Moral Reasoning with Multiple Effects: Justification and Moral Responsibility for Side Effects

Neele Engelmann (neele.engelmann@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Germany

Abstract

Many actions have both an intended primary effect and unintended, but foreseen side effects. In two experiments we investigated how people morally evaluate such situations. While a negative side effect was held constant across conditions in Experiment 1, we varied features of the positive primary effect. We found that judgments of moral justification of actions were sensitive to the numerical ratios of helped versus harmed entities as well as to the kind of state change that was induced by an agent's action (saving entities from harm versus improving their status quo). Judgments of moral responsibility for side effects were only sensitive to the latter manipulation. In Experiment 2, we found initial support for a subjective utilitarian explanation of the moral justification judgments.

Keywords: Moral Reasoning, Causal Reasoning

Introduction

Research on moral judgments often probes people's intuitions about moral dilemmas. One of the most famous and well-studied dilemmas is the so-called trolley problem (Foot, 1967). In the side effect variant of trolley dilemmas, agents have a choice between letting a runaway trolley kill several people or an action that redirects the trolley to a different track where it would kill fewer people. The primary question in these studies is typically whether it is morally permissible to act. Many factors have been identified that influence people's intuition about this question (for an overview see Waldmann, Nagel, & Wiegmann, 2012).

The two dominant normative ethical approaches, utilitarianism and nonconsequentialism, largely agree in this situation. According to utilitarian recommendations, the action should be performed whenever its positive consequences outweigh the negative effects. Nonconsequentialist theories, such as the *Doctrine of Double Effect* (DDE, see Mikhail, 2011), arrive at similar conclusions for this case. The focus of the DDE and nonconsequentialism in general lies on the causal structure mediating acts and outcomes. In the side effect variant of the trolley dilemma, acting is considered permissible because the negative effect is not an intended means, but merely a foreseen side effect, and is not out of proportion to the positive effect. Psychological research on the side effect dilemma has shown that subjects indeed take the alternative outcomes into account when assessing the action's permissibility (e.g., Mikhail, 2011; Cohen & Ahn, 2016).

Evaluating Actions and their Side Effects

The focus of research on trolley dilemmas is on how people evaluate the permissibility of an action that causes two outcomes. All theories assume that in the side effect dilemma, both outcomes are compared and affect the moral evaluation, but little is known about the functional form of this comparison. A typical claim is that harming is permissible if the good outweighs the bad, but it is unclear whether this decision is just based on a simple categorical decision about which value is larger, or whether gradual differences between outcome values affect the decision. Few studies have systematically manipulated the numbers of victims that are saved or harmed in moral dilemmas (but see Cohen & Ahn, 2016; Waldmann & Wiegmann, 2012).

Cohen and Ahn (2016) postulate a subjective utilitarian analysis. For each item or set of items (e.g., 5 people) subjects provided an estimate of their personal value. The personal values were affected by the type of item and their number, although the number turned out to have a relatively small effect. These estimates of the personal values were then used to predict subjects' judgments about choice situations in which one set of items is about to be destroyed (or killed) when no action is taken but saved when the agent acts, which in turn would destroy (kill) a second set of items. According to the categorical utilitarian decision strategy, the action is chosen that saves items with the higher personal value. The model also predicts reaction times: Given that the comparison is typically influenced by uncertainty, a faster reaction time is predicted when the difference between values becomes larger.

One key goal of our project is to provide further tests of the subjective utilitarian model. A salient problem of the current version of the model is that it lacks generality. Its predictions are based on the personal values of the items involved in the outcomes but this model neglects that actions cause transitions between states. An evaluation of an action thus needs to take into account the values of the states of the items in the presence versus the absence of the action. Cohen and Ahn (2016) did not consider how subjects assess the personal values of the items in their destroyed or dead states, probably because this was the standard state in the absence of an action across all item sets. However, actions can also improve the state of items that otherwise would be in a normal state, or they could be saved from a disease that would harm, but not kill them. To provide a full utilitarian account of how outcomes of actions should be evaluated we suggest that people compute contrasts between the personal values of the outcomes in the presence versus the absence of the target action. We will also argue that sometimes more than two states need to be considered. We will present an experiment that presents a wider range of actions, which allows us to test our subjective utilitarian model against theories that are not sensitive to different types of states in the presence and absence of the

target action.

A further focus of our study is to investigate how the relation between the number of people that are positively or negatively affected by the action influences the degree to which people find the action morally justifiable and the agent morally responsible for the outcomes, especially the negative side effect. We systematically manipulated the numbers involving the positive primary effect while holding the negative side effect constant (see also Waldmann & Wiegmann, 2012, for a similar design but different tasks). For example, in one of our experimental conditions, ten members of a tribe are harmed by an action that would save a varying number of members of a different tribe. According to Cohen and Ahn's (2016) model, an act involving a negative side effect should lead to faster reaction times the more entities are helped compared to harmed. If reaction times indicate certainty about an act's permissibility, one can also derive from this theory the prediction that justification ratings should be affected in a similar manner.

One limitation of trolley studies is that so far they have focused on a particular type of situation in which the primary goal is to save victims that otherwise would be killed. It may well be that acts that lead to negative side effects are only considered justified when the primary effect targets entities that, prior to the intervention, are threatened to be harmed. The primary effect may be less effective as a justification when the act is supererogatory and just improves the states of entities that prior to the act are in a normal state. For example, instead of saving varying numbers of victims from grave harm, the people may be fine prior to the act, with the act just improving their health and living conditions. The theory proposed by Cohen and Ahn (2016) does not make predictions here because it only takes into account the personal values of the entities in their intact state. We will in Experiment 2 test a modified account that postulates that subjects take into account personal values of states in both the presence and the absence of an action. This account makes predictions for the difference between saving entities or improving their states.

Another limitation of the typical trolley dilemma studies is that they have focused on situations in which saving and harming are causally achieved by redirecting a harmful entity (the runaway trolley). In order to widen the range of studied dilemmas and to be able to manipulate the prior state of the entities involved in the primary goal, we tested a different causal structure in which a helpful act rather than a threat was redirected (see also Ritov & Baron, 1999; Bartels & Medin, 2007). For example, in the condition involving two tribes, a dam may be opened that redirects water from one tribe to the other. Redirecting might save tribe members from a negative state or improve their normal situation.

Finally, a limitation of previous research is that the test question typically focuses only on the act leading to two outcomes. We are also interested in how people evaluate the two outcomes individually. We therefore added as test questions requests to judge moral responsibility for the negative side effect. Our goal was to test whether these judgments are also influenced by the value of the primary effect (e.g., number of victims). If subjects just focus on the side effect, the primary effect should not have an influence. However, if the status quo or the number of affected entities are used as exonerating factors, their impact should also be seen in moral responsibility ratings for the side effect.

Together, these manipulations and the studied judgments widen the focus of previous work on people's moral intuitions about cases with multiple effects. The aim of the first experiment was to test whether the relation between primary and side effect of an action influences moral justification assessments. Moreover we were interested in whether the primary effect influences moral responsibility assessments for a bad side effect. We tested whether these two types of moral queries are affected by the kind and number of entities that are potentially harmed or saved, and by their state change due to a possible intervention. Experiment 2 inquires to what extent the results of Experiment 1 can be explained by a subjective-utilitarian framework.

Experiment 1

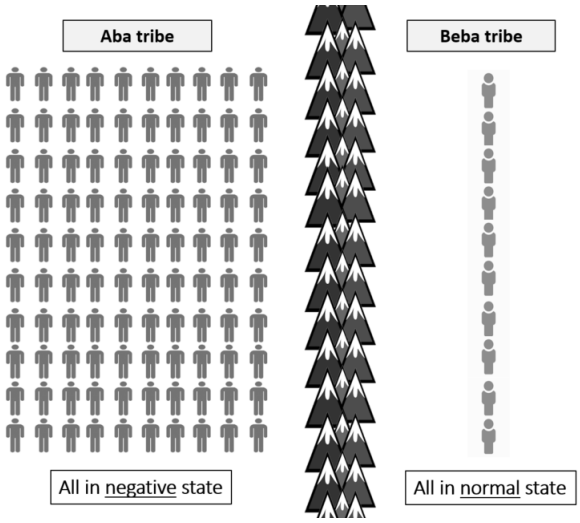
We constructed three scenarios in which an agent decides to perform an action with a positive, intended primary effect and a negative, unintended (but foreseen) side effect. The negative side effect was held constant across conditions and always consisted in killing 10 entities (people, animals, or plants). We varied whether 1, 5, 20 or 100 entities benefitted from the action. Furthermore, we manipulated whether these entities were in a negative or a neutral state prior to the action. In the situations in which the entities were in a negative state, they would have died without the agent's action; in the contrasted normal state condition, the action would merely cause additional benefit (e.g., people improving their living conditions or plants growing better).

Design, Material and Procedure[1] 450 participants were recruited via the UK based platform Prolific Academics for a compensation of £0.25 (£6 per hour). Inclusion criteria were a minimum age of 18 years, English as a first language, a study approval rate on the platform of at least 90%, and not having participated in previous studies with similar material. Participants were randomly allocated to one of 24 conditions (primary effect: saving vs. improving; number of helped entities: 1 vs. 5 vs. 20 vs. 100; affected entities: people vs. animals vs. plants). Here is an example vignette from the *saving* conditions. The example describes a condition in which 100 people are saved by the action, who otherwise would die:

Suzy is the prime minister of Tolosia, a mountainous country with many distant and small villages. The villages are populated by different indigenous tribes. She is authorised to make all decisions about the country's welfare that she deems appropriate. One day, she learns that a mountain village has

[1]The full material and data for both experiments are available under `https://osf.io/jcux6/`

suffered from an ongoing drought that left its inhabitants, the Aba tribe, in poor health due to lack of water. Exactly 100 people belong to the Aba tribe, all of whom are in critical condition and will die if nothing is done. Suzy could order to open a dam that would redirect a mountain river towards the Aba tribe. With a quick water supply, the 100 members of the Aba tribe could recover. However, the redirection of the river could also cause a lack of water in another mountain village, home to the Beba tribe, causing its 10 members to die of thirst within a few days. All of the 10 members of the Beba tribe are fine at the moment. Since both mountain villages are inaccessible to any means of transport, redirecting the river is the only currently available measure to influence the well-being of the two tribes. Here is a schematic representation of the two tribes and the current state of their members:



Suzy is aware of all the facts. She wants the 100 members of the Aba tribe to recover, but also not to cause any harm to the 10 members of the Beba tribe. She decides to open the dam and redirect the mountain river. All of the 100 members of the Aba tribe recover. However, all of the 10 members of the Beba tribe die within a few days.

The figure was followed by the instruction: "Here is a schematic representation of the tribes and their state after the river has been redirected" along with the same figure as above in which the lower labels now read "all in normal state" for the Aba tribe and "all dead" for the Beba tribe. In the corresponding *improving* condition, the vignette stated that the Aba tribe could vastly improve their health and lifespan with an extra water supply (no threat by a drought was mentioned). In the subsequent test phase participants were asked to rate the extent to which they saw the agent's action as morally justified ("To what extent was Suzy's action morally justified ?"). The moral responsibility question focused on the side effect ("To what extent is Suzy morally responsible for the members of the Beba tribe dying?"). As a control, we also asked about the primary goal ("To what extent is Suzy morally responsible for the members of the Aba tribe improving their health?"). Ratings were given on a 10-point Likert scale with the endpoints labelled "not at all" (1) and "fully" (10). Justification and responsibility questions were presented on two separate pages, with page order counterbalanced between participants; order of the two responsibility questions within the respective page was randomized. Subsequently, two manipulation check questions assessed whether people had correctly understood how many entities were harmed and helped in the scenario.

Results and Discussion 18 participants were excluded for failing at least one of the manipulation check questions, leaving data of 432 participants for the analysis (mean age = 34.4, SD = 11.93). We conducted a 2 (primary effect) x 3 (entity) x 4 (numbers) x 2 (test question order) ANOVA for each of the three dependent variables. Since our study is partly exploratory, we used a conservative significance threshold that takes into account the number of tests in the models (here: $p<.003$). Results for the 432 valid subjects can be seen in Figure 1.

Moral justification ratings were higher the more entities were helped compared to harmed, $F_{(3,\,384)} = 8.81$, $p<.001$, $\eta^2 = .06$. Additionally, a large effect was obtained between the conditions saving and improving, $F_{(1,384)} = 130.74$, $p<.001$, $\eta^2 = .25$. The interaction was not significant (p=.37). Participants gave the highest justification ratings when the primary effect was an instance of saving and more entities were saved than killed.

Post hoc tests (Newman-Keuls) for the saving condition revealed that the case in which only one entity was saved as a primary effect was judged significantly less morally justified than the cases in which twenty or a hundred entities were saved. The other cases did not differ significantly from each other. In the *improving* condition, post hoc tests showed no significant differences.

There was also a main effect of vignette. Subjects considered the action as most morally justified when the affected entities were plants (M = 5.23, SD = 2.6), followed by animals (M = 4.41, SD = 2.52), and people (M = 3.84, SD = 2.77), $F_{(2,384)} = 14.39$, $p<.001$, $\eta^2 = .07$. A possible reason for this ordering might be that harming people may be seen as a harsher moral violation than harming plants and therefore less justifiable by good effects. Animals seem to be in the middle.

Additionally, a small unexpected order effect was found. Ratings were slightly higher when the moral justification question was presented after the moral responsibility questions (M = 4.88, SD = 2.71) compared to before (M = 4.12, SD = 2.62), $F_{(1,384)} = 12.51$, $p<.001$, $\eta^2 = .03$.

Moral responsibility ratings for the negative side effect were generally high, but not detectably influenced by the number of helped entities, $F_{(3,384)} = 0.35$, $p = .79$ (see Fig. 1). However, the ratings were lower when the action's primary effect was an instance of saving (M = 8.09, SD = 2.23) rather than improving (M = 9.12, SD = 1.59), $F_{(1,384)} = 33.51$, $p<.001$, $\eta^2 = .08$. The interaction was not significant (p=.61). *Moral responsibility ratings* for the positive primary effect were

high (M = 8.23, SD = 2.31) and not influenced by any manipulation.

In sum, the moral justification ratings of the action were sensitive to the relation between the primary and the side effect. The more entities were helped as a primary effect, the more justified the action was judged. This pattern shows that moral justification is a continuous quantity that is sensitive to the relative size of the outcomes. A novel result concerns the comparison between different status quos, which generated the largest effect. If entities are saved from a threat, the action was seen as substantially more justified than when the primary goal is just to improve states starting from a neutral state.

The fact that subjects took into account both the primary and the side effect in their justification judgments is predicted by both nonconsequentialist and utilitarian accounts. However, the specific theory proposed by Cohen and Ahn (2016) does not predict the largest effect in our experiment: Subjects clearly differentiated between saving entities versus improving their state. Simply using assessments of personal values of the entities does not predict these effects without taking into account the personal values of the states of the entities in the absence of the action. We will test a modified model that is sensitive to state changes in Experiment 2.

An interesting unexpected finding was that moral responsibility ratings proved insensitive to the number of helped entities, but were reduced when the action's primary effect was an instance of saving rather than improving. This latter effect makes it unlikely that the lack of an effect of number is due to a ceiling effect. A possible interpretation of this pattern may be that subjects tried to focus on the side effect alone but were influenced by features of the primary effect that have a large impact on justification, such as the status quo, rather than only a small effect, such as the numbers.[2]

Experiment 2

The aim of the second experiment is to investigate to what extent the effects observed in Experiment 1 could be explained by a a variant of a subjective utilitarian theory that in crucial aspects differs from the one proposed by Cohen and Ahn (2016). Cohen and Ahn (2016) modeled choices as decisions based on the personal values of the entities involved in the alternative outcomes. For example, the task in their second study was to choose which of two sets of items should be saved and which destroyed in a dilemma. The model claims that the differences between the personal values of the two sets of items predict judgments. The focus on the personal values of the items seems appropriate here because all actions represented a choice between leaving the items intact or destroying (or killing) them. This restriction of the task allowed Cohen and Ahn (2016) to focus on the personal values of the affected items. However, the model is a too restrictive as a general model of moral reasoning. We suggest that the focus should be on actions, which can cause transitions between various states, not only between the states dead and alive or intact and destroyed. For example, in our Experiment 1 we presented cases in which actions improved states of entities that prior to the intervention were in a normal state.

To overcome the limitations of the model proposed by Cohen and Ahn (2016), we here propose a variant of a subjective utilitarian theory that focuses on actions and models them as state changes. When people evaluate an action, they should be sensitive to both the outcomes in the presence of the action but also to what happens in the absence of the action. For example, an action that improves the state of an entity can be represented as the difference between the personal values of the improved state and the normal state prior to the action. More complex state transitions are conceivable, and in fact in Experiment 1 we presented scenarios in which the entities shifted between four possible states (normal, threatened, improved, dead). In the present study we collected assessments of personal values of all the entities for these four states and used these assessments to predict the justification judgments obtained in Experiment 1.

Figure 2 shows how we adapted our model to the cover stories in Experiment 1. In the example in Figure 2, 100 people are under the threat of dying prior to any action. In the absence of an action (i.e., omission) they would die, which is modeled here as the contrast of the personal values between death and a critical state (second component of Figure 2a). In the presence of the action, the people in critical state would be shifted into a normal, healthy state, here represented as the difference between the personal values of a critical versus a normal state (first component of Figure 2a). The overall utility of saving the people is modeled as the sum of these contrasts because the action both prevents the people from being killed and puts them from a critical into a healthy state. Thus, the representation of the saving action considers both the effects of the potential action and of its omission. In the case of improving (not depicted), the model simplifies to a contrast between the values of the improved versus the normal states. The second component in the equation in Figure 2a would amount to 0 in this case because there is no threat to the normal state. Finally, Figure 2b shows how we model the total utility of the action in a scenario with multiple effects: It is the sum of the median utilities of the primary effect (saving) and the harmful side effect (killing 10 people).

Design, Material and Procedure The design of our basic value estimation task largely follows the methodology described in Cohen and Ahn (2016) but assesses a wider range of possible states of entities. Like Cohen and Ahn (2016), we tested the influence of the numbers of entities (1 vs. 5 vs. 10 vs. 20 vs. 100) on personal value assessments in

[2] In this experiment, moral justification was assessed globally (i.e., for a whole action), while responsibility was assessed separately for the single effects. One might worry that this does not allow us to tell whether the differences between the two judgments are driven by the type of judgment or by the focus of the question on global or separate outcomes. We therefore conducted a follow-up study in which we fully crossed these two factors. We found that the type of judgment seems to be the driving factor. The study is available online along with materials and data.

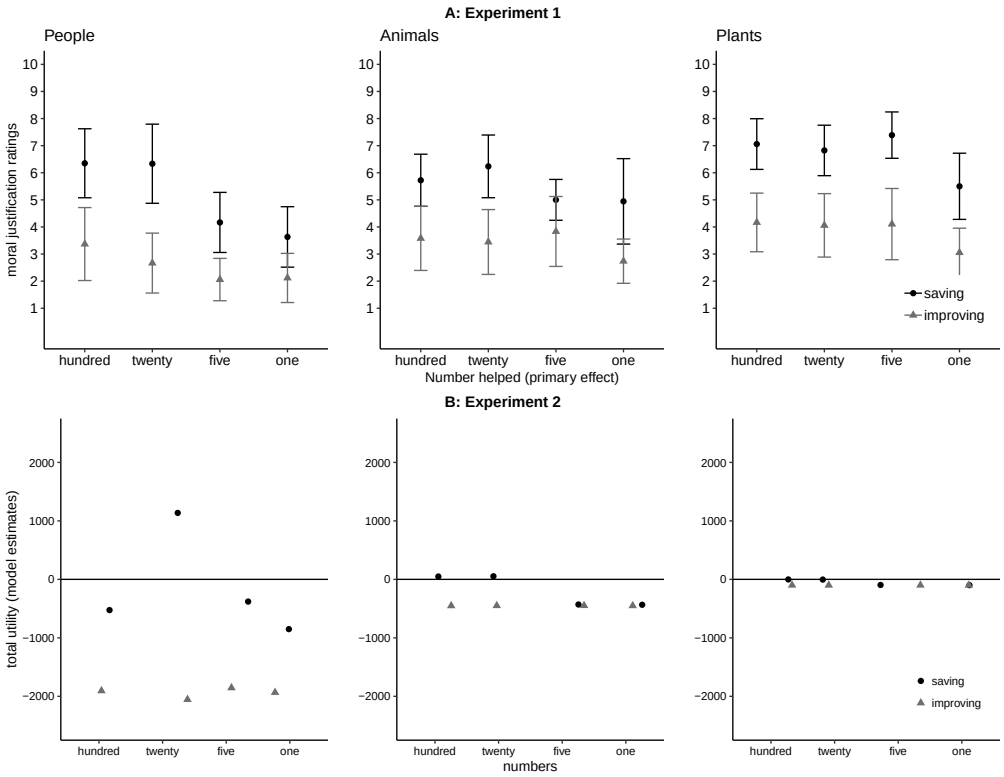


Figure 1: A: Means and 95% confidence intervals for moral justification ratings in Experiment 1, B: Total utility estimates generated by our model in Experiment 2.

separate experimental groups to avoid demand characteristics (i.e., participants feeling pressured to assign exactly five times the value of one entity to a group of five of the same entities). Within each group, we presented instances of people, fish and roses, each of them in all of the states that were described in Experiment 1 (normal vs. threatened vs. improved vs. dead). Thus, each participant judged 12 stimuli, in randomised order.[3] Like Cohen and Ahn, we presented people with a measuring standard to calibrate their value estimates. They were told that "one healthy chimpanzee" should be taken to have a value of 1000. If they valued any item half (or twice or any other ratio) as much as one healthy chimpanzee, they should assign the corresponding value to the item (e.g., a value of 500 if they value an item half as much as the chimpanzee). Participants were further instructed that "personal value" does not necessarily correspond to monetary value and that they should judge the entities' value in their *current* state. 250 participants (mean age = 36.6, SD = 13.5, 67% female, 32% male, 1% other) were recruited on Prolific Academics and completed the survey for a compensation of £0.40 (£6 per hour). Inclusion criteria were identical to Experiment 1, and not having participated in Experiment 1.

Results and Discussion To test our model, we used the value estimates of the four states of the entities to generate predictions for the justification assessments. Following the rationale outlined in Figure 2 we generated predictions for all 24 experimental conditions.The results are shown in Figure 1B. The total utilities overall capture the patterns found in Experiment 1, even though the maximal range of values was much wider for people cases compared to animals and plants (see Fig. 1A). Most importantly, the total utility estimates reflected the differences between improving versus saving, at least for people (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$) and animals (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$)[4]. In both cases the total utility for saving was larger than for improving, which mirrors the effects in Experiment 1. The corresponding effect for plants was not significant when correcting for multiple testing. Moreover, we did not find significant effects for the manipulation of the number of the affected entities for either people, animals or plants. But note that this effect was fairly small in Experiment 1 (and also in Cohen & Ahn, 2016). Also, this factor was the only one manipulated between subjects, which may have led to reduced sensitivity to this factor.

As an overall test of the fit of our model to the data of Experiment 1, we conducted a linear regression analysis with to-

[3] With the exception of the "10 entities" condition, which referred to the constant side effect. Here, we only needed estimations of each set of entities in their normal and dead states since the side effect entities never were in other states.

[4] We used again a conservative significance threshold that takes into account that we tested each factor separately for each entity category (here: $p < .017$).

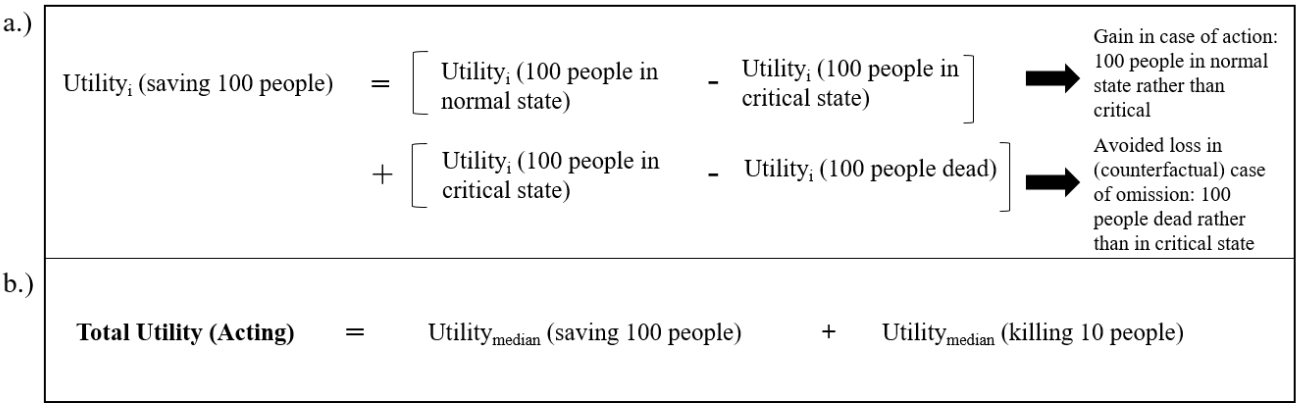


Figure 2: Rationale of our calculation of an action's total utility, spelled out for the example of the *saving 100 people* scenario. See text for explanation.

tal utilities estimated by our model as the predictor and mean moral justification ratings obtained in Experiment 1 as the criterion. The model fit the data well and explained a substantial amount of variance in the criterion, $F_{(1,22)} = 16.31$, $p<.001$, $R^2 = .43$, $RMSE = 1.14$.

General Discussion

The main goal of our study was to provide more fine-grained evidence on how moral judgments are influenced by characteristics of multiple effects of an action in dilemma situations. Experiment 1 showed that judgments of moral justification for the agent's action increased with more favourable ratios of helped compared to harmed entities, but were even more influenced by the change of state that was induced by the agent's action (saving vs. improving). Moral responsibility judgments for the negative side effect were only affected by the latter manipulation but not by the number of affected entities.

In Experiment 2 we tested a novel subjective utilitarian model that goes beyond previous proposals. Whereas Cohen and Ahn (2016) claimed that moral decisions are based on the personal values of the affected entities in their healthy or intact states, we argued that this assumption restricts their model to a small set of situations in which actions destroy or kill entities. Our goal was to propose a model that is more general. A basic assumption of our model is that actions can be modelled as state changes and that moral judgments are sensitive to both the states that entities are in prior and following a target action. This model allowed us to not only model cases of killing and saving but also, for example, cases of improvement.

Although our results in Experiment 2 showed that the new model explains a substantial amount of variance, it does not capture all effects. One reason for this may have been the necessary differences in the designs of Experiments 1 and 2. But there may be other reasons: For example, to demonstrate the increase of expressiveness of our model, we suggested a model for the cover stories of Experiment 1 that captures transitions between the four possible states mentioned there. Given that utility measurements are unreliable and influenced by additional factors, making the model more complex will certainly reduce its fit to the data.

Future research will also have to investigate whether there are alternative models that may also capture the results. As in the case of improving, we could, for example, generally use a more basic utilitarian model that only compares the two states in the presence versus absence of the action (e.g., dead vs. alive in the case of saving). Future research will need to test in greater detail the assumptions entering the different variants of the model.

We labeled our model "subjective utilitarian" because it was inspired by the theory of Cohen and Ahn (2016). However, we mentioned in the introduction that both utilitarian and nonconsequentalist theories predict that in side effect dilemmas the outcomes should be compared. Thus, our model may also be viewed as a component of a nonconsequentialist account. One possible way to test the two alternative theoretical possibilities is to take a closer look at the assumption that actions can be modeled as state changes. This assumption embodies the utilitarian claim that it is only the outcomes that matter, not the type of action leading to the outcomes. We suspect, however, that the type of action and the type of causal relations leading to the changes may also matter (see Kamm, 2007; Waldmann, Wiegmann, & Nagel, 2017). Future research will have to further explore these issues.

Acknowledgements

We thank Alex Wiegmann for helpful discussions about our utility model.

References

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*(1), 24–28.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381.

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

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

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational behavior and human decision processes*, *79*(2), 79–94.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The Oxford handbook of thinking and reasoning*, 364–389.

Waldmann, M. R., & Wiegmann, A. (2012). The role of the primary effect in the assessment of intentionality and morality. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, pp. 1102–1107). Austin, TX: Cognitive Science Society.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Tremolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge/Taylor Francis Group.