

Sacrificing objects instead of persons: Order effects without emotional engagement

Emilian Mihailov, Ivar R. Hannikainen & Alex Wiegmann

To cite this article: Emilian Mihailov, Ivar R. Hannikainen & Alex Wiegmann (2023): Sacrificing objects instead of persons: Order effects without emotional engagement, *Philosophical Psychology*, DOI: [10.1080/09515089.2023.2195043](https://doi.org/10.1080/09515089.2023.2195043)

To link to this article: <https://doi.org/10.1080/09515089.2023.2195043>



Published online: 23 Apr 2023.



Submit your article to this journal [↗](#)



View related articles [↗](#)




View Crossmark data [↗](#)

ARTICLE



Sacrificing objects instead of persons: Order effects without emotional engagement

Emilian Mihailov ^{a,b}, Ivar R. Hannikainen^{b,c} and Alex Wiegmann^d

^aResearch Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest, Romania; ^bBioXPhi Lab, Oxford Uehiro Centre for Practical Ethics, UK; ^cDepartment of Philosophy I, University of Granada, Spain; ^dInstitute for Philosophy II, Ruhr University Bochum, Germany

ABSTRACT

In this paper we develop test cases to adjudicate between dual-process and the causal mapping explanations of order effects. Using dilemmas with minimized emotional force, we explore new conditions for order effects to occur. Overall, the results support causal model theory. We produced novel evidence that order effects extend not only to cases with low emotional engagement, but also to specialized judgments about whether an action violates a rule. However, when objects are sacrificed instead of persons the order effect either disappears or becomes symmetrical, contrary to previous theorizing that it is an asymmetrical transfer effect. Causal model theory needs to be developed to include interplays between the moral status of sacrificed entities and computational models of causal mapping. Symmetric order effects remain a puzzle, motivating future research. Though we do not know how to explain them yet, we discuss how symmetric order effects can influence policy decision making.

Keywords

order effects; causal model theory; dual process; moral judgment; symmetric order effects; moral ontology

Introduction

Issuing a moral judgment should be based on the relevant qualities of the target behavior: How good or bad were its outcomes? Were the outcomes brought about intentionally or accidentally? For instance, it is well-known that people generally approve of diverting a train onto a sidetrack to save five innocent persons (in the Switch scenario), but oppose actively killing one person as a result of redirecting the train (in the Push scenario).

Yet there is growing evidence that the order in which behaviors are evaluated – a seemingly *irrelevant* quality – can influence the content and severity of those moral judgments (Lombrozo, 2009; Petrinovich & O’neill, 1996; Schwitzgebel & Cushman, 2012; Wiegmann & Waldmann, 2014; Wiegmann et al., 2012). Diverting the train in the Switch scenario becomes

less acceptable when it is evaluated immediately after the Push scenario. An action that is usually deemed acceptable can become morally unacceptable without any changes to its relevant features. Strikingly, even professional philosophers are susceptible to order effects in their judgments about moral scenarios and their endorsement of moral principles (Schwitzgebel & Cushman, 2012; Wiegmann et al., 2020). Philosophers' susceptibility raises concerns that they are not immune to post-hoc rationalization (Bortolotti, 2011; Greene, 2008; Mihailov, 2016).

What psychological mechanisms account for this undesirable influence on moral judgment? The most prominent explanations focus on (i) *emotion elicitation*, and (ii) the *causal structure* of actions and their default evaluations. Initial research by Greene and collaborators showed that Push cases are more likely to activate brain regions associated with emotional processing than Switch cases (Greene et al., 2001, 2004). According to early formulations of the dual-process theory (Greene, 2008), automatic emotional processes cause deontological judgments ("wrong to kill one person to save five"), while cognitive control processes underlie utilitarian judgments ("right to kill one person to save five"). On this view, order effects are explained by a *carry-over effect* of emotional activation onto Switch cases (Greene, 2014). When the Push case is presented first, an aversive response contributes to the moral condemnation of the sacrificial action. The lingering affective response can heighten condemnation of cases that, in normal circumstances, would not elicit an aversive response—e.g., the Switch case – if presented immediately after. In the opposite order, when the Switch case precedes the Push case, people initially engage in cost-benefit analysis to endorse the utilitarian action of saving more lives. If the Push case is presented next, a spontaneous affective response arises which (unconstrained by cognitive control processes) promotes opposition to the utilitarian action of saving more lives. Dual-process theory explains the asymmetric pattern of order effects by a "priming" affect in a Push case that modulates subsequent responses to Switch cases, whereas priming deliberation in Switch cases does not affect subsequent responses to Push cases.

Alternatively, Alex Wiegmann and Michael Waldmann (Wiegmann & Waldmann, 2014) elaborated a causal model theory of order effects, motivated by research outside moral psychology, namely how people interpret ambiguous images (Medin et al., 1993). The Push dilemma is an unambiguous moral dilemma because there is only one causal path (a chain-like structure) that involves the intervention (push), the bad outcome (one person dies) and the good outcome (five persons are saved). The Switch case, on the other hand, is an ambiguous moral dilemma because there are two, in some sense independent, causal paths underlying it (see [Figure 1](#)). In the Push dilemma, the victim is used as a means in the causal chain to reach

A. Wiegmann, M.R. Waldmann/Cognition 131 (2014) 28–43

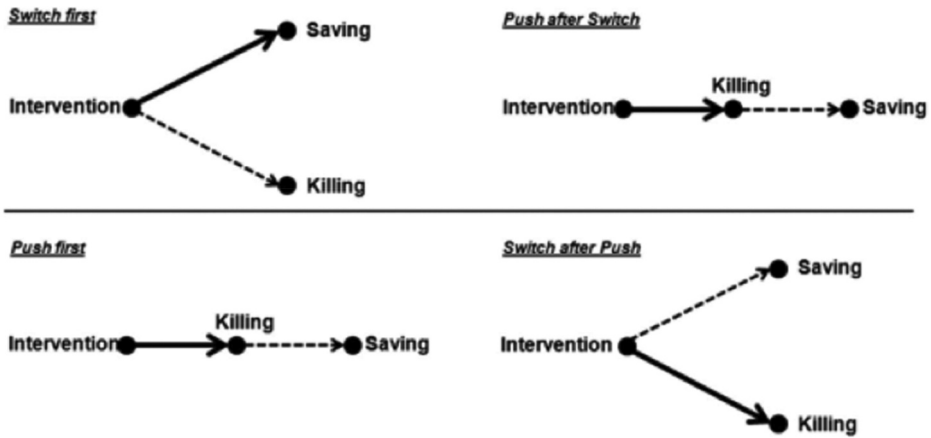


Figure 1. The highlighted path of the causal structure of switch and push in the default evaluation (left side) vs. when preceded by the other dilemma (right side). Bold solid lines represent the highlighted part of the causal structure.

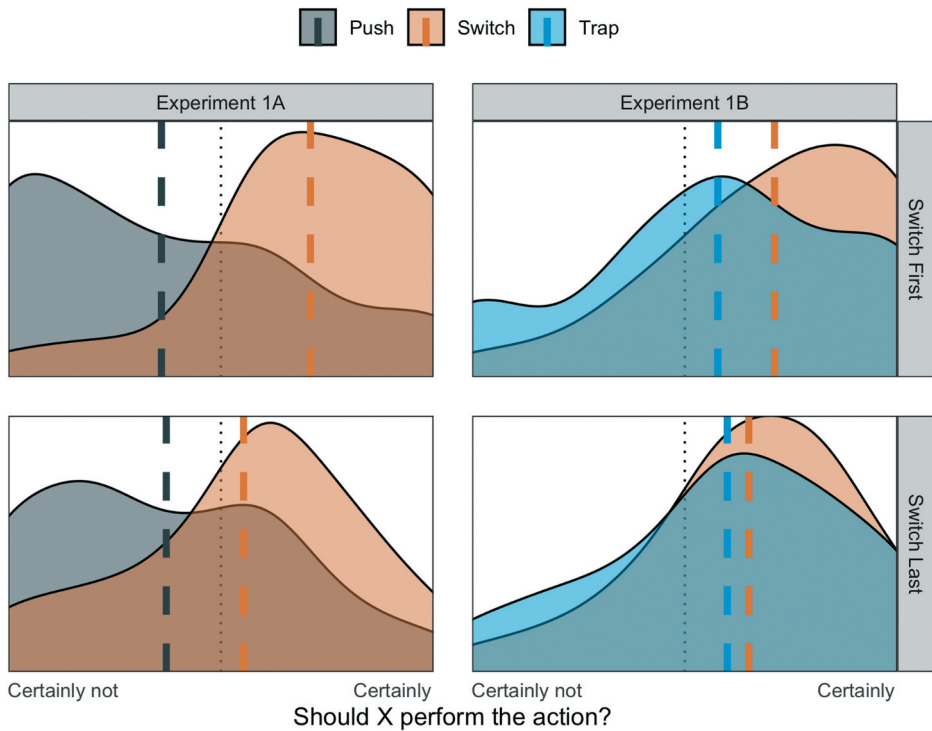


Figure 2. Density plots of moral judgment in experiments 1A and 1B (columns) by order (rows).

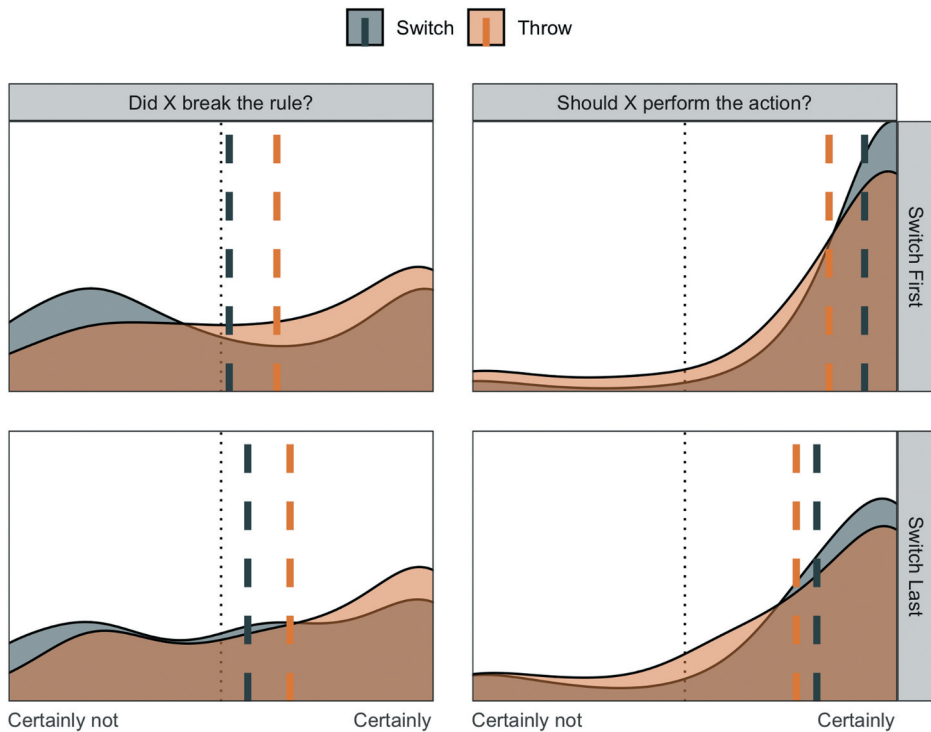


Figure 3. Density plots of moral judgment in experiments 2A and 2B (columns) by order (rows).

the goal of saving the five, so the bad outcome lies between and on the same causal path as the intervention and the good outcome. Order effects take place depending on whether the default highlighted causal path from the first scenario can be mapped onto the causal structure of the second scenario. If people are presented first with Switch, the causal path from the intervention to *saving* lives is highlighted by default (people's judgments favor redirecting the train from five persons to the one person). But this causal path from the intervention to the good outcome cannot be mapped analogously onto the causal structure of Push because the bad outcome (killing the one person) lies on the causal path from the intervention to the good outcome. By contrast, when Push is presented first, the causal path to *killing* is highlighted by default (people's judgments favor not intervening). Since the causal structure of Switch also includes a direct path from the intervention to the bad outcome, the previously highlighted path from the intervention to the bad outcome can be mapped onto the causal structure of Switch. Now the bad outcome of the intervention becomes more salient and Switch is evaluated more negatively, compared to the first judgment, resulting in an asymmetrical transfer effect.

We don't have enough data to adjudicate between these competing explanations (Wiegmann & Waldmann, 2014). developed a test case to

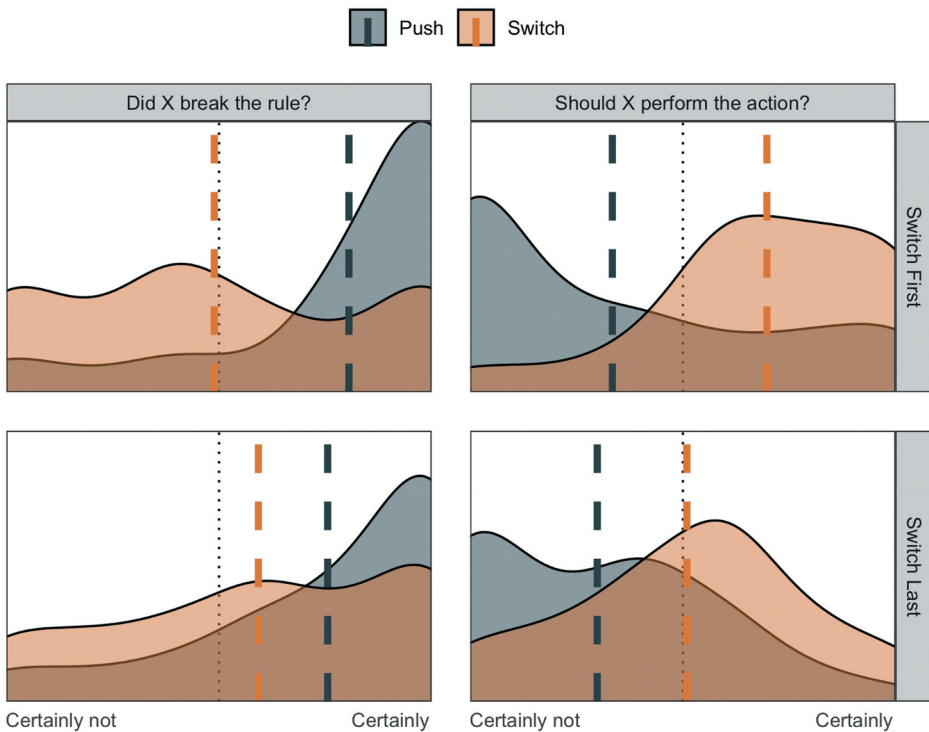


Figure 4. Density plots of moral judgment in experiments 3 by dependent measure (columns) and order (rows).

show that causal model theory outcompetes dual-process theory. They imagine a positive scenario of Push (PosPush) in which saving five persons from being run over by a train can be achieved by pushing them directly off the tracks. Unfortunately, another person further down the tracks who had not been in danger before will die. Here, the primary action is pushing people to save them. The causal model theory correctly predicts the lack of an order effect between Push and PosPush because they have unambiguous causal structures which cannot be mapped onto one another (Wiegmann & Waldmann, 2014). Emotional engagement cannot explain the lack of an order effect between Push and PosPush. If order effects are triggered by emotions, order effects should be observed between Push and PosPush, which is not the case. However, no strong conclusion can be drawn from just one example. Moreover, we are lacking evidence from cases with reduced emotional engagement which have the potential to advance the discussions on emotion elicitation versus causal structure explanations.

In this paper we further develop test cases for the dual-process and the causal mapping explanations and explore the scope of order effects. We create an impersonal experimental design that keeps the relevant causal features and outcomes of the trolley type scenarios consistent. Dilemmas

with minimized emotional force provide new conditions under which asymmetrical transfer effects can occur. The affective turn in neuroscience and moral psychology has revealed a substantial role for emotions in moral judgments (Haidt, 2013). People's responses to moral dilemmas vary systematically in the extent to which they engage emotional processing. Following the way these variations influence moral judgment, psychological research has focused on emotionally charged scenarios. However, the focus on emotion-based explanations has overlooked other non-emotional factors (Nichols & Mallon, 2006). have shown, for example, that moral judgments about trolley cases can be explained in terms of what moral rules do and do not forbid. They designed scenarios that were impersonal analogues of the Switch and Push cases. Instead of deciding whether to sacrifice people for the greater good, the scenarios described a trolley-like situation in which one had to "sacrifice" physical objects and decide whether a rule had been broken. Further, using impersonal versions of the scenario can make people less condemning of sacrificing one person for the greater good. The action of dropping a victim onto the tracks using a trap door and a remote switch is considered more permissible than using personal force to push the victim, which triggers strong emotional reactions (Cushman & Greene, 2012; Greene et al., 2009).

Impersonal versions of trolley-like decision making are critical to test emotional engagement and causal mapping explanations. We used impersonal analogues of the Switch and Push cases which involve damaging property, as well as versions of footbridge which involve remote killing. Our experiments aim to elucidate which account is most plausible. If order effects occur under impersonal conditions, then the dual-process explanation is implausible. When there is no prepotent emotion to exert a strong influence on moral judgment, dual-process explanations cannot appeal to differential emotional engagement. If order effects are not documented in impersonal scenarios, then the causal model theory needs further development to include interplays between emotional input, judgments about rule-application and computational models of causal mapping.

Experiment 1A: asymmetrical order effect

In Experiment 1A, we sought to replicate existing work on order effects in moral judgments of trolley dilemmas (Wiegmann & Okan, 2012; Wiegmann & Waldmann, 2014; Wiegmann et al., 2012, 2020). We used the standard Switch and Push paradigms. The hypothesis was that presenting people first with Push would affect their judgment for Switch while presenting people first with Switch would not affect their judgment for Push.

Methods

162 English-speaking participants (mean age = 26.3 years old; 41% women) were recruited on Prolific. In a 2 between (order: Switch First, Push First) x 2 within (case: Switch, Push) -subjects design, participants viewed both Switch and Push dilemmas in a randomized order. After each case, participants judged whether the agent should perform the action on a six-point scale from 1: “Certainly not” to 6: “Certainly”.

Results

A mixed ANOVA revealed a main effect of case and a case-by-order interaction (see Table 1). The interaction indicated that the distinction between Push and Switch was larger when Switch preceded Push (Cohen’s $d = 1.08$) than when Push preceded Switch (Cohen’s $d = 0.81$, see Figure 2). As in previous studies, Switch cases were judged less acceptable when preceded by Push cases, $t = 4.38$, $p < .001$, Cohen’s $d = 0.60$. No corresponding order effect arose for Push cases, $t = 0.42$, $p = .68$.

Discussion

Experiment 1A replicated the asymmetric order effect observed in prior literature. Judgments of Switch cases depended on case order, whereas judgments of Push cases did not.

Experiment 1B: order effect without personal force?

Our aim in Experiment 1B was to produce initial evidence for discriminating between dual-process and causal model explanations. Dual-process explanations appeal to the failure of cognitive deliberative processes to control automatic emotional processes, while causal mapping explanations appeal to a potential similarity in the underlying intentional structure of presented scenarios. We used the standard Switch and the Trapdoor version of Push dilemma, both of which do not involve personal force; in the Trapdoor version the victim is dropped onto the tracks using a trapdoor which is activated remotely.

Table 1. Split-plot ANOVAs for experiments 1a and 1b.

	Experiment 1a			Experiment 1b		
	<i>df</i>	<i>F</i>	<i>p</i>	<i>df</i>	<i>F</i>	<i>p</i>
<i>case</i>	(1, 160)	148.9	<.001	(1, 175)	27.65	<.001
<i>order</i>	(1, 160)	3.38	.068	(1, 175)	0.25	.62
<i>case*order</i>	(1, 160)	14.47	<.001	(1, 175)	5.51	.020

Methods

168 participants (mean age = 26.2 years old; 46% women) were recruited on Prolific. In a 2 between (order: Switch First, Trap First) x 2 within (case: Switch, Trap) -subjects design, participants viewed both Switch and Push dilemmas in a randomized order. Participants judged whether the agent in each case should perform the action on a six-point scale from 1: “Certainly not” to 6: “Certainly”.

Results

As in Experiment 1A, a mixed ANOVA revealed a main effect of case and a case-by-order interaction (see Table 1). We observed a moral distinction between Trap and Switch cases, which was larger when Switch preceded Trapdoor (Cohen’s $d = 0.52$) than when Trapdoor preceded Switch (Cohen’s $d = 0.24$, see Figure 1). Unlike in Experiment 1A, however, neither simple effect of order was significant – whether on Switch, $B = 0.30$, $t = 1.44$, $p = .15$, or on Trapdoor, $B = -0.11$, $t = -0.53$, $p = .60$.

Discussion

Experiment 1B revealed that order effects occur even when the element of personal force – known to engender affective responses – is absent. Demonstrating that order effects can arise in reaction to low affect cases casts doubt on emotion-based explanations (but is consistent with causal structure explanations).

The moral distinction between Switch and Push in Experiment 1A was larger than the distinction between Switch and Trap in Experiment 1B. This variation in the default difference between cases could help explain the weaker order effect in Experiment 1B (in comparison to Experiment 1A). Specifically, the weaker distinction between Trap and Switch creates a smaller “upper bound” on the order effect – since the magnitude of the order effect cannot exceed the default difference between cases.

Experiment 2A: sacrificing teacups instead of people

In Experiment 2A we devised a further test to discriminate between dual-process and causal model explanations by replacing human victims with mere objects. In previous research (Nichols & Mallon, 2006), devised material object versions of the Switch and Push cases. These cases describe trolley-like situations in which a teacup can be “sacrificed” to prevent the destruction of a greater number of teacups. We explored whether order effects could extend to scenarios involving trade-offs between objects. These

cases have the same causal structure as the cases in Experiment 1A involving human victims, but elicit a much weaker affective response – or perhaps no affect at all. Therefore, the causal mapping account predicts that an order effect will arise, while the dual-process account predicts that it will not.

Methods

333 participants (mean age = 26.0 years old; 60% women) were recruited on Prolific. In a 2 between (order: Switch First, Throw First) x 2 within (case: Switch, Throw) -subjects design, participants viewed both Switch and Throw versions of a teacup dilemma in a randomized order. In the impersonal Switch case, an agent sees that if a toy train continues on its present course, it will run through and break five cups. The agent cannot get to the cups or the off-switch in time, but he can reach a lever, which will divert the train to a side track, breaking one of the cups. In the impersonal Push case, an agent sees that a toy truck is about to wreck the cups. He is standing next to the counter with the remaining teacups and realizes that the only way to stop the truck in time is by throwing one of the teacups at the truck, which will break the cup he throws. Experiment 2A employed a new dependent measure: Participants were asked to decide whether a household rule against breaking teacups had been violated, using a six-point Likert scale from 1: “Certainly not” to 6: “Certainly”.

Results

A mixed ANOVA revealed a main effect of case, but no main effect of order or case-by-order interaction (see Table 2). The main effect of case revealed that throwing a teacup violated the rule more so than redirecting toward a teacup (Cohen’s $d = 0.38$). However, unlike in the previous experiments, the magnitude of the distinction was roughly the same in both orders, both $ps < .001$ (see Figure 3).

Discussion

We observed no significant order effect, despite the fact that participants recognized a moral distinction in breaking a rule between throwing vs

Table 2. Split-plot ANOVAs for experiments 2A and 2B.

	Experiment 2A			Experiment 2B		
	<i>df</i>	<i>F</i>	<i>p</i>	<i>df</i>	<i>F</i>	<i>p</i>
<i>case</i>	(1, 331)	48.36	<.001	(1, 309)	22.03	<.001
<i>order</i>	(1, 331)	1.10	.30	(1, 309)	11.95	<.001
<i>case*order</i>	(1, 331)	0.17	.68	(1, 309)	1.55	.21

redirecting. These scenarios keep the same differences in causal structures as in normal trolley dilemmas in which order effects are robustly documented.

Experiment 2B: sacrificing teacups instead of people

In Experiment 2A, participants' judgments of whether a rule prohibiting the destruction of tea cups had been violated were not subject to an order effect. Whether the absence of an order effect stems from the fact that participants judged whether the agent had *violated the rule* (and not what they should have done), or from the fact that the actions were directed at *inanimate objects* (instead of people). Our aim in Experiment 2B was to understand this question. In Experiment 2B, we asked participants to consider these object cases, and issue a prescriptive judgment ("Should one teacup be broken in order to save five teacups?"; in line with Experiments 1A and 1B) instead of a transgression judgment.

Methods

311 participants (mean age = 26.2 years old; 68% women) were recruited on Prolific. In a 2 between (order: Switch First, Throw First) x 2 within (case: Switch, Throw) -subjects design, participants viewed both Switch and Throw versions of a tea cup dilemma in a randomized order. After each case, participants judged whether the agent should perform the action on a six-point scale (as in Experiments 1A and 1B).

Results

A mixed ANOVA revealed main effects of case and order (but no case*order interaction; see Table 2). The main effect of case revealed that participants were more likely to say that the agent should redirect the toy train in the Switch case than that the agent should throw a teacup in the Throw case (Cohen's $d = 0.27$) in order to prevent the destruction of five teacups.

The main effect of order revealed a symmetrical order effect – such that both teacup dilemmas were affected by order: The Switch dilemma was deemed less acceptable when preceded by Throw, $t = 3.64$, $p < .001$, Cohen's $d = -0.45$. Additionally, the Throw dilemma was deemed more acceptable when preceded by Switch, $t = 2.50$, $p = .013$, Cohen's $d = 0.27$.

Discussion

The results of Experiment 2B suggest that order effects can arise when evaluating trade-offs between *inanimate objects* (instead of people). These results also suggest that the absence of an order effect in Experiment 2A was

due largely to changes in the dependent measure: namely, that *transgression* judgments (of whether an agent had violated the rule) are less susceptible to order effects than *prescriptive* judgments (about what the agent should have done).

Surprisingly, the results of Experiment 2A revealed a *symmetrical* order effect – unlike the previous experiments that revealed an asymmetry. In conditions of high ratings of permissibility with the same causal structures as in Switch and Push, Teacup Push can become more permissible when Switch is first, whereas in normal trolley dilemmas Push does not become more permissible when preceded by Switch.

Experiment 3: do order effects extend to rule violation judgments?

Taken together, the results of Experiments 2A and 2B raise the possibility that order effects do not extend to judgments about rule violation: When we asked participants whether throwing a teacup or redirecting a toy train to minimize damage violates the household rule, we did not observe an order effect. Then, when participants were asked whether the agent should throw a teacup or redirect the toy train, the order effect reappeared. This pattern suggests that judgments about rule violation are less susceptible to order effects than prescriptive judgments. Our aim in Experiment 3 was to corroborate this phenomenon in the context of sacrificial dilemmas: i.e., by comparing the magnitude of order effects on prescriptive judgments (about what the agent facing the dilemma should do) versus judgments of whether they violated the rule “not to kill”.

Methods

310 participants (mean age = 25.6 years old; 57% women) were recruited on Prolific. In a 2 (order: Switch First, Push First) x 2 between (judgment: transgression, prescription) x 2 within (case: Switch, Push) -subjects design, participants viewed both Switch and Throw versions of a teacup dilemma in a randomized order. Participants were also randomly assigned to either judge whether the agent should perform the action, or whether they had violated the rule not to kill, on a six-point scale.

Results

Replicating Experiment 1A, we found a dilemma*order interaction (i.e., an asymmetric order effect) when asked whether the agent should carry out the action (see Table 3). This asymmetric order effect arose also when participants reported whether the agent had violated the rule not to kill (see Table 3).

Table 3. Split-plot ANOVAs for experiment 3 (separately for each dependent measure).

	Experiment 3: prescriptive judgments			Experiment 3: transgression judgments		
	<i>df</i>	<i>F</i>	<i>p</i>	<i>df</i>	<i>F</i>	<i>p</i>
<i>case</i>	(1, 155)	133.64	<.001	(1, 151)	60.14	<.001
<i>order</i>	(1, 155)	7.45	.007	(1, 151)	0.38	.54
<i>case*order</i>	(1, 155)	9.22	.003	(1, 151)	6.39	.012

The pattern of simple effects once again indicated that Push cases were unaffected by order (p s > .32), whereas Switch dilemmas demonstrated an effect of order: When preceded by a Push case, participants were less likely to report that the agent in the Switch dilemma should perform the action ($B = -0.94$, $t = -3.91$, $p < .001$, Cohen's $d = -0.69$) and more likely to report that they broke the rule not to kill ($B = -0.52$, $t = -1.95$, $p = .053$, Cohen's $d = 0.29$; at the marginally significant level) (see [Figure 4](#)).¹

Discussion

Experiment 3 investigated whether more impersonal judgments like rule judgments are susceptible to order effects. We obtained evidence that participants were more likely to state that the Switch cases violate the no killing rule when preceded by the Push case, suggesting that order effects can arise in response to transgression judgments, as well as prescriptive judgments. This is the first evidence that order effects apply to rule breaking judgments. Furthermore, when statistically comparing the magnitude of rule and moral judgments, we did not detect a difference between the simple effects of order across dependent measures.

General discussion

Addressing the replication crisis

Our studies successfully replicated order effects. Many efforts to replicate past findings frequently fail to show the same results, generating what has been called a replication crisis (Maxwell et al., 2015; Shrout & Rodgers, 2018), in which an initial study shows a statistically significant result but the replication does not. Contrary to concerns about the reliability of the published literature in psychology, our results strengthen the evidence for order effects in moral judgments.

The ways in which psychological factors affect our moral intuitions has attracted the interest of experimental moral philosophers (Earp et al., 2020, 2021; Lewis, 2020; Königs, 2022; Mihailov et al., 2021; Pözlner & Paulo, 2021). In a recent estimation (Cova et al., 2021), found that x-phi studies were successfully replicated about 70% of the time, as compared with less

than 50% in social psychology (Open Science Collaboration, 2015). Our work, thus, contributes to a high reproducibility in the experimental research on moral intuitions.

From replication to new results

Apart from the effort to replicate previous findings, we documented new, subtle effects, as well as negating the order effect, wherein participants reacted differently to changes in the scenarios. On the one hand, we produced the first evidence that order effects extend to cases with low emotional engagement such as remote killing and to specialized judgments about whether an action breaks a rule when it affects human beings. On the other hand, the order effect was not significant when the action of breaking a rule involved objects. We also produced first evidence that the order effect can be symmetrical, against previous theorizing that it is an asymmetrical transfer effect (though future work should aim to replicate the symmetrical transfer). Overall, the results support causal model theory but with a few caveats.

The need to develop causal model theory

Our research developed test cases for adjudicating between dual-process and the causal mapping explanations and further explored the scope of order effects. Experiment 1B investigated the extent to which order effects occur in dilemmas with minimized emotional force. We tested emotional engagement and causal mapping explanations by using the Trapdoor version of the Push dilemma. We reasoned that if order effects occur in these impersonal conditions, then the dual-process explanation is implausible. But if order effects do not occur in impersonal scenarios, then the causal model theory needs further development. The results show that order effects still occur in less emotional scenarios; in the Trapdoor version you have no physical contact and you do not exert personal force to push someone onto the tracks to save five innocent persons. Because it is less emotionally engaging, the action of dropping a victim onto the tracks using a trapdoor is considered more permissible than using personal force to push the victim onto the tracks (Cushman & Greene, 2012; Greene et al., 2009). For this reason, the result of Experiment 1B speaks against emotional engagement explanations. The fact that order effects occur when there is no prepotent emotion to exert a strong influence on moral judgment does not support dual-process explanations that appeal to differential emotional engagement.

We do not want to suggest that emotions play no role in how the order effect influences moral judgment. Our moral judgments heavily depend on emotions (Nichols, 2004; Prinz, 2007). We saw in our first experiment how

differences in emotional salience were manifest. In Experiment 1A the order effect was clear because there was a big difference between Switch and Push (highly emotionally salient). But in Experiment 1B, the difference in moral ratings between Switch and Trapdoor (less emotionally salient) is smaller, and, thus, the order effect was smaller. This confirms the prediction of causal model theory that order effects depend on the default evaluation of the scenarios. The bigger the difference in default evaluations, the bigger the order effect. When the default evaluations do not differ very much, the effect will be smaller.

Yet, the fact that the strength of the effect depends on differences in emotional salience should not be interpreted as evidence for a dual-process explanation. It rather shows that causal model theory needs to be developed to include interplays between emotional engagement and computational models of causal mapping. In fact, its theoretical framework has the potential to clarify where exactly emotions intervene. The causal mapping explanation consists of two parts: (i) default evaluations and highlighting and (ii) selective highlighting and mapping. There needs to be different default evaluations of the scenarios in order to highlight different causal paths from an intervention to its outcome. For example, most people approve of redirecting a runaway train, killing one person, but saving five, whereas they tend to disapprove of pushing one person onto the tracks to save five. Switch highlights the causal path that leads from the intervention to the good outcome, whereas Push highlights the causal path that leads from the intervention to the bad outcome. The order effect emerges when the highlighted causal structure can be mapped from one moral dilemma to another. Looking at the two parts of causal model theory, emotions can contribute to the default evaluation of a scenario, influencing the severity of the moral judgment, but they do not generate the order effect. Our moral judgments change when the highlighted causal path from the first dilemma can be mapped onto a similar path of the causal model of the second dilemma. Notice that the severity of a moral judgment can be caused by other factors, not only emotions. So, causal model theory should look at how multiple psychological factors highlight different causal structures in the initial computational stage – the default evaluation.

The relevance of causal structure depends on moral ontology

In the second experiment, we further tested if order effects occur in impersonal scenarios. Shaun (Nichols & Mallon, 2006) argued that emotion-based explanations of moral intuitions neglect the contribution of rules to judging moral dilemmas. In their empirical studies, they found the same pattern of moral evaluation about versions of the dilemmas that have minimized emotional force. Participants were more likely to say that the rule was

broken in the impersonal Push case than in the impersonal Switch case. We used their impersonal scenarios and replicated the results. Throwing a cup to save five cups breaks the rule more than redirecting the toy train to a side track, destroying one cup, but saving five. But even though the participants in our experiment recognized a slight moral distinction between throwing vs redirecting, there was no order effect. This result seems to indicate potential limits for the causal model theory because the impersonal cases we used have the same causal structures as in the classical trolley dilemmas, where order effects are robustly documented.

One potential explanation is that the default evaluations differed slightly. Participants were more likely to say that the rule was broken in the impersonal Push case than in the impersonal Switch case, but this contrast was not strong in our study. However, in Experiment 1B, the difference in moral ratings between standard Switch and Trapdoor (less emotionally salient) was also smaller and we still obtained an order effect (which was consequently smaller). A more plausible explanation, which opens new directions of research, is that changes in the moral ontology of the scenarios makes the causal structures from an intervention to an outcome less relevant. Substituting objects for persons can shift the focus from causal paths to outcomes because objects do not have the same moral status as persons. The individual rights of persons constrain which actions are permissible for the greatest good. If this explanation is right, then the lack of an order effect in the teacups scenarios does not oppose the causal model theory. It only circumscribes the scope of the theory. If the causal paths from an intervention to an outcome are relevant given the moral status of the entities involved, then we should expect an order effect. But if the individual entities have no intrinsic value attached, then causal paths become less relevant, and so, we can predict the lack of an order effect. Future research can explore how different categories of moral status contribute to the relevance of causal structures in generating an outcome. For example, we should compare saving the environment by sacrificing parts of the environment, saving private property by sacrificing parts of the private property, or saving works of art by sacrificing other works of art.

Symmetric order effects: a puzzle

Order effects are assumed to be essentially asymmetrical. Moral intuitions about Switch are subject to order effects, whereas moral intuitions about the Push dilemma stay constant regardless of its position in a series of moral dilemmas presented to subjects. A high approval of redirecting a threat does not reduce the condemnation of using someone as a means to stop a threat, but the condemnation of using someone as a means reduces the approval of redirecting a threat. This is why every theory of transfer effects between

evaluations of moral dilemmas focuses on explaining this asymmetry (e.g. (Horne et al., 2013; Lanteri et al., 2008; Lombrozo, 2009; Petrinovich & O’neill, 1996; Schwitzgebel & Cushman, 2012; Wiegmann et al., 2012). Contrary to established theory, experiment 2B has shown that transfer effects can be symmetrical. Instead of asking whether an action violated a rule, we asked participants whether the action of breaking teacups *should* be performed. Both actions (throwing a teacup to save five and redirecting the train) received very high approval rates. In conditions of high ratings of permissibility, and keeping the same causal structures as in Switch and Push, Teacup Push can become more permissible when Teacup Switch is presented first and *vice versa*. It is hard to explain this symmetry because the theoretical frameworks developed so far have been molded on asymmetric transfer effects.

We acknowledge several possibilities. If causal model theory does not predict the symmetrical effects, we might need a different theory to explain these effects and once we have the theory to explain symmetrical effects, it will provide us with a general framework for understanding order effects. Another possibility is that we will still need causal model theory to explain order effects in life and death situations, but it is not applicable to object-scenarios; this is supported by our contention that causal paths depend on the moral status of sacrificed entities. We might have to look at other psychological concepts to understand why the transfer is symmetrical in object-scenarios. Maybe, some version of the anchor effect (Kahneman, 1992) is generating the transfer in both directions. The high permissibility of any scenario presented first suggests an anchor of permissibility that is transferred to the subsequent evaluation. However, it is not clear what factors generate the transfer between two scenarios with different causal structures. Symmetric order effects in impersonal scenarios remain a puzzle.

Extending the scope of order effects to rule judgments

By asking participants whether an action violates a rule, instead of asking whether an action is impermissible, the salient features of the scenarios would become less emotionally engaging. In experiment 2A, when we asked participants whether throwing a teacup or redirecting the train to minimize overall damage violates a rule, we did not obtain an order effect. This result raised issues about the scope of order effects. It seemed that the effect did not extend to specialized judgments about rule violation. However, in Experiment 3 we wanted to investigate whether rule judgments are susceptible to order effects when persons are involved. In contrast to the lack of order effects in Experiment 2A, participants were more likely to state that the Switch dilemma violates the no killing rule when preceded by the Push dilemma. We obtained evidence that order effects extend to rule-breaking

judgments, when the actions affect persons. How do we explain the lack of order effect on rule judgments in Experiment 2 vs the presence of order effect on rule judgments in Experiment 3? One attractive explanation is the changes in the moral ontology of the scenarios. We stated previously the hypothesis that the moral ontology of the scenarios makes the causal structures from an intervention to an outcome less relevant. Comparative results from Experiments 2 and 3 show major differences in the default evaluations of breaking teacups as a means violates the rule versus sacrificing people as a means violates the rule. Most participants considered that pushing a person off the footbridge definitely violates the rule “do not kill”. Whereas most participants were tolerant of breaking teacups, showing less severe judgments that damaging a teacup to minimize negative outcomes violates a rule. The moral status of persons make the causal paths more relevant and, consequently, we see greater default evaluations between the standard trolley dilemmas. So, framing trade-offs in terms of rule-breaking does not nullify the order effect, but substituting objects for persons does.

Extending the scope of order effects to rule judgments has implications for theories of moral learning about how we acquire moral rules and judge their application. An adequate moral psychology must explain the susceptibility of rule judgments to order effects. According to causal model theory, ambiguous causal paths from an intervention to its outcome are essential for obtaining the order effect. So, moral learning has to be sensitive to the processes highlighted by causal model theory. But what features of learning moral rules preserve the ambiguous causal paths from an intervention to its outcome? The distinction made by (Nichols, 2021) between narrow scope rules and wide scope rules might be useful to understand why rule judgments are subject to order effects. A rule has a narrow scope when the rule being taught prohibits agents from producing a certain consequence (e.g., it is wrong for an agent to scratch a car). A rule has a wider scope when the rule being taught prohibits agents from producing a certain consequence and also from allowing such a consequence to come about or persist (e.g., it is wrong for an agent to scratch a car or allow a car to be scratched). It is plausible to think order effects influence rule judgments because rules with wider scope have more causal paths from interventions to outcomes.

Objects instead of persons: implications for policy decision making

Though we do not know how to explain them yet, symmetric order effects have implications for policy decision making. If a highly permissible action makes the subsequent evaluation of another action more permissible than it is judged in isolation, then policy makers and other stakeholders might be interested in shaping people’s judgments in their favor. By adding high permissibility scenarios alongside moderately permissible scenarios we can

make the latter even more permissible. Our results also show that substituting objects for people increases significantly the approval rate of sacrificing one kind of entity to save many other entities of the same kind. If people's willingness to minimize collateral damage greatly increases when trade-offs involve objects, then this effect can be used to make people more willing to produce the greater good. For example, the outcomes of proposed actions and policies can be reframed in terms of trade-offs between numbers, material costs, property, and in general terminology that objectifies the outcomes, instead of affecting individual lives.

Conclusion

The findings of our experiments advance the discussions on emotion elicitation versus causal structure explanations of order effects. Against the emotion elicitation explanation, order effects extend to cases with low emotional engagement, thus supporting causal model theory. Moreover, we produced novel evidence that order effects influence specialized judgments about whether an action violates a rule. It seems that order effects are so powerful that they can influence different kinds of moral judgments (judgments about the impermissibility of actions and judgments about conformity to moral rules), which makes the effect more surprising. However, when agents sacrifice objects instead of persons, the order effect either disappears or becomes symmetrical, contrary to previous theorizing that it is an asymmetrical transfer effect. This means that causal model theory needs to be developed to include interplays between the moral status of sacrificed entities and computational models of causal mapping. Symmetric order effects remain a puzzle, motivating future research for finding an explanation.

Note

1. After reverse-coding rule judgments so that higher values would indicate compliance, we found that the difference between the simple effects of order across dependent measures was not statistically significant, $B = 0.42$, $t(529) = 1.18$, $p = .24$.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS – UEFISCDI, project number PN-III-P4-ID-PCE-2020–0521, within PNCDI III; a grant from the Spanish Ministry of Science and Innovation (grant number:

PID2020.119791RA.I00); an Emmy Noether grant of the German Research Foundation (DFG) [grant number 391304769].

ORCID

Emilian Mihailov  <http://orcid.org/0000-0002-0221-7209>

References

- Bortolotti, L. (2011). Does reflection lead to wise choices? *Philosophical Explorations*, 14(3), 297–313. <https://doi.org/10.1080/13869795.2011.594962>
- Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., & Zhou, X. (2021). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*, 12(1), 9–44. <https://doi.org/10.1007/s13164-018-0400-9>
- Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, 7(3), 269–279. <https://doi.org/10.1080/17470919.2011.614000>
- Earp, B. D., Demaree-Cotton, J., Dunn, M., Dranseika, V., Everett, J. A., Feltz, A., & Tobia, K. (2020). Experimental philosophical bioethics. *AJOB Empirical Bioethics*, 11(1), 30–33. <https://doi.org/10.1080/23294515.2020.1714792>
- Earp, B. D., Lewis, J., Dranseika, V., & Hannikainen, I. R. (2021). Experimental philosophical bioethics and normative inference. *Theoretical Medicine and Bioethics*, 42(3), 91–111. <https://doi.org/10.1007/s11017-021-09546-z>
- Greene, J. D. (2008). The secret joke of Kant's soul. *Moral Psychology*, 3, 35–79.
- Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*, 124(4), 695–726. <https://doi.org/10.1086/675875>
- 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(3), 364–371. <https://doi.org/10.1016/j.cognition.2009.02.001>
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400. <https://doi.org/10.1016/j.neuron.2004.09.027>
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108. <https://doi.org/10.1126/science.1062872>
- Haidt, J. (2013). Moral psychology for the twenty-first century. *Journal of Moral Education*, 42(3), 281–297. <https://doi.org/10.1080/03057240.2013.817327>
- Horne, Z., Powell, D., & Spino, J. (2013). Belief updating in moral dilemmas. *Review of Philosophy and Psychology*, 4(4), 705–714. <https://doi.org/10.1007/s13164-013-0159-y>
- Kahneman, D. (1992). Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes*, 51(2), 296–312. [https://doi.org/10.1016/0749-5978\(92\)90015-Y](https://doi.org/10.1016/0749-5978(92)90015-Y)
- Königs, P. (2022). *Problems for moral Debunkers: On the logic and limits of empirically informed ethics*. Walter de Gruyter GmbH & Co KG.
- Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics*, 83(4), 789–804. <https://doi.org/10.1007/s10551-008-9665-8>

- Lewis, J. (2020). From x-phi to bioxphi: Lessons in conceptual analysis 2.0. *AJOB Empirical Bioethics*, 11(1), 34–36. <https://doi.org/10.1080/23294515.2019.1705430>
- Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, 33(2), 273–286. <https://doi.org/10.1111/j.1551-6709.2009.01013.x>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, 70(6), 487. <https://doi.org/10.1037/a0039400>
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254. <https://doi.org/10.1037/0033-295X.100.2.254>
- Mihailov, E. (2016). Is deontology a moral confabulation? *Neuroethics*, 9(1), 1–13. <https://doi.org/10.1007/s12152-015-9244-5>
- Mihailov, E., Hannikainen, I. R., & Earp, B. D. (2021). Advancing methods in empirical bioethics: Bioxphi meets digital technologies. *The American Journal of Bioethics*, 21(6), 53–56. <https://doi.org/10.1080/15265161.2021.1915417>
- Nichols, S. (2004). *Sentimental rules: On the natural foundations of moral judgment*. Oxford University Press.
- Nichols, S. (2021). *Rational rules: Towards a theory of moral learning*. Oxford University Press.
- Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, 100(3), 530–542. <https://doi.org/10.1016/j.cognition.2005.07.005>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 4716. <https://doi.org/10.1126/science.aac4716>
- Petrinovich, L., & O’neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17(3), 145–171. [https://doi.org/10.1016/0162-3095\(96\)00041-6](https://doi.org/10.1016/0162-3095(96)00041-6)
- Pölzler, T., & Paulo, N. (2021). Thought experiments and experimental ethics. *Inquiry*, 1–29. <https://doi.org/10.1080/0020174X.2021.1916218>
- Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.
- Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27(2), 135–153. <https://doi.org/10.1111/j.1468-0017.2012.01438.x>
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69(1), 487–510. <https://doi.org/10.1146/annurev-psych-122216-011845>
- Wiegmann, A., Horvath, J., & Meyer, K. (2020). Intuitive expertise and irrelevant options. *Oxford Studies in Experimental Philosophy*, 3(3), 275.
- Wiegmann, A., & Okan, Y. (2012). Order effects in moral judgment searching for an explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 34(34), 1143–1148.
- Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, 25(6), 813–836. <https://doi.org/10.1080/09515089.2011.631995>
- Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, 131(1), 28–43. <https://doi.org/10.1016/j.cognition.2013.12.004>