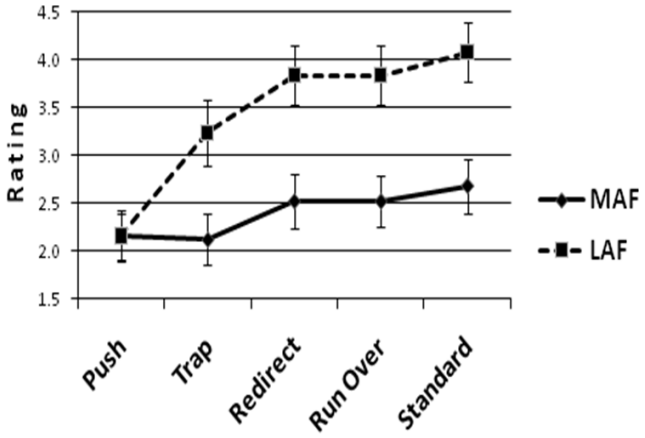

Figure 1: Mean ratings of agreement (1 stands for "not at all, 6 stands for "absolutely") with the proposed action in the five scenarios when evaluated sequentially, as a function of the order of presentation. Error bars indicate SEM. MAF = Most Aversive First; LAF = Least Aversive First.

This finding motivated a closer look at the results at the level of individual participants. In particular, the data was explored treating the ratings as a set of binary choices made by each participant (i.e., treating ratings ≤ 3 as indication of disagreement and ratings ≥ 4 as indication of agreement with the proposed action). The following tendency was observed: A disagreement with an action was virtually always "transferred" to the judgment of the action in the next scenario. That is, an action receiving a positive rating when judged independently received lower ratings when presented as part of a sequence if the preceding action was rated negatively by the same participant. However, positive ratings did not affect the ratings of the next action (by changing them into positive ones) if this action was rated negatively in independent ratings. Reformulating the pattern this way allows order effects to occur not only for actions rated positively when judged independently, but also for actions rated negatively on average. It just has to be the case that the number of participants that disagree with the proposed action in a particular scenario is sufficiently higher than the number of participants that disagree with the action in the subsequent scenario. This excess of "disagreements" can be transferred to the next scenario and cause an order effect. On the flipside, an order effect might also occur when a particular dilemma is preceded by another one where the proposed action is judged positively. Again, it just has to be the case that the number of disagreements in the preceding scenario is sufficiently higher than in the following scenario.

Although the pattern outlined at the individual level fits the data and allows making accurate predictions, the psychological mechanisms underlying order effects are still opaque. The experiments reported below pinpointed some of the features of the Push scenario that could be affecting other scenarios, and tested the effect of each of them individually. Features include differences in an emotional activation associated with Push, the activation of moral principles (e.g., "do not kill") and the trade-off of lives.

Experiment 1

In this experiment we aimed to test the hypothesis that the order effect described can be explained in terms of an emotional activation linked to Push, which would affect judgments in subsequent scenarios. As Green and his collaborators have shown, dilemmas like Push are more likely to activate brain regions associated with emotional processing than dilemmas like Standard (Green, Sommerville, Nystrom, Darley, & Cohen, 2001). Thus, in the sequence of scenarios described above (MAF) subjects might first experience a negative emotion when they are presented with Push, and once this negative emotion is in place, it might lead subjects to judge all the actions proposed in the remaining scenarios as morally unacceptable (cf. Haidt, 2001, Prinz, 2007). If the activation of such negative emotion is sufficient to cause order effects, then the presentation of other aversive scenarios that elicit a

similar emotion should also affect the judgment for other less aversive dilemmas (e.g., Standard).

Participants 259 subjects were recruited for a compensation of £ 0.50 via an online database located in the U.K..

Design, Materials, and Procedure Subjects were randomly assigned to one of four conditions. In two conditions subjects first had to read an aversive story, and then they were asked to judge the proposed action in Standard. The story was different in each of these two conditions. The following two stories were used:

Incest (Haidt, 2001):

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other.

Starving Child (actual newspaper article[1]):

> The 41-year-old man and 25-year-old woman, who met through a chat website, reportedly left their infant unattended while they went to internet cafes. They only occasionally dropped by to feed her powdered milk.
>
> According to the Yonhap news agency, South Korean police said the couple had become obsessed with raising a virtual girl called Anima in the popular role-playing game Prius Online. The game, similar to Second Life, allows players to create another existence for themselves in a virtual world, including getting a job, interacting with other users and earning an extra avatar to nurture once they reach a certain level.

In the two remaining conditions subjects had to either judge Standard alone (to obtain a baseline rating), or Standard after having judged Push. Additionally, as the study was conducted online, at the end of the questionnaire subjects completed a simple logical task to identify those who did not pay sufficient attention to the task.

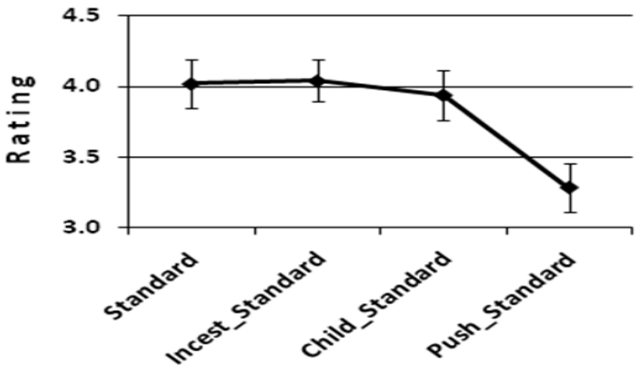

Figure 2: Mean Ratings for Standard as a function of the preceding scenario. Error bars indicate SEM.

Results Fifty subjects dropped out because they did not answer the test question, failed to solve the logical question or went through the whole survey in less than 40 seconds.

The results for the remaining subjects are summarized in Figure 2. The mean rating for Standard when judged independently was 4.02 (SD=1.18), while it was lower (M=3.28, SD=1.39) when it was preceded by Push (Push_Standard), as predicted. A planned-contrast test confirmed this difference to be significant, $F_{1,205}$=9.68, p<0.001. In contrast to this, reading and judging Newspaper or Incest did not have any effect on Standard (Standard vs. Incest_Standard: $F_{1,205}$=0.01, p>0.9; Standard vs. Child_Standard: $F_{1,205}$=0.10, p>0.7).

Discussion The findings obtained suggest that an emotional activation may not be sufficient to cause the kind of order effect described above. That is, judgments for Standard were not affected by the prior presentation of different scenarios that are likely to have elicited negative emotions. Further evidence for this idea comes from ongoing research conducted in Spain in which participants were presented with selected pictures of unpleasant affective valence and high arousal, before judging Standard. In line with the study described above, preliminary results revealed that the emotional priming did not affect judgments for Standard.

Experiment 2a and 2b

The two following experiments test the hypothesis that the order effect described is related to differences in the activation of principles associated with each dilemma. In particular, Push could trigger the urge for subjects to justify their judgment, or the principle "Do not kill!", while Standard may not. If the activation of such principle can account for the carry over effect of judgments when Push is presented first, it is reasonable to expect that forcing the activation of a principle relevant for Standard when this dilemma comes first (e.g., "Save the greater number") should lead to an order effect in the opposite direction (i.e., people should be more likely to agree with the action proposed in subsequent scenarios).

Experiment 2a

The rationale for this experiment was as follows: When subjects judge the action proposed in Push as morally unacceptable they articulate, so to say, a prohibition or imposing a ban. Actions which we prohibit are generally accompanied by a justification. For instance, one often has to justify or explain to kids why something is forbidden. In contrast, there are fewer situations where one has to explain why something is allowed. Justifications for allowed actions generally only happen when the actions were expected to be forbidden. In Push, the vast majority of subjects judge the proposed action as forbidden while in Standard they don't. Hence, it could be the case that subjects first judging Push have a stronger urge to justify their judgment. If the justification is "You are not allowed to kill innocent people" (or something similar), judgments for subsequent scenarios could accommodate this justification, explaining why all

proposed actions are regarded as not acceptable. In contrast, subjects starting with Standard might not have an urge to justify their initial judgment, explaining why subsequent judgments are not affected.

Participants 36 subjects were recruited from a student subject pool. Participants were compensated with course credits.

Design, Materials, and Procedure Subjects were presented with the LAF condition. However, in contrast to the original LAF condition described above, here participants were required to justify their judgment for Standard. Since participants were recruited from the same student pool (mainly psychology students) and this experiment was conducted only a few weeks after the one conducted by Wiegmann and colleagues (2012), the original LAF condition was used as a control condition.

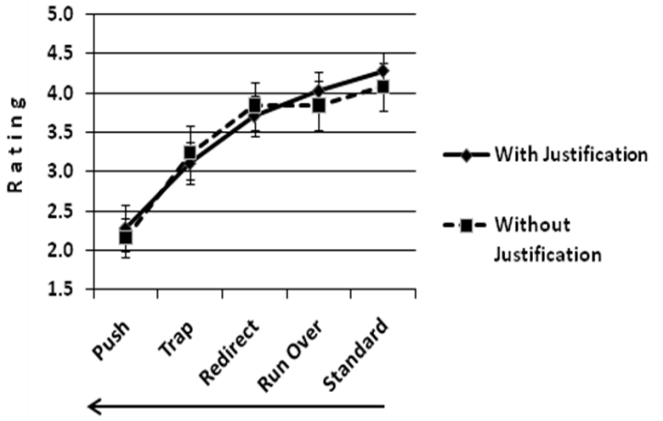

Figure 3: Mean Ratings for the five scenarios as a function of whether subjects had to justify their judgment for Standard. Arrow indicates order of presentation. Error bars indicate SEM.

Results One subject failed to answer all questions for unknown reasons. Results for the remaining subjects are summarized in Figure 3. As can be seen, being forced to justify the judgment for Standard did not influence judgments for the actions proposed in the following scenarios. An ANOVA with justification (required vs. not required) as a between-subjects factor and scenario as a within-subjects factor revealed that there was neither a main effect of justification ($F_{1,58}$=.02, p=.88) nor an interaction between justification and scenario ($F_{4,232}$=.62, p=.65).

Experiment 2b

The rationale for this experiment was as follows: When judging Push the principle "Do not kill!" comes easily to mind because the proposed action in this dilemma is a paradigmatic case of killing a person. In contrast, Standard might not trigger such a clear principle. In philosophy, the permissibility to intervene in Standard is often justified by rather subtle principles like the Doctrine of Double Effect or even more subtle principles (Kamm, 2007). In this experiment we sought to trigger a clear principle (saving the greater number) and test whether it would be carried over to the following scenarios, thereby affecting judgments.

Participants 63 subjects were recruited from a student subject pool. Participants were compensated with course credits.

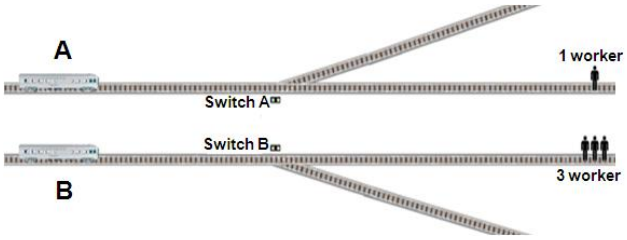

Figure 4: Illustration of the Rescue scenario

Design, Materials, and Procedure Two of the three conditions tested were LAF and MAF. The third condition included a new scenario called Rescue that was placed before Standard and was intended to trigger the principle "Save the greater number of lives" (see Figure 4). In this scenario a train is threatening one person and another train is threatening three persons. There is not enough time to throw the switch for both trains so that everyone is saved. We assumed that in such a case virtually everyone would agree to throw the switch that will save three persons rather than only one.

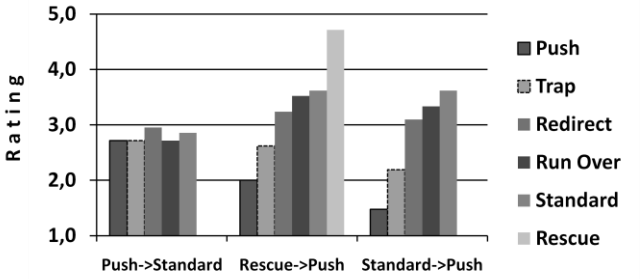

Figure 5: Mean ratings for the 5 (6) scenarios as a function of both the order of presentation and whether the Rescue scenario was present. For example, Push->Standard means subjects started with Push and ended with Standard.

Results The results are shown in Figure 5. As expected, a transference in judgments was observed for MAF, but not for LAF ($F_{4,160}$=11.37, p<.001). However, introducing Rescue before Standard did not affect judgments for Standard, as evidenced by the absence of an interaction of condition and scenario between LAF and the new condition starting with Rescue ($F_{4,160}$=.60, p=.67).

Discussion The results of these two experiments can be interpreted in at least two ways. First, it could be the case that the order effect in MAF is neither based on a stronger urge to justify one's judgment in Push (2a) nor due to Push triggering a clear principle (2b). Alternatively, the order effect in MAF could indeed be driven by a principle

triggered by Push which is carried over. However, the principle triggered by Standard or Rescue may not be, so to say, strong enough to override intuitions in other scenarios. In other words, while a principle like "Do not kill!" may be carried over once it has been triggered, a principle like "Save the greater number!" may not be potent enough to influence moral judgments in following scenarios (cf, e.g., Gert, 2007).

Experiment 3

In this experiment we took a further step aiming to investigate which features of Push are necessary to cause an order effect. In particular, we examined the impact of the number of lives that are traded-off in Push.

Participants 343 subjects, each receiving £ 0.50, were recruited via an online database located in the U.K..

Design, Materials, and Procedure Subjects were randomly assigned to one of four conditions. In one condition Standard was judged alone (to obtain a baseline rating), while in a second condition Standard was judged after Push, involving the same number of potential victims as in previous experiments (Standard_Push_3). In the third and fourth conditions we manipulated the number of potential victims that would be saved by the intervention in Standard. In one condition nobody would be saved by pushing the person from the bridge (Standard_Push_0), while in the other condition a group of one hundred people would be saved (Standard_Push_100).

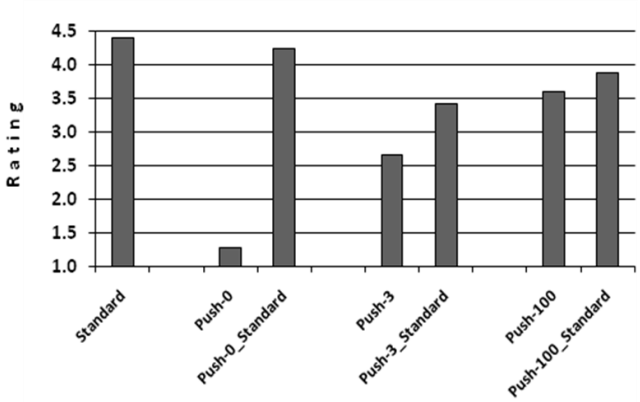

Figure 6: Mean ratings for Standard and its preceding scenario, i.e. Push-0 is the rating for Push if nobody is recued by killing the one person and Push-0_Standard is the rating for Standard if preceded by Push-0.

Results 96 subjects did not answer the test question, gave the wrong answer to the logical task, or took less than 40 seconds for completing the task.

The results for the remaining subjects are summarized in Figure 6. The baseline rating for Standard was the highest in descriptive terms (M=4.4, SD=1.26). The next highest rating for Standard was delivered if it was preceded by the version of Push where pushing the man from the bridge would not save anyone (i.e., Standard_Push_0) (M=4.26, SD=1.46). The difference between Standard in this condition and the baseline rating was not significant, $F<1$. In contrast, ratings for Standard where lower in Standard_Push_3 and in Standard-Push-100 than for the baseline ($F_{1,243}$=18.14, $p<0.001$ and $F_{1,243}$=4.57, $p<0.05$, respectively).

Discussion One might wonder why the rating for Standard is lowered to a lesser extent when it is preceded by Push involving saving 100 people than by Push involving saving three. However, the pattern outlined above can account for this finding. Recall that most of the time only negative ratings were transferred to the next scenario. Since the rating for Push_100 (M=3.61, SD=1.50) was already high, it is not surprising that the rating for Standard_Push_100 was relatively high too. There were seemingly just not enough negative ratings for Push_100 to be transferred and lower the rating of Standard_Push_100 to the same extent as for Standard_Push_3. Interestingly, the results show that Standard is not influenced by a version of Push in which killing the person does not save anyone. Since Push_0 and Push_3 only differ with regards to whether there is a trade-off of lives involved, we can infer that a dilemma must contain such trade-off to influence judgments for Standard.

Experiment 4

The results of experiment 3 suggest that a scenario preceding Standard needs to contain a trade-off of lives in order to influence Standard. In this experiment we aimed to investigate whether Standard could be influenced by a preceding scenario similar to Push with regards to being aversive and containing a trade-off, but with a different cover story. Furthermore, we tested whether just reading (and not judging) Push or similar scenarios would be sufficient to influence Standard.

Participants 321 subjects were recruited for a compensation of £ 0.50 via an online database located in the U.K..

Design, Materials, and Procedure Subjects were randomly assigned to one of five conditions. As in previous experiments there was one condition to get a baseline for Standard and another where Standard was preceded by Push. In a third condition (Organ) Standard was preceded by a scenario in which a doctor can save three patients by transplanting organs from a healthy person into them.. The last two conditions, (Push_readonly) and (Organ_readonly) were identical to Push and Organ, respectively, except that subjects were not given the opportunity to judge the proposed action.

Results 43 subjects did not answer the test question, gave the wrong answer to the logical task, or took less than 40 seconds for completing the task.

The results for the remaining subjects are summarized in Figure 7. The baseline rating for Standard was the highest in

descriptive terms (M=4.35, SD=1.25). Again, the difference between Standard_Push_3 (M=3.37, SD=1.37) and the baseline rating was, as expected, significant ($F_{1,282}$=14.46, p<0.001). Interestingly, ratings for Standard when preceded by Organ were also significantly lower than the baseline ($F_{1,282}$=6.13, p<0.05). As Figure 7 shows, it did not make a difference whether subjects just read or also judged the action proposed in Push or Organ.

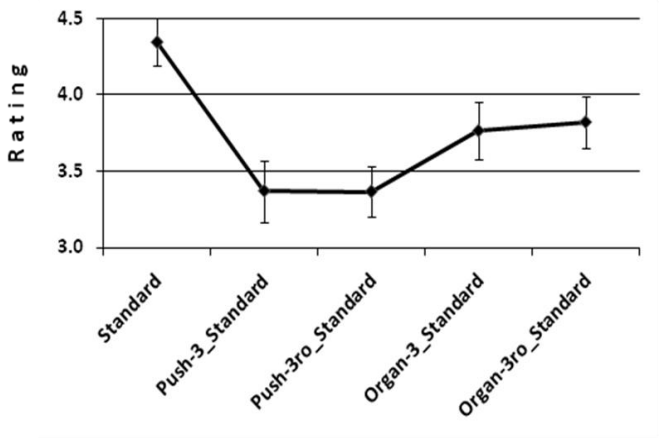

Figure 7: Mean Ratings for Standard as a function of the preceding scenario. Error bars indicate SEM. "ro" stands for "read only".

Discussion The results showed that Standard can be influenced by a scenario with a different cover story than Push, but that also involves a trade-off of lives. This influence is not as strong as the influence of Push but it is still significant. Moreover, whether subjects only read a scenario or also judged it did not affect judgments.

General Discussion

In this paper we sought to search for an explanation of order effects in moral judgment. The overarching question guiding our experiments was: Which features of the Push scenario are the ones that enable Push to influence other scenarios such as Standard? Although we still do not have a complete explanation, we think we have made some progress towards this aim.

Our findings suggest that the order effect is likely not be caused by negative emotions activated by Push. Presenting subjects with aversive stories which were likely to elicit such emotions did not have any effect on the judgment for Standard (Experiment 1).

Forcing participants to justify their judgment for Standard or saliently triggering the principle "Save as many lives as possible!" did not affect judgments for Standard either (Experiments 2a and 2b). As noted above, these results can be explained in two ways: First, it might be that the order effect is not driven by the activation of a principle for the Push scenario, which is then applied to subsequent scenarios. Second, it might be that the principle "Save lives" affects subsequent judgments only to a lesser extent than the principle triggered in Push (presumably: "Do not kill"" or "Do not kill in order to save lives!").

Interestingly, we found a feature of Push that seems necessary to influence a subsequent scenario: The trade-off of lives. A version of Push which did not involve such trade-off had no influence on judgments for Standard. Hence, apart from being judged more negatively than Standard (see Introduction) containing a trade-off of lives seems to be a necessary feature for a scenario to influence subsequent scenarios.

Interestingly, when we presented subjects with a scenario (Organ) similar to Push in that it contained a very aversive action and a trade-off of lives, but that differed with regards to the cover story, the agreement with the action proposed in Standard was also reduced. Future research should examine other similarities between Push and Organ which could be related to their potency to cause order effects.

In summary, our findings suggest that the features that a moral dilemma must exhibit in order to affect judgments in subsequent dilemmas (here, in Standard) are:

1. It must receive significantly more negative ratings than the following dilemma (see Introduction)
2. It must contain a trade-off of lives

Furthermore, if the preceding scenario exhibits these two features the influence on Standard becomes even stronger if the superficial features of the cover story resemble the ones in Standard.

Acknowledgments

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).

References

Gert, B. (2007). *Common morality: Deciding what to do*. New York: Oxford University Press.

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.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108, 814–834

Kamm, F. M. (2007). Intricate ethics. Oxford, England: Oxford University Press.

Prinz, J. J. (2007). The emotional construction of morals. New York: Oxford University Press.

Schwitzgebel, E., & Cushman, F. (in press). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. Mind and Language.

Wiegmann, A., Okan, Y., & Nagel, J.(2012). Order effects in moral judgment. Philosophical Psychology,