Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

How to weigh lives. A computational model of moral judgment in multiple-outcome structures

Neele Engelmann *, Michael R. Waldmann

Department of Psychology, University of Göttingen, Germany

ARTICLE INFO

Keywords:
Moral judgment
Moral dilemmas
Moral reasoning
Deontology
Consequentialism
Utilitarianism

ABSTRACT

When is it allowed to carry out an action that saves lives, but leads to the loss of others? While a minority of people may deny the permissibility of such actions categorically, most will probably say that the answer depends, among other factors, on the number of lives saved versus lives lost. Theories of moral reasoning acknowledge the importance of outcome trade-offs for moral judgments, but remain silent on the precise functional form of the psychological mechanism that determines their moral permissibility. An exception is Cohen and Ahn's (2016) subjective-utilitarian theory of moral judgment, but their model is currently limited to decisions in two-option life-and-death dilemmas. Our goal is to study other types of moral judgments in a larger set of cases. We propose a computational model based on sampling and integrating subjective utilities. Our model captures moral permissibility judgments about actions with multiple effects across a range of scenarios involving humans, animals, and plants, and is able to account for some response patterns that might otherwise be associated with deontological ethics. While our model can be embedded in a number of competing contemporary theories of moral reasoning, we argue that it would most fruitfully be combined with a causal model theory.

1. Introduction

Most of us will never be in the unlucky position of the agent in a trolley dilemma (Foot, 1967). Our moral concerns are usually much more mundane than the question of whether or not we should let one person get run over by a train in order to save five others from the same fate, for example. Some people, however, routinely make life-and-death decisions. Many political actions, take the allocation of healthcare resources as just one example, have outcomes that can be quantified in terms of lives saved versus lives lost. While most of us do not actively get a say in these large-scale matters, we judge those who do. Everyday moral discourse, be it in person or on social media, is rife with both condemnation and justification of actions which, more or less directly, trade off lives or other goods. Examples of such trade-offs are policies implementing speed limits in traffic, the introduction of social distancing measures during the Covid-19 pandemic, or the European Union closing its borders to refugees.

Trolley dilemmas have, in recent years, often been criticized for lacking such real-life context or for poorly predicting actual moral behaviour (see, for example, Bauman, McGraw, Bartels, & Warren, 2014; Bostyn, Sevenhant, & Roets, 2018; Schein, 2020). Against this criticism, others have argued that moral psychology does not only address the question of how people behave in real-world situations, but also what they judge to be right and wrong. Moral judgment, so the argument, is an interesting psychological phenomenon in its own right (Białek, Turpin, & Fugelsang, 2019). Furthermore, moral dilemmas are not always meant to be representative of actual situations. As Plunkett and Greene (2019) argue, contrasts between different artificial moral dilemmas can serve the same purpose as contrasts between visual stimuli in artificial optical illusions. They can expose the core mechanics that are untraceable in more content-laden "realistic" situations.

Inspired by an initially exclusively philosophical debate ignited by Foot (1967) and Thomson (1985), moral psychologists have now spent at least two decades empirically investigating people's intuitions about moral dilemmas. Mirroring the philosophical debate about trolley dilemmas, the dominant research strategy in psychology has been to keep the outcomes of an action constant and vary other factors of interest. This strategy has revealed some relatively stable patterns (see May, 2018; Waldmann, Nagel, & Wiegmann, 2012, for detailed overviews). Everything else being equal, people find it morally worse if a negative outcome is brought about intentionally rather than by accident (Cushman, 2008; Cushman, Young, & Hauser, 2006; Young & Saxe, 2011),

* Corresponding author at: Department of Psychology, University of Göttingen, Gosslerstraße 14, 37073 Göttingen, Germany.
E-mail address: neele.engelmann@uni-goettingen.de (N. Engelmann).

https://doi.org/10.1016/j.cognition.2021.104910
Received 15 December 2020; Received in revised form 15 September 2021; Accepted 16 September 2021

through an action rather than an omission (Cushman et al., 2006; Cushman & Young, 2011; Spranca, Minsk, & Baron, 1991, but see Willemsen & Reuter, 2016), as a causal means for a positive primary outcome rather than a side-effect (Cushman et al., 2006; Cushman & Young, 2011; Feltz & May, 2017; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007), and by so-called "personal force" or "battery" rather than indirectly (Greene et al., 2009; Hauser et al., 2007; Mikhail, 2007, 2011). Overall, all of these features taken together may constitute the prototype of a harmful, morally bad action (see Greene, 2013, p. 247).

In contrast to these studies, the focus of the present research is on the role of outcomes in moral judgments. A common response is to associate outcomes with consequentialist and acts with deontological ethical theories. However, outcomes play a role in all ethical frameworks, including deontological theories. For example, the deontological Doctrine of Double Effect (see Alexander & Moore, 2016, for an overview) holds that an action which causes serious harm (such as a person's death) can be morally permissible given that, among other things, the harm is outweighed by the action's positive effects. But can one death be considered as outweighed when two other lives are saved? Are there degrees of permissibility when a larger or smaller number of lives are saved? Further complications arise when the lives involved in a trade-off belong to different categories (e.g., people vs. animals) or lives are traded off against other goods, such as inanimate objects or abstract values. Any rule based on a simple numerical comparison will fail to be applicable as soon as trade-offs involve more than one kind of entity (while causing the death of one person to save five others may be permissible, it may not be permissible to cause one person's death in order to save five fish, for example). Normative philosophical theories cover a wide range of positions on both the kind of trade-offs that are allowed and the circumstances under which they are allowed (see Alexander & Moore, 2016). Psychologically, judging trade-offs between different kinds of entities can certainly be requested from subjects, as has recently been strikingly demonstrated by the "moral machine" experiment (Awad et al., 2018). Here, participants made choices in dilemmas pitting a wide range of possible victims against each other (differing in number, age, role in society, and other features). Some stable patterns emerged, for example a preference to save more rather than fewer lives, or to save humans rather than animals. However, this study does not answer the question of how different entities are compared.

1.1. The role of outcomes in psychological theories of moral judgment

While most of the general psychological theories of moral judgment do not spell out an outcome integration mechanism in detail, all of them assume such a mechanism. Dual-process accounts posit that there are two competing modes of moral reasoning, with the first one reacting to situational features, such as personal force, intentionality, or the distinction between action and omission. In the theory of Greene and colleagues (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001) a slow and deliberative second process follows and rationally determines whether the outcome trade-off is favourable or not. In Cushman's (2013) and Crockett's (2013) versions of dual-process theories, this second process is described in more detail and characterized as a model-based algorithm which evaluates an action based on all immediate outcomes in a specific situation. Still, the focus of these theories is on the "big picture" of moral judgment, for example on explaining to what extent it is driven by affective and cognitive processes. Figuring out the details of outcome trade-offs is not the main aim (but see Shenhav & Greene, 2010).

A competitor of dual-process theories is Mikhail's Universal Moral Grammar theory (Mikhail, 2007, 2011), which is inspired by deontological ethics. In this theory, the *Doctrine of Double Effect* (DDE) plays a central role with its focus on the distinction between intended and foreseen harm. As mentioned above, the DDE addresses outcome trade-offs in its *proportionality condition*: for an action that causes serious harm to be morally permissible, the harm in question must, among other specified conditions, not be "out of proportion" to the action's positive effects. Mikhail has proposed a formalism for comparing outcomes, the so-called *Moral Calculus of Risk* (Mikhail, 2011, pp. 140–142). It consists of the values of the positive and negative outcomes of a candidate action and considers their respective probabilities. Furthermore, the "Necessity of the Risk" is included, which is the probability that the agent's purpose (default: bringing about the positive effect) would not be achieved without risking the negative effect. Briefly put, an action should become more permissible with a better expected value, but less permissible when it is more likely that the positive effect could also have been produced without risking the negative effect at all. Dilemmas are defined by a limited set of options: the agent cannot bring about a positive effect without also causing a negative one. Therefore, the Moral Risk Calculus will, in most dilemma scenarios, come down to a simple expected value calculation: the actual numbers of lives saved versus lost, weighted by the respective probabilities of them being saved versus lost given the action. To our knowledge, the Moral Risk Calculus has not been subjected to a systematic empirical investigation.

Cohen and Ahn (2016) recently defended a novel one-system approach to reasoning about outcomes in moral scenarios, inspired by philosophical utilitarianism and decision theory (see Steele & Stefánsson, 2020, for an overview). According to their *Subjective-Utilitarian Theory of Moral Judgment* (henceforth: STMJ), *only* outcomes matter for evaluating a moral dilemma, which is contrary to all psychological theories of moral judgments discussed above. More specifically, the value that an observer attaches to the outcomes of each available course of action determines, according to STMJ, the probability that this course of action is selected as morally preferable. The proposed mechanism is formalized, and yields quantitative predictions about judgments in moral dilemmas. Applied to the standard trolley case with five lives saved and one life lost, the typical majority opinion that acting is permissible is explained by the fact that, all else being equal, people think that five lives are more important or valuable than just one. However, STMJ does not claim that people simply count and compare lives saved and lost. Instead, it is possible that one particular life (e.g., of a close friend) has a higher subjective value to someone evaluating the dilemma than the lives of five others combined. In this case, the theory predicts that the action that saves this one person is favoured.

The underlying decision process is described as a cumulative sampling of values from internal distributions until a decision criterion is reached. The form of the mechanism is inspired by a random-walk decision process, a model that has been confirmed in other domains, such as visual perception (e.g., Ratcliff & Rouder, 1998). Spelled out for the standard trolley case, STMJ would claim that an observer has some internal representation of the value of one life, and also of the value of five lives. These representations take the form of Gaussian distributions. The mean of the value distribution for five lives is likely to be higher than the mean of the value distribution of one life, but the two distributions might also overlap to some extent. When an observer is faced with the task of identifying the higher-valued stimulus of a pair, they repeatedly sample and compare value pairs from both distributions. At some point, enough evidence will have been accumulated to consciously conclude that five lives have the higher value. Crucially, the more two distributions overlap, the longer this process will take, resulting in the experience of a harder decision and in longer reaction times. More overlap between two distributions also creates noise, sometimes leading to prediction errors in which the option with a lower mean value dominates.

Cohen and Ahn (2016) had participants explicitly indicate the subjective values of a variety of stimuli: people, animals, and inanimate objects. Values were elicited by asking participants to compare each item against a standard with a fixed, arbitrary value (a chimpanzee with a value of 1000). From the values participants generated, a distribution for each item and the overlap between any two distributions was

determined. Different participants then completed a series of moral dilemma tasks using the pretested set of stimuli. In each trial, two stimuli were randomly drawn and presented together in a situation in which only one of them could be saved, and the other one would be killed or destroyed. Participants had to answer the question "*Would you save [Item A], causing [Item B] to be killed/destroyed?*" Their choices as well as response times were recorded for each trial. The overlap between value distributions of any two items turned out to predict both measures very well. Based on these results, Cohen and Ahn (2016) conclude that people are subjective utilitarians when it comes to moral judgments – that is, that they base their moral judgment only on the subjective values of an action's outcomes. Predictions of STMJ converge with findings from different lines of research. For example, when weighing different numbers of lives against each other, STMJ would not predict that the mean of the value distribution for "five lives" is five times higher than the mean for "one life". Instead, a concave relationship between the number of lives and values is assumed (see also Cromley & Cohen, 2019). And indeed, the distributions of some items, such as "one adult" and "five adults", showed a near complete overlap in Cohen and Ahn's (2016) studies, indicating that five lives were only valued marginally higher than one life. In brief, STMJ is parsimonious, firmly grounds moral judgment in well-established domain-general mechanisms, and makes quantitative predictions about choices in moral dilemmas that could be confirmed in several experiments.

Nonetheless, there are also some shortcomings and open questions. In its current form, the model is only applicable to classic moral dilemmas in which the action under consideration leads to a trade-off between saving and killing (or destroying). While such dilemmas are important, there are many decisions with multiple outcomes that are not life-and-death dilemmas. For example, a political action may benefit some groups at the expense of others (such as tax alleviations for top incomes), while nothing at all would have changed if the action had not been performed. Moreover, killing versus saving does not exhaust the realm of moral actions. An agent may also consider improving people's lives and compare the outcomes with an act that simply retains the status quo (e.g., health-related policy interventions). In these situations, the value of people's lives is not the only relevant quantity, but their status in the presence versus absence of a potential action needs to be compared. It is therefore desirable to generalize the model, and make it applicable to these other kinds of multiple-outcome situations as well.

Next, it is questionable whether participants in Cohen and Ahn's (2016) experiments actually provided *moral* judgments. After all, the test question in all experiments was "Would *you* save [Item A], causing [Item B] to be killed/destroyed?" (emphasis added). What people say they would do can be very different from what they think is morally right. For example, people might say that they would save their best friend rather than five strangers, while at the same time denying that this is the correct thing to do from a moral point of view (Kahane & Shackel, 2010, see also Tassy, Oullier, Mancini, & Wicker, 2013, Soter, Berg, Gelman, & Kross, 2021). Royzman and Hagan (2017) demonstrated that the "would you…" question used in many experiments may actually not track moral judgment, but a self-assessment of the likelihood that one would act in the described situation. Matters are further complicated by the fact that many dilemma studies, including those conducted by Cohen and Ahn (2016), frame the participant as the actor in a dilemma. However, people can give different judgments about a case when they are mere observers and thus morally evaluate someone else's action (Nadelhoffer & Feltz, 2008). Arguably, a large proportion of day-to-day moral judgments, and certainly the examples cited in the introduction, concern the actual behaviour of other people rather than hypothetical scenarios about oneself. Whether subjective utilities of outcomes predict judgments that (1) are actually about morality, and (2) concern the behaviour of other people thus remains an open question.

1.2. A Generalized Subjective-Utilitarian Model (GSUM)

To address these concerns, we propose and evaluate a *Generalized Subjective-Utilitarian Model* (GSUM). GSUM is based on the sampling of values, like the model proposed by Cohen and Ahn (2016). As described above, their model compares values of relevant entities in their alive or intact state, for example the value of five lives against the value of one life. An underlying assumption seems to be that when killed or destroyed, the value of entities reduces to zero (or another constant), and is therefore cancelled out when comparing the action alternatives. This may be a plausible simplification in life-and-death dilemmas, but it limits the range of applicability of the model. To generalize the model, all relevant actual, hypothetical, or counterfactual states of entities need to be explicitly represented. Imagine that an action improves the lives of five people, but also leads to the death of one person (henceforth: *improving* cases). Here, the gain of the first group needs to be traded off against the death of one person. In the case of a retrospective moral evaluation of an already executed action, the relevant comparison is between the actual state of affairs after the intervention, and the counterfactual state that would have obtained in the absence of the intervention. However, the same comparison can be made for a prospective evaluation of moral permissibility, in which case the predicted states in the presence and absence of an intervention are both hypothetical.

In a case in which two groups of people (or animals, plants) are affected by an action, our model therefore considers four subjective utilities[1]: (1) the state of Group 1 without intervention, (2) the state of Group 1 after intervention, (3) the state of Group 2 without intervention, and (4) the state of Group 2 after intervention. From these four values, the subjective utility of acting in this particular scenario (henceforth *scenario utility*) can be calculated. In the case of a classic moral dilemma, and assuming that subjective utilities of dead entities cancel out, the model reduces to the comparison of alive or intact values, as described by Cohen and Ahn (2016). But other cases require it to explicitly represent the values of entities in the contrasted states. Here is an example with a classic life-and-death dilemma case in which five lives are saved at the expense of one (SU = subjective utility):

$$\text{Scenario Utility } (\textit{saving}) = [\text{SU (5 normal)}–\text{SU (5 dead)}]+ [\text{SU (1 dead)}–\text{SU (1 normal)}]$$

And for an *improving* case with the same numbers:

$$\text{Scenario Utility } (\textit{improving}) = [\text{SU (5 improved)}–\text{SU (5 normal)}]+ [\text{SU (1 dead)}–\text{SU (1 normal)}]$$

If the action has more favourable outcomes than inaction, the scenario utility becomes positive in both cases.

GSUM takes as its input subjective utility assessments for items in different numbers and states. To make predictions for a particular scenario in which two items are traded off, four values are randomly sampled from the relevant pool of utility estimates (for example: one value for "five people in normal condition", one value for "five dead people", and so on), and the scenario utility is calculated. If a scenario utility is positive, a value of 1 is stored, otherwise it is represented as 0. To arrive at a robust prediction for each scenario, a large number of sampling iterations and scenario utility calculations are performed for each scenario (we are going to use 10,000 iterations). The proportion of positive scenario utilities among this large number of iterations is used as the predictor for a scenario's moral evaluation. The higher the proportion of positive scenario utilities, the higher are the predicted moral permissibility ratings for acting. Fig. 1 illustrates the procedure for a moral dilemma (*saving*) and for a case in which the action leads to an improvement of otherwise unchanged entities (*improving*).

GSUM thus embodies straightforward intuitions about the functional

[1] Any number of outcomes can be added to this equation.

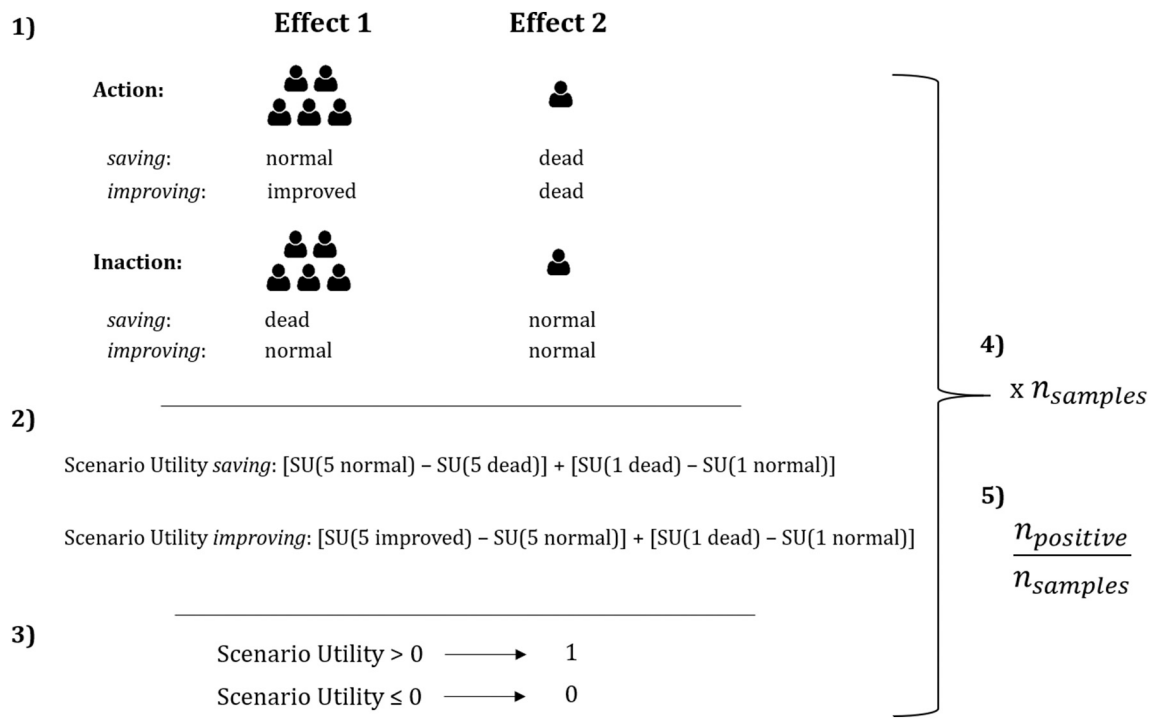

Fig. 1. Illustration of GSUM (with example values for subjective utilities) for a moral life-and-death dilemma (saving) and a case with two effects that is not a life-and-death dilemma (improving). SU = subjective utility.

form of a psychological outcome integration mechanism in the context of moral judgment. In a single sampling iteration, the model considers the aggregated value of all changes that are brought about by an action, and compares it to the aggregated value of an inaction. The crucial question for a moral evaluation of the action is whether the outcomes of acting outweigh the outcomes of inaction (or of an alternative action). GSUM represents this as a binary as well. As more and more samples are drawn, uncertainty caused by similar values of action and inaction or by large variations of the estimates becomes represented.

While our model is inspired by the model of Cohen and Ahn (2016), there are some key differences. The most obvious difference is the explicit modelling of state changes, resulting in a consideration of four rather than two values in each sample. Other differences arise due to the focus on moral instead of action preference judgments. Cohen and Ahn (2016) focus on binary choices. A choice counts as correctly predicted when the item with the higher mean utility (as identified in their independent utility estimation task) is saved. By contrast, we are interested in the extent to which people regard another person's action as morally permissible. We take moral permissibility to be a continuous evaluative reaction ranging from stark opposition to strong approval, rather than a binary choice. To predict moral judgments from participants' subjective utilities, we thus do not need to define a correct choice against which responses are compared. Our hypothesis is that the size of moral permissibility ratings will be proportional to the difference between the valuations of acting versus not acting. Formally, this is reflected in our model in the following way: we count the proportion of samples in which the outcomes of simulated actions outweigh the outcomes of simulated inactions. We use this proportion as a direct predictor of continuous moral permissibility judgments for actions.

2. Utility estimation study

In this study, we aimed to elicit the input data for our model, that is, subjective utility estimates for different entities in a range of numbers and states. The stimuli whose values we asked participants to estimate are the same ones that were used in the subsequent moral judgment tasks of Experiment 1 (life-and-death dilemmas) and Experiment 2 (life-and-death dilemmas vs. improving cases). Different kinds of entities (people, animals, plants) were compared in order to elicit a wide range of values, which allowed us to model a wide range of permissibility judgments in the subsequent moral judgment tasks.

2.1. Methods

2.1.1. Participants and design

We varied the number (one, five, ten, twenty, hundred), state (normal, dead, improved), and kind (people, monkeys, fish, trees, roses) of entities, all within-subject. We aimed for a sample size of 120 valid responses. Sample size was determined via simulation based on effects observed in a pilot study (small effect of numbers, $\eta_g^2 = 0.02$, large effects of state, $\eta_g^2 = 0.39$, and entity, $\eta_g^2 = 0.24$, two-way interactions between number and state, $\eta_g^2 = 0.01$, state and entity, $\eta_g^2 = 0.04$, and a three-way interaction, $\eta_g^2 = 0.004$). With 120 participants in a fully within-subject design with a conservative estimate for the correlation between repeated measures ($r = 0.1$), we achieve a power of at least 80% to detect each of these effects. Note, however, that the principal aim of this experiment was to collect input data for our model, not to test any specific hypotheses. All analyses should be therefore regarded as exploratory.

We invited 125 participants on *prolific* (www.prolific.co). Inclusion criteria were being at least 18 years old and a native English speaker, having an acceptance rate of previous studies on the platform of at least 90%, and not having participated in any previous studies using similar materials. Participants were paid £1.50 for an estimated 15 min of their time.

2.1.2. Materials and procedure

Participants were presented with the following instructions (see also Cohen & Ahn, 2016):

In the following study, your task will be to provide numerical value estimates for certain stimuli that will be presented to you. These stimuli can be people, animals, plants, or objects. You can understand the values that we will

ask you to estimate as an indication of how important, valuable or meaningful something /someone is, or how good or bad it is that something/ someone exists or does not exist, in your opinion. These values do not need to correspond to monetary value. For example, the first teddy bear you had as a child might have a high value to you, but only a very low monetary value. Likewise, something expensive could mean very little to you personally.

For example, an item in the experiment could be "a new bicycle". If you think that this is something good, then you should assign a positive value to this item. If you think that this is something bad, then you should assign a negative value. You could also assign a value of 0, to indicate that you are indifferent about the item. Moreover, the size of the value that you assign should reflect how positive or negative an item is, in your opinion. For example, assume you assigned a positive value of 10 to some item. If you value a second, different item ten times as much as this first item, you should assign a value of roughly 100 to the second item. The same is true for the negative direction. If you assign a negative value of − 10 to some item, and there is another item that is ten times worse than the first, in your opinion, you should assign a value of − 100 to the second item.

To help you come up with the numerical estimates, the task will be structured as follows:

You will see all items whose value we will ask you to estimate at once, on the same page. We encourage you to read through the whole list of items before assigning any values. When assigning the values, please use the following benchmarks as a reference:

- *Assume that "pieces of a broken tea cup" would be assigned a value of zero*
- *The highest possible value is + 1000*
- *The lowest possible value is − 1000*

Note that you can, but do not have to make use of the full range of the scale.

We chose "pieces of a broken tea cup" as a representative example for the scale value zero because we expected this item to be both familiar and naturally associated with a value of zero (worthless). Before the main task began, we presented participants with some practice trials ("a dead penguin", "two diamonds", "your best friend", "three healthy elephants", "a house that is burned down") and four instruction check questions (see Supplementary Materials). Participants were able to proceed to the main task once they had answered all instruction check questions correctly. Before entering any value estimates, participants had to scroll through the list of all 75 items (to help them calibrate their value estimates to the provided scale). On the next page, all items were presented again, and participants entered their value estimate for each item into a text field. The entries into text fields were not restricted, but participants were reminded to stick to the instructed scale (from −1000 to +1000).

2.2. Supplementary Materials

Data, materials, and code for this and all following experiments are available at https://osf.io/682uc/ (from here on: Supplementary Materials). For all statistical analyses and figures, we used R (R Core Team, 2019) and RStudio (RStudio Team, 2016) in combination with the following packages (in alphabetical order): *car* (Fox & Weisberg, 2019), *effsize* (Torchiano, 2020), *ez* (Lawrence, 2016), *faux* (DeBruine, 2020), *ggpubr* (Kassambara, 2019), *lmtest* (Zeileis & Hothorn, 2002), *MASS* (Venables & Ripley, 2002), *MBESS* (Kelley, 2019), *nlme* (Pinheiro et al., 2020), *nls2* (Grothendieck, 2013), *rcompanion* (Mangiafico, 2019), *reshape2* (Wickham, 2007), and the *tidyverse* (Wickham et al., 2019).

2.3. Results and discussion

Two participants were excluded because they failed a simple attention check,[2] resulting in a final sample size of 123 participants (mean age = 34.35, $SD = 13.16$, 56% women, 43% mean, 1% non-binary or no answer). Prior to the analyses we checked whether participants' entries conformed to the instructed response format (only numbers between −1000 and 1000, no text) and excluded those entries that did not. This resulted in the exclusion of 26 entries (0.3% of all entries). Fig. 2 shows the results. For all species, dead entities were predominantly assigned negative utilities (i.e., disutilities), and these values became more negative with higher numbers of dead entities. Normal and improved entities were assigned positive values that increased with larger numbers. Moreover, normal and improved states were valued very similarly overall. The highest values were assigned to people and the lowest to roses. Stepwise model comparisons revealed that the data were best described by a model containing main effects of number, entity, and state, the two-way interactions number × state and entity × state, plus the three-way interaction (see Table 1 for the output of the final model). The model explained 56% of the variance of the responses (Cragg & Uhler Pseudo-R^2). The number x state interaction reflects the fact that estimates became more positive with higher numbers for the improved and normal states, but more negative with higher numbers for the dead states (post-hoc tests[3] revealed that the effect was roughly medium-sized for all states, $\epsilon^2 = 0.07$ for dead states, 0.08 for normal states, and 0.1 for improved states, all $ps < .001$, just the direction changed; see Mangiafico, 2016, for benchmarks of ϵ^2). Likewise, the entity × state interaction reflects that when in a normal or improved state, the highest values were provided for people, then monkeys and trees, then fish, and then roses (all $p < .001$, with Bonferroni-adjustment). When entities were dead, however, this order was reversed, with the most negative values assigned to people, then monkeys and trees, then fish, then roses (all $p < .001$, with Bonferroni-adjustment). Again, the size of the effect was medium for all three states ($\epsilon^2 = 0.14$ for dead states, 0.10 for normal states, and 0.13 for improved states, all $p < .001$). The three-way interaction indicates that the difference in slopes for the manipulation of numbers of normal, improved, and dead states differed slightly between entities.

We also compared the fit of linear and nonlinear (exponential) models to the utility estimates, separately for each entity for alive (combining normal and improved) vs. dead states (see Supplementary Materials for the models and plots). We found that exponential models described the trajectory of utilities better than linear models for all entities and states. Utility estimates rise (or, for the dead states, fall) more quickly in the lower compared to the higher numerical ranges, thus showing patterns of diminishing marginal (dis-)utility or numbing (Slovic, 2007).

The main purpose of collecting this dataset was to use it as input for GSUM. We now turn to collecting moral permissibility judgments for a range of scenarios involving the stimuli whose values we have assessed in the Utility Estimation Study. We will also generate predictions for these cases using GSUM and compare them to participants' responses.

3. Experiment 1: Life-and-death dilemmas

The purpose of this experiment was to collect data from a new sample of participants for an initial evaluation of our model in an actual moral judgment task. We examined dilemmas in which ten entities

[2] "If Peter is taller than Alex, and Alex is taller than Max, who is the shortest among them?" This attention check was used in all subsequent experiments (presented on the final page).

[3] Friedman rank sum tests based on the data of all participants who provided no invalid entries ($N = 114$). *P*-values are Bonferroni-adjusted for the number of Friedman tests conducted (6 tests).

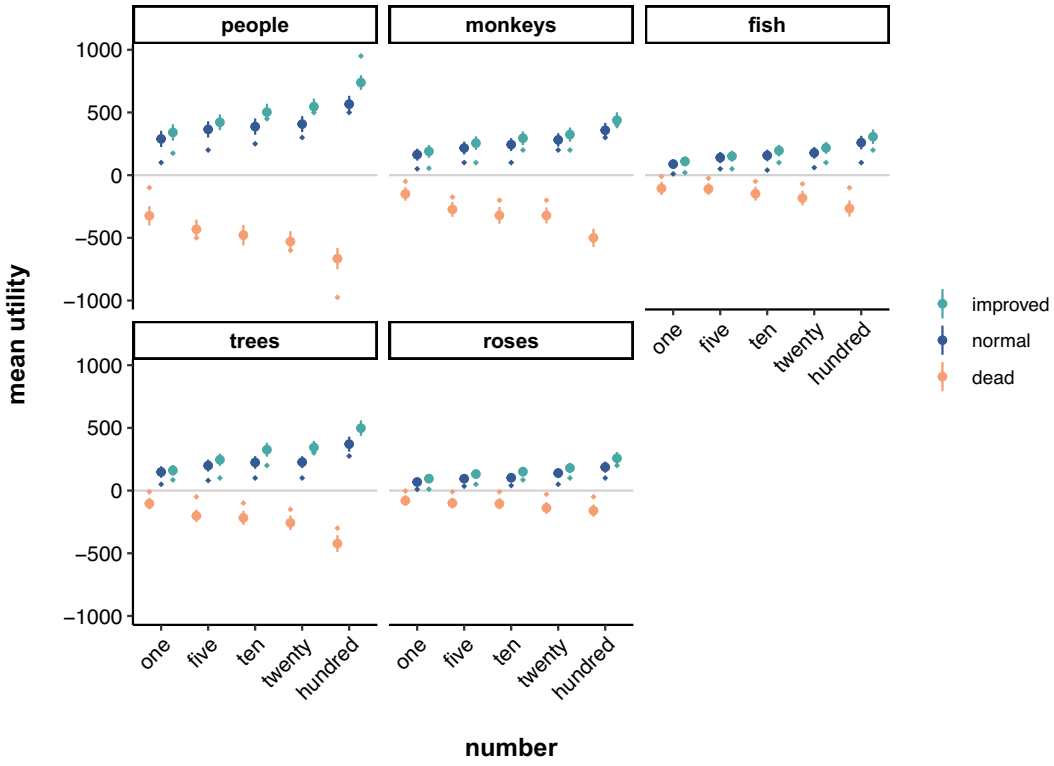

Fig. 2. Mean and median utilities assessed in the Utility Estimation Study. The large dots are means, the error bars are 95% confidence intervals. The small dots are medians. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(people, animals, or plants – two scenarios for each category) were killed in order to save either one hundred, twenty, or five others, or just one. Previous work (Cohen & Ahn, 2016) has only tested questions about personal action preferences, judged from the actor's perspective ("Would you…"). We systematically varied both the number and kind of entities (people, animals, plants) involved in a trade-off. This design allowed us to investigate whether the numerical ratio of lives saved versus lost influences moral judgments about trade-offs between human lives similarly as trade-offs between lives of animals and plants. The main goal was to explore whether potential value differences between humans, animals, and plants in different states explain differences in permissibility judgments.

3.1. Methods

3.1.1. Participants and design

We employed a 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people 1 (*foodtruck* case) vs. people 2 (*river* case)[4] vs. monkeys vs. fish vs. trees vs. roses, within subject) design. We expected that the between-subjects effect of the number of affected entities will be the smallest effect in the design. We invited 615 participants to participate in our survey on the platform *prolific* (www.prolific.co). To be included in the experiment, participants had to be native speakers of English, not have participated in any previous studies using similar materials, and have a 90% acceptance rate of previous tasks on the platform at least. Participants were paid 0.50 GBP for an estimated five minutes of their time (6 GBP/h). 21 participants were excluded for failing a simple attention check, leaving data of 594 participants for the analyses (mean age = 36.3, $SD = 12$, 60% female, 39% male, 1% another identity/no answer). This sample size yielded a power of approximately 80% to detect a between-subjects effect of numbers at Cohen's $f = 0.14$ ($\eta_p^2 = 0.019$), and a power of approximately 90% to detect the effect at Cohen's $f = 0.16$ ($\eta_p^2 = 0.025$; determined with G*Power 3.1.9.2, Faul, Erdfelder, Lang, & Buchner, 2007, and Superpower, Lakens & Caldwell, 2021).

3.1.2. Materials and procedure

In each of the six vignettes an agent is facing a dilemma. By performing a certain action, they can save a number of lives (hundred, twenty, five, or one), but will inevitably also cause ten deaths (this number was kept constant across all scenarios and conditions). The threat to one group was described as resulting from external circumstances such as natural disasters or illness. The sole means of saving was a re-allocation of limited resources (e.g., food, water), where receiving extra resources would save the threatened group. Given that these resources are limited, re-allocating more to the threatened group would lead to the death of the other, formerly unthreatened group (by lack of food or water, for example). Thus, harming was a side-effect of helping, never a means. We described agents as authorized to make the decision in question (via roles in government or management) in order to preclude participants from making judgments about legal rather than moral permissibility. Personal force or physical contact were not part of the scenarios. Moreover, the consequences of acting were never self-beneficial to agents. In each vignette, all entities are of the same kind (all human, all animals, or all plants). The agent is aware of all the outcomes and is motivated by the positive, but not the negative outcomes. In all cases, the agent decides to act, and both outcomes occur. Scenarios were presented in random order. After reading each scenario, participants were asked to provide a rating of the moral permissibility of

[4] Since we included two scenarios about animals (monkeys, fish) and plants (trees, roses), we also included two scenarios about people. These only differed in terms of the cover story: In the *food truck* case, lives could be saved by redirecting a food truck from one village to another; in the *river* case, people could be saved by redirecting a river.

Table 1
Summary of the selected regression model of the data collected in the Utility Estimation Study.

Random effects: participant ID					
		Intercept			Residual
SD		136.04			275.14
Fixed effects:					
	Estimate	SE	*df*	*t*	*p*
(Intercept)	−323.87	27.88	9002	−11.62	<0.001
Five	−108.02	35.37	9002	−3.05	0.002
Ten	−153.89	35.37	9002	−4.35	<0.001
Twenty	−205.94	35.37	9002	−5.82	<0.001
Hundred	−341.02	35.37	9002	−9.64	<0.001
Monkeys	173.93	35.3	9002	4.93	<0.001
Fish	218.58	35.37	9002	6.18	<0.001
Trees	220.17	35.37	9002	6.22	<0.001
Roses	244.44	35.37	9002	6.91	<0.001
Normal	612.88	35.3	9002	17.36	<0.001
Improved	664.36	35.3	9002	18.82	<0.001
Five, normal	183.81	49.92	9002	3.68	<0.001
Ten, normal	252.01	49.92	9002	5.05	<0.001
Twenty, normal	329.13	49.97	9002	6.59	<0.001
Hundred, normal	618.27	49.92	9002	12.38	<0.001
Five, improved	189.37	49.92	9002	3.79	<0.001
Ten, improved	315.77	49.92	9002	6.33	<0.001
Twenty, improved	411.92	49.92	9002	8.25	<0.001
Hundred, improved	739.07	49.92	9002	14.8	<0.001
Monkeys, normal	−299.41	49.87	9002	−6	<0.001
Fish, normal	−419.6	49.92	9002	−8.4	<0.001
Trees, normal	−361.01	49.97	9002	−7.22	<0.001
Roses, normal	−465.14	49.92	9002	−9.32	<0.001
Monkeys, improved	−324.86	49.87	9002	−6.51	<0.001
Fish, improved	−446.12	49.97	9002	−8.93	<0.001
Trees, improved	−399.61	49.92	9002	−8	<0.001
Roses, improved	−489.84	49.92	9002	−9.81	<0.001
Five monkeys	−14.89	49.97	9002	−0.3	0.766
Ten monkeys	−16.33	49.97	9002	−0.33	0.744
Twenty monkeys	34.25	49.92	9002	0.69	0.493
Hundred monkeys	−8.93	49.97	9002	−0.18	0.858
Five fish	102.98	49.97	9002	2.06	0.039
Ten fish	111.56	50.02	9002	2.23	0.026
Twenty fish	127.97	50.02	9002	2.56	0.011
Hundred fish	180.35	49.97	9002	3.61	<0.001
Five trees	10.59	50.02	9002	0.21	0.832
Ten trees	39.88	50.02	9002	0.8	0.425
Twenty trees	52.79	50.02	9002	1.06	0.291
Hundred trees	21.55	49.97	9002	0.43	0.666
Five roses	87.29	50.03	9002	1.74	0.081
Ten roses	128.86	50.02	9002	2.58	0.01
Twenty roses	147.69	50.08	9002	2.95	0.003
Hundred roses	260.43	49.97	9002	5.21	<0.001
Five monkeys, normal	−8.09	70.6	9002	−0.11	0.909
Ten monkeys, normal	−1.52	70.57	9002	−0.02	0.983
Twenty monkeys, normal	−39.77	70.57	9002	−0.56	0.573
Hundred monkeys, normal	−73.78	70.57	9002	−1.05	0.296
Five fish, normal	−126.69	70.57	9002	−1.8	0.073
Ten fish, normal	−140.86	70.6	9002	−2	0.046
Twenty fish, normal	−162.41	70.67	9002	−2.3	0.022
Hundred fish, normal	−286.04	70.57	9002	−4.05	<0.001
Five trees, normal	−35.31	70.64	9002	−0.5	0.617
Ten trees, normal	−62.72	70.67	9002	−0.89	0.375
Twenty trees, normal	−97.57	70.67	9002	−1.38	0.167
Hundred trees, normal	−76.55	70.6	9002	−1.08	0.278
Five roses, normal	−136.65	70.6	9002	−1.94	0.053
Ten roses, normal	−193.18	70.6	9002	−2.74	0.006
Twenty roses, normal	−199.63	70.68	9002	−2.82	0.005
Hundred roses, normal	−419.01	70.57	9002	−5.94	<0.001
Five monkeys, improved	−0.36	70.57	9002	−0.01	0.996
Ten monkeys, improved	−40.66	70.57	9002	−0.58	0.564
Twenty monkeys, improved	−105.98	70.53	9002	−1.5	0.133
Hundred monkeys, improved	−140.66	70.57	9002	−1.99	0.046
Five fish, improved	−146.22	70.6	9002	−2.07	0.038
Ten fish, improved	−190.62	70.64	9002	−2.7	0.007
Twenty fish, improved	−229.28	70.64	9002	−3.25	0.001
Hundred fish, improved	−384.56	70.6	9002	−5.45	<0.001
Five trees, improved	−7.08	70.6	9002	−0.1	0.92
Ten trees, improved	−36.99	70.6	9002	−0.52	0.6
Twenty trees, improved	−76.18	70.6	9002	−1.08	0.281
Hundred trees, improved	−82.76	70.57	9002	−1.17	0.241
Five roses, improved	−132.09	70.6	9002	−1.87	0.061
Ten roses, improved	−234.94	70.6	9002	−3.33	<0.001
Twenty roses, improved	−268.6	70.64	9002	−3.8	<0.001
Hundred roses, improved	−495.69	70.57	9002	−7.02	<0.001
AIC	129,970.4				
Pseudo-R^2 (Cragg & Uhler)	0.56				

the action ("To what extent was [agent]'s action morally permissible?") on a scale ranging from 1 ("not at all") to 10 ("fully"). For each scenario, illustrations were shown indicating the numbers of entities as well as their states before and after the agent's action. Here is an example of a scenario in which 100 people are saved and ten are killed (see Supplementary Materials for all other scenarios):

Olivia is the prime minister of Tolosia, a mountainous country with many distant and small valleys. She is authorised to make all decisions about the inhabitants' welfare.

One day, she learns that one valley, Morhall, is suffering from a drought that left its inhabitants in poor health due to lack of water. Exactly 100 people live in Morhall, all of whom are in critical condition and will die if nothing is done.

Olivia could order to open a dam that would redirect a mountain river towards Morhall. With a quick water supply, the 100 inhabitants would recover. However, the redirection of the river would also cause a lack of water in another mountain village, Lorness, causing its 10 inhabitants to die of thirst within a few days. All of the 10 inhabitants of Lorness are fine at the moment.

Since both valleys are inaccessible to any means of transport, redirecting the river is currently the only available measure to influence the wellbeing of the inhabitants.

Here is an illustration of the two valleys and the current state of their inhabitants (Fig. 3).

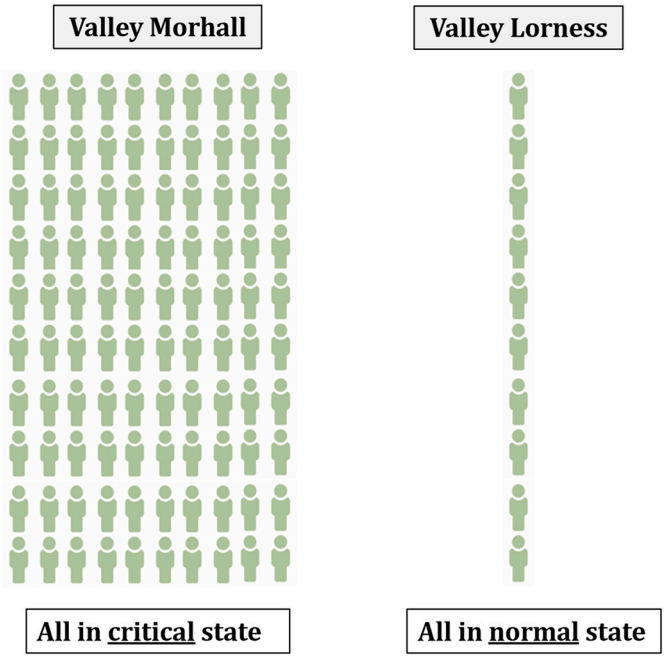

Fig. 3. Example of illustrations used in Experiment 1: States of affected groups before intervention.

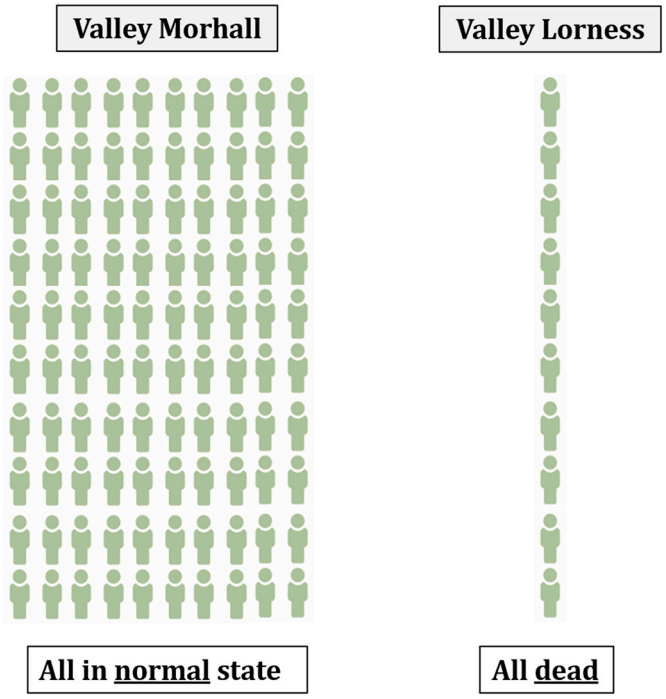

Fig. 4. Example of illustrations used in Experiment 1: States of affected groups after intervention.

Olivia is aware of all the facts. She wants the 100 inhabitants of Morhall to recover, but also not to cause any harm to the 10 inhabitants of Lorness. She decides to open the dam and redirect the mountain river. All of the 100 inhabitants of Morhall recover. However, all of the 10 inhabitants of Lorness die within a few days.

Here is an illustration of the two valleys and the state of their inhabitants after the river has been redirected (Fig. 4).

After completing all six scenarios, demographic variables were assessed, and participants were presented with the same attention check as in the previous study.

3.2. Results and discussion

Fig. 5 shows the mean moral permissibility ratings per condition, along with GSUM's predictions. The scenarios elicited judgments across the whole range of the rating scale. The action was judged as least permissible in the case of an unfavourable trade-off (saving one and killing 10) and when the affected entities were people ($M = 2.85$, $SD = 2.34$). It was judged as most permissible, nearly at ceiling, when the trade-off was favourable (saving 100 and killing 10) and the affected entities were plants ($M = 8.56$, $SD = 1.81$). In between, permissibility ratings increased as a function of the numerical ratio of saved compared to killed entities (more permissible with more entities saved compared to killed) and of the kind of affected entities (more permissible when plants were concerned than animals, and more permissible for animals than for people). This pattern indicates that people are more willing to trade off saving with harming when plants are involved than when the trade-offs concern animals. The strongest reluctance can be seen with humans.

A mixed 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people/foodtruck, people/river, monkeys, fish, trees, roses; within subject) ANOVA confirmed the impression from the visual inspection. There was a large main effect of the number of saved entities, $F(3, 590) = 137.62$, $p < .001$, $\eta_p^2 = 0.41$ [0.36; 0.45],[5] as well as a somewhat smaller, but still large effect of scenario, $F(5, 2950) = 186.76$, $p_{GG} < 0.001$, $\eta_p^2 = 0.24$ [0.22; 0.26].

There was also an interaction effect, $F(15, 2950) = 6.78$, $p_{GG} < 0.001$, $\eta_p^2 = 0.03$ [0.02; 0.04], indicating that the number of saved entities did not have an equally strong effect on moral permissibility ratings in all scenarios (the ANOVA results do not change when adjusting p-values for multiple testing). We followed up on this interaction with contrasts checking for an overall linear trend for the number variable, and possible interactions of this trend with the scenario factor. As expected, moral permissibility ratings showed an overall linear trend, increasing with more entities saved compared to harmed ($D = 2.64$, $t = 13.46$, $p < .001$). The significant interactions revealed that this linear trend was stronger when the involved entities were fish rather than people ($D = 0.81$, $t = 2.92$, $p = .003$), trees rather than people ($D = 1.10$, $t = 3.97$, $p < .001$), and roses rather than people ($D = 0.80$, $t = 2.89$, $p = .003$). The strength of the trend did not differ between the two scenarios involving people ($D = -0.09$, $t = -0.32$, $p = .75$.), nor between people and monkeys ($D = 0.35$, $t = 1.25$, $p = .21$).[6] Thus, the number of saved compared to killed entities mattered less for moral permissibility ratings in scenarios involving trade-offs among human lives compared to those of nearly all other entities. Detailed descriptive statistics for all conditions can be found in the Supplementary Materials.

To test GSUM, we generated permissibility predictions for all experimental conditions (see Supplementary Materials for the code). The model predicts participants' judgments well. We compared the fit of linear, exponential, and sigmoid functions to describe the relationship between model predictions and participants' mean moral evaluations of the scenarios. An exponential function ($y = ax^b$, $a = 12.16$, $t_{22} = 10.28$, $p < .001$, $b = 1.19$, $t_{22} = 7.64$, $p < .001$, normalized[7] RMSE = 0.16, Cragg & Uhler $R^2 = 0.77$) described the relationship best. Instead of group means, the model can also be fit to the group medians, which results in a virtually identical fit (here, a linear model described the relationship best, $b = 16.13$, $t_{22} = 8.29$, $p < .001$, normalized RMSE = 0.16, $R^2 = 0.76$).

We also generated a separate set of predictions in which values for the dead states of all entities were replaced by zeroes. This model corresponds to the one proposed by Cohen and Ahn (2016) in which only alive/intact states were compared. This model fits the data of the present study on life-and-death dilemmas roughly equally well, regardless of whether means or medians were used as criterion (means: $y = ax^b$, $a = 15.74$, $t_{22} = 8.19$, $p < .001$, $b = 1.45$, $t_{22} = 8.06$, $p < .001$, normalized RMSE = 0.15, Cragg & Uhler $R^2 = 0.79$; medians: $y = ax^b$, $a = 21.20$, $t_{22} = 6.23$, $p < .001$, $b = 1.94$, $t_{22} = 7.80$, $p < .001$, normalized RMSE = 0.15, Cragg & Uhler $R^2 = 0.79$). Again, we compared the fit of linear, exponential, and sigmoid functions, and reported the best-fitting relation, which was the exponential function). The next experiment will provide a better test between the models.

The results of Experiment 1 show that our generalized subjective utilitarian model (GSUM) predicts people's moral permissibility judgments of the actions of *other* agents. The better the outcomes of acting compared to inaction in a scenario, the higher participants' ratings of

[5] We report 90% confidence intervals for all eta squared effect sizes, see Steiger (2004).

[6] There was also a significant negative cubic trend ($D = -0.47$, $t = 2.43$, $p = .015$) for the manipulation of the numbers (overall, no interactions with scenario). This trend is likely due to the fact that ratings increased more steeply between five and twenty than between the other numerical conditions. The trend analyses were not adjusted for multiple testing and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests involved in the polynomial contrasts of the numbers variable (18 tests), only the following trends remain significant: the overall linear trend ($p < .001$) and the interaction with the trees scenario ($p = .001$).

[7] RMSEs were normalized by the range of the criterion on all occasions where they are reported.

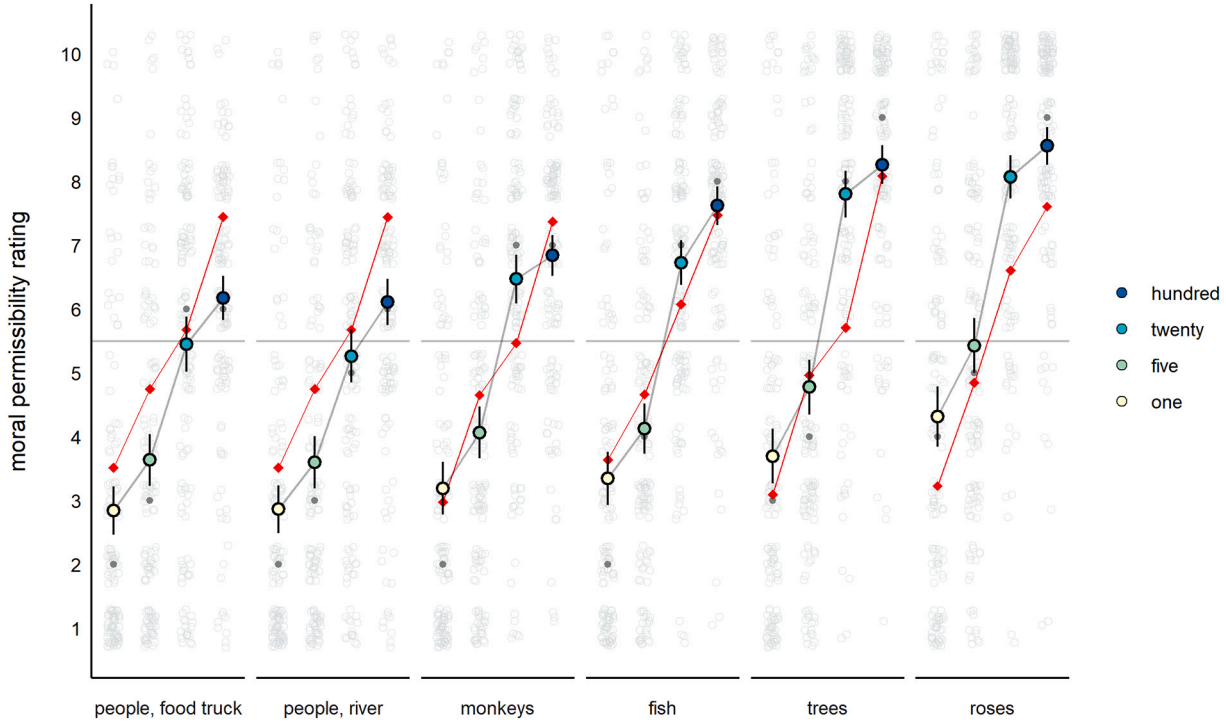

Fig. 5. Mean moral permissibility ratings (large points in blue colors) per condition in Experiment 1. Error bars are 95% confidence intervals. Medians are displayed in dark grey, individual data points (jittered) in light grey. GSUM predictions (fitted to means) are shown in red. The light grey line indicates the scale midpoint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

moral permissibility. Thus, it seems that subjective utilities do not only predict judgments about what people think *they* would *do* in a dilemma (Cohen & Ahn, 2016), but also of how they morally evaluate *other* people's behaviour.

It is noteworthy that participants' moral judgments did not show a strict split (i.e., uniformly low whenever fewer lives are saved than lost, uniformly high otherwise). Instead, moral permissibility ratings linearly increased with higher numbers of saved lives, even though the strength of the trend differed between entities. This pattern suggests that people's intuitions about the cases may be driven by the subjective values of the outcomes (relative to the outcomes of inactions) rather than, say, by a categorical principle.

4. Experiment 2: Saving versus improving

The aim of the second experiment was to extend the scope of investigated situations to cases beyond simple life-and-death dilemmas. Many actions with multiple morally relevant outcomes are not just about trade-offs between life and death. Other cases can be understood in terms of state differences, too. For example, an action might improve the lives of 100 people, but cause the deaths of ten others. The gain that is obtained by making the lives of 100 people somewhat better has to be traded off against the loss of 10 lives. If the perceived gain is higher than the perceived loss, the action should be seen as morally permissible. While decisions like this are more common than life-and-death dilemmas, previous models like the one proposed by Cohen and Ahn (2016) do not address them. By explicitly modelling the state changes that all entities undergo due to an action, GSUM can fill this gap. If moral judgments about improving scenarios are also driven by the subjective value of outcomes, an action should be seen as more morally permissible, the stronger its outcomes outweigh the outcomes of inaction (in the case of improving scenarios, retaining the status quo). An alternative possibility is that such actions are categorically impermissible, independent of the relation between losses and gains. Such a constraint might be justified deontologically, for example by positing that causing death can never be allowed when the positive outcome is a mere improvement of other's lives. In this case, participants permissibility judgments about such cases should be uniformly low.

4.1. Methods

4.1.1. Participants and design

The design was identical with the one of Experiment 1, except for the addition of a new between-subjects condition (improving). Here, the scenario was not described as a life-and-death dilemma; rather, the agent in the scenario had to decide whether to perform an action that would improve the states of some entities (people, animals, or plants, whose numbers varied as in Experiment 1) while causing the deaths of ten others. Thus, the full design was 2 (saving vs. improving, between-subjects) × 4 (number saved: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people 1 (*foodtruck* case) vs. people 2 (*river* case) vs. monkeys vs. fish vs. trees vs. roses, within subject). We decided to aim for a sample size of 300 participants in both the saving and the improving condition ($N = 600$ in total). We invited 621 participants to take part in our survey via *prolific* (www.prolific.co), who had not participated in Experiment 1. Otherwise, the inclusion criteria were the same as in Experiment 1. Participants were paid £0.50 for an estimated five minutes of their time (6 GBP/h). 14 participants were excluded for failing the attention check, leaving data of 607 participants for all analyses (mean age = 37.4, $SD = 13.3$, ca. 55% female, ca. 45% male, < 1% no answer). With 303 participants (rounded down) in both the *saving* and the *improving* conditions, we achieved a power of approximately 80% to detect a between-subjects effect of numbers at a size of Cohen's $f = 0.20$ ($\eta_p^2 = 0.038$), and a power of approximately 90% to detect this effect at a size of Cohen's $f = 0.22$ ($\eta_p^2 = 0.046$) in each condition (determined with GPower 3.1.9.2, Faul et al., 2007, and Superpower, Lakens & Caldwell, 2021). Note that these effects are the

smallest effects of interest in our design (the power is even higher for the within-factor "kind of affected entities" and for the main effect of "saving vs. improving" on moral permissibility ratings in an overall ANOVA).

4.1.2. Materials and procedure

In the saving conditions, we used the same vignettes as in Experiment 1. In the improving conditions, a different positive primary effect was described. As in Experiment 1 and as in the saving conditions, the action in the improving scenarios was a re-allocation of resources. This feature allowed us to keep all scenario features comparable to the saving conditions, with the exception that the agent did not re-allocate the resources to save a threatened group from death, but to improve a non-threatened group's condition while causing another group's death due to a lack of a resource. In the case of inaction, both groups of entities would remain in their normal, non-threatened state. For the example presented earlier (in which 100 people were saved), the improving version of the vignette included the following changes (see Supplementary Materials for the full text of all scenarios):

(..) One day she learns that the health of the 100 inhabitants of one valley, Morhall, could be even better and their lifespan vastly extended if extra water was available to them. Olivia could order to open a dam that would redirect a mountain river toward Morhall. With a quick water supply, the 100 inhabitants of Morhall could improve farming and hygiene and thereby reach an even better level of health and longer life than before. (…)

Olivia is aware of all the facts. She wants the 100 inhabitants of Morhall to improve their health and extend their lifespan, but also not to cause any harm to the 10 inhabitants of Lorness. She decides to open the dam and redirect the mountain river. All of the 100 inhabitants of Morhall improve their health and extend their lifespan. However, all of the 10 inhabitants of Lorness die within a few days.

As in the saving conditions, the improving versions of the vignettes included illustrations of numbers and states. Moral permissibility ratings and demographics were assessed in the same manner as in Experiment 1.

4.2. Results and discussion

Fig. 6 provides an overview of results, along with model predictions (see Supplementary Materials for all descriptive statistics). The results in the *saving* condition showed roughly the same patterns as in Experiment 1. In the *improving* conditions, the permissibility ratings were generally low. In most conditions, participants found an *improving* action not permissible (i.e., ratings below scale midpoint). However, within the lower half of the rating scale, permissibility ratings in the improving conditions still tended to increase when more entities' conditions were improved, as would be expected by GSUM. Trading off human lives was again least permissible, followed by animals, and plants.

The statistical analyses confirmed the descriptive patterns. In a

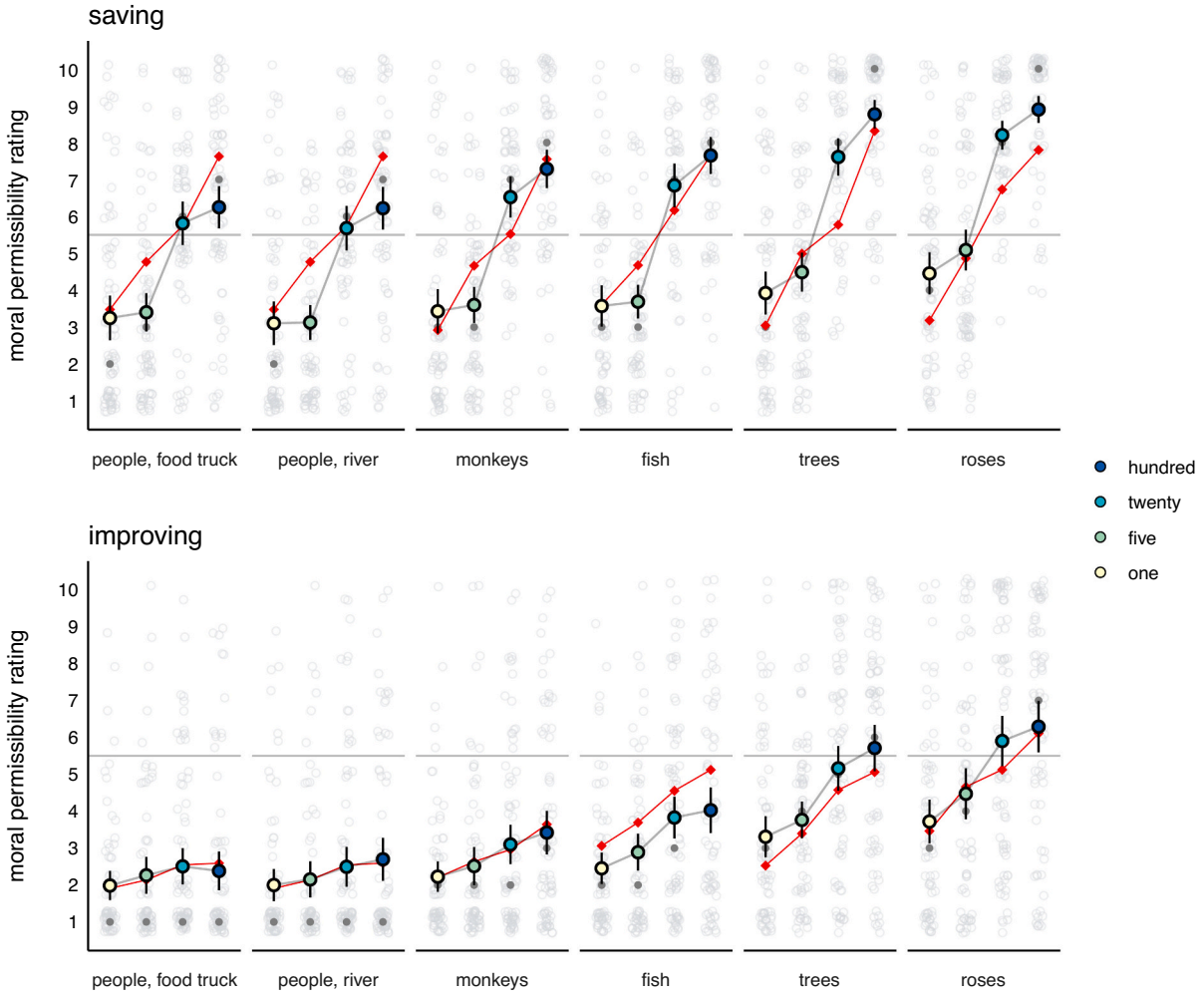

Fig. 6. Mean moral permissibility ratings (large points in blue colors) per condition in Experiment 2 (upper panel: saving conditions, lower panel: improving conditions). Error bars are 95% confidence intervals. Medians are displayed in dark grey, individual data points (jittered) in light grey. GSUM predictions (fitted to means) are shown in red. The light grey line indicates the scale midpoint. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Results of the overall ANOVA for Experiment 2. (*p*-values are Greenhouse-Geisser-corrected, degrees of freedom are unadjusted. Bonferroni-adjusting p-values for multiple testing did not change the results.)

Effect	*df*	*F*	*p*	η_p^2 [90% CI]
(Intercept)	1,599	3501.09	<0.001	
structure	1,599	192.34	<0.001	0.24 [0.20; 0.30]
number saved	3, 599	80.63	<0.001	0.29 [0.24; 0.33]
scenario	5,2995	224.49	<0.001	0.27 [0.25; 0.29]
structure:number saved	3, 599	19.33	<0.001	0.09 [0.05; 0.12]
structure:scenario	5,2995	6.84	<0.001	0.01 [0.01; 0.02]
number saved:scenario	15,2995	7.70	<0.001	0.04 [0.02; 0.04]
structure:number saved: scenario	15, 2995	0.41	0.937	<0.01

mixed 2 (structure: *saving* vs. *improving*, between-subjects) × 4 (number helped: hundred vs. twenty vs. five vs. one, between-subjects) × 6 (scenario: people/foodtruck vs. people/river vs. monkeys vs. fish vs. trees vs. roses, within subject) ANOVA, all main effects and two-way interactions were significant (all *p*'s < 0.001, see Table 2). To follow up on the differences between *saving* and *improving* cases (indicated by the main effect of structure and the two interactions involving structure), we conducted separate mixed ANOVAs for the two conditions. For the saving condition, we replicated the results from Experiment 1 with very similar effect sizes. People judged an action to be more permissible the more entities were saved compared to killed, $F(3, 299) = 86.79$, $p_{GG} < 0.001$, $\eta_p^2 = 0.47$ [0.40; 0.52], but again also differentiated between groups, with low permissibility ratings for the killing of people, higher permissibility ratings for animals, and the highest permissibility ratings for harming plants, $F(5,1495) = 90.28$, $p_{GG} < 0.001$, $\eta_p^2 = 0.23$ [0.20; 0.26]. Again, there was a small two-way interaction effect, $F(15,1495) = 3.61$, $p_{GG} < 0.001$, $\eta_p^2 = 0.03$ [0.01; 0.04], indicating that the number of saved compared to killed entities did not influence permissibility ratings equally for all groups (Bonferroni-adjusting *p*-values for multiple testing did not change the results). As in Experiment 1, contrasts revealed an overall positive linear trend in the moral permissibility ratings with increasing numbers of saved entities ($D = 2.55$, $t = 9.62$, $p < .001$), and this trend was stronger in the scenarios about fish ($D = 0.89$, $t = 2.37$, $p = .018$), trees ($D = 1.40$, $t = 3.72$, $p < .001$), and roses ($D = 1.12$, $t = 3.0$, $p = .003$) compared to people. The two people scenarios did not differ from each other ($D = 0.11$, $t = 0.30$, $p = .77$), and neither did the people and monkey scenarios ($D = 0.69$, $t = 1.84$, $p = .07$).[8]

The ANOVA for the *improving* condition confirmed that the number of affected entities also led to higher permissibility ratings in *improving* cases, although the effect was smaller than in the *saving* condition, $F(3,300) = 11.05$, $p_{GG} < 0.001$, partial $\eta^2 = 0.10$ [0.05; 0.15]. As in the saving condition, trade-offs among lives of people were seen as least permissible, followed by animals, and then plants, $F(5,1500) = 135.32$, $p_{GG} < 0.001$, $\eta_p^2 = 0.31$ [0.28; 0.34]. A small two-way interaction effect indicated that the influence of the number of improved entities did not affect moral judgments equally for all entities, $F(15,1500) = 4.39$, $p_{GG} < 0.001$, $\eta_p^2 = 0.04$ [0.02; 0.05] (these results did not change when Bonferroni-adjusting p-values for multiple testing). Follow-up contrasts showed that this time, there was no significant overall linear trend for the influence of numbers on permissibility ratings, but a linear trend emerged for the scenarios about fish ($D = 0.95$, $t = 2.41$, $p = .016$), trees ($D = 1.61$, $t = 4.08$, $p < .001$) and roses ($D = 1.72$, $t = 4.39$, $p < .001$), when compared to people.[9]

We generated predictions for the permissibility judgments in all experimental conditions using GSUM. Again, the model fit the data well. As in the previous study, we tested the fit of linear, exponential, and sigmoid functions. Exponential functions described the relationships best, and the fit was better for improving ($y = ax^b$, $a = 18.92$, $t_{22} = 7.35$, $p < .001$, $b = 1.12$, $t_{22} = 11.65$, $p < .001$, normalized RMSE = 0.11, Cragg & Uhler $R^2 = 0.90$) than for saving scenarios ($y = ax^b$, $a = 12.75$, $t_{22} = 9.62$, $p < .001$, $b = 1.25$, $t_{22} = 7.42$, $p < .001$, normalized RMSE = 0.17, Cragg & Uhler $R^2 = 0.76$). A similar fit is obtained overall when group medians were used as the criterion, with slightly better predictions of saving scenarios ($y = ax^b$, $a = 17.14$, $t_{22} = 8.62$, $p < .001$, $b = 1.72$, $t_{22} = 8.51$, $p < .001$, normalized RMSE = 0.13, Cragg & Uhler $R^2 = 0.82$), and slightly worse predictions of improving scenarios ($y = ax^b$, $a = 47.07$, $t_{22} = 3.82$, $p = .001$, $b = 1.90$, $t_{22} = 9.39$, $p < .001$, normalized RMSE = 0.12, Cragg & Uhler $R^2 = 0.85$).

The model predictions captured the patterns that we observed in the moral judgments. For improving scenarios, permissibility ratings and model predictions increased less steeply with higher numbers of entities benefitting from an action, compared to saving scenarios. Moreover, the model predictions reflected the differences between people, animals, and plants. The permissibility was generally lowest for people, and increased only very little with higher numbers of lives improved in this case. The predictions were higher for monkeys, trees, fish, roses (in this order), and also increased more steeply with numbers of lives improved for these groups.

To test GSUM against Cohen and Ahn's (2016) model, we again replaced all valuations of the dead states with zeroes, as their model solely took into account the valuations of the alive or intact states. The model was roughly equivalent to GSUM for the saving scenarios, regardless of whether means or medians were used as criterion (means: $y = ax^b$, $a = 16.63$, $t_{22} = 7.54$, $p < .001$, $b = 1.51$, $t_{22} = 7.70$, $p < .001$, normalized RMSE = 0.16, Cragg & Uhler $R^2 = 0.77$; medians: $y = ax^b$, $a = 24.62$, $t_{22} = 6.44$, $p < .001$, $b = 2.08$, $t_{22} = 8.56$, $p < .001$, normalized RMSE = 0.13, Cragg & Uhler $R^2 = 0.82$).[10] The model is not applicable for improving cases, as these cases require comparisons between more than just the two alive/intact states of entities. In the improving cases, three states are traded off against each other (dead, normal, improved), which is beyond the scope of the Cohen and Ahn model.

Interestingly, the predictions of GSUM were able to account for two patterns that might otherwise be attributed to deontological constraints. First, the model correctly predicted that in improving scenarios, acting was generally seen as impermissible. A possible account of this difference could have been that people regard causing death to merely improve other's lives as categorically impermissible, regardless of the extent of the benefit to one group. GSUM makes this prediction based on

[8] As in Experiment 1, there was also a significant negative cubic trend ($D = -0.95$, $t = 3.52$, $p < .001$) for the numbers factor (overall, no interactions with scenario). This trend is likely due to the fact that ratings increased more steeply between five and twenty than between the other numerical conditions. The trend analyses were not adjusted for multiple testing and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests (18 tests), only the following trends remain significant: the overall linear trend ($p < .001$), the overall cubic trend ($p = .008$), the interaction of the linear trend with the trees scenario ($p = .004$), and the interaction of the linear trend with the roses scenario ($p = .05$).

[9] No other trends for the numbers factor were significant. Trend analyses were not adjusted for multiple comparisons and should be regarded as exploratory. When Bonferroni-correcting for the number of trend tests (18 tests), only the following trends remain significant: the interaction with the trees scenario ($p < .001$), and the interaction with the roses scenario ($p < .001$).

[10] Only linear and exponential functions were compared for the relationship between the predictions by Cohen and Ahn's model and mean moral judgments, as sigmoid models did not converge here. Exponential functions described the relationship better for saving as well as improving scenarios and are therefore reported.

the fact that gains generally do not outweigh losses in improving cases, when the alternative state is normal. If a categorical constraint against acting in improving scenarios governed people's judgments, we should have observed equally low permissibility ratings in all numerical conditions and for all entities. We instead observed that permissibility ratings generally increased when larger numbers benefitted, suggesting that subjective utilities still influence permissibility judgments here. Second, this increase was weaker for higher-valued entities than for lower-valued entities, and not statistically detectable at all in scenarios about human lives. Again, this difference between species might be attributed to a deontological constraint shielding human lives from being traded off. Note however that our model predicts both the generally lower permissibility ratings for humans in improving scenarios, and the weak-to-absent increase of permissibility ratings with higher numbers in scenarios about human lives (see Fig. 6). GSUM makes these predictions based on the differences of the subjective utilities alone: Improving scenarios are generally fairly impermissible because here the losses (i.e., deaths) are not outweighed by the gains. However, they become gradually more permissible the larger the perceived gains are in relation to the perceived losses. Within the class of improving scenarios, acting is less permissible when people are concerned because losses are especially large for this group at all levels of the numerical manipulation, while at the same time the differences between normal and improved states (the gains) are more similar for all species groups (see Fig. 2).

5. Interindividual differences as a possible boundary condition

Based on the results we have described for the utility estimation data, the two experiments, and the fit between model predictions and data, we can derive additional hypotheses about subsets of participants for which better or worse correspondence between model predictions and data can be expected.[11] An inspection of the utility estimation data shows that participants differed in their use of the scales. This raises the question whether interindividual differences in the way the entities are valued may have generated noise that negatively affects the fit of our model. It is therefore interesting to test whether the predictions of GSUM change for different subsets of participants. We generated another set of predictions based on just the utilities of participants who assigned the minimal value of −1000 to any number of human deaths ($N = 66$). This corresponds to participants anchoring the scale at "dead people" = −1000, and determining the values of the other items from there. We also explored other anchors, such as "dead people = -1000 and improved people = 1000" (again for at least one of the numerical conditions). As for the relationship between the original GSUM predictions and the moral judgment data, we investigated the fit of several functions (linear, exponential, sigmoid), and we used both group means and medians as criterion.

The upshot of these analyses is that in four out of six cases, the best-fitting model based on the utilities of a homogeneous subset of participants fit the data better than the best-fitting model based on all participants' utilities (based on comparing normalized RMSE's, see Table 3). In two cases, the fit was identical, and there was only one case (improving, means as criterion) in which the predictions based on all participants' utilities fit the data slightly better. Thus, homogenizing the predictor variable improved model fit. Of course, these analyses should be regarded as exploratory, especially since the subgroups comprised just slightly more than between half and a third of participants in the utility estimation study.

A second focus of our analyses was on the minority of people who may have strict deontological constraints about intervening in a moral dilemma, even when more lives are saved than lost (cf. Thomson, 2008). GSUM's predictions will fail to describe the judgment of people who are insensitive to consequences in a moral dilemma. We used the data of Experiments 1 and 2 to estimate the upper bound of the proportion of such people. Typically, deontological constraints are applied to actions that harm or kill humans, not animals or plants. To use a lenient criterion, we thus determined the proportion of participants who thought that intervening was completely impermissible when human lives were at stake (rating = 1 on the scale ranging from 1 to 10), even though the ratio of lives saved compared to lost was favourable (conditions 100 vs. 10 and 20 vs. 10). 17% of participants in Experiment 1 and also 17% in Experiment 2 conformed to this criterion. Thus, 17% of the participants in our samples provided moral judgments that cannot be explained by GSUM. However, our experiments did not exhaust the space of possible outcome trade-offs. It may be the case that even though the threshold is higher for subjects classified as "deontologists"; they may ultimately waver in their judgment when outcome trade-offs in sacrificial dilemmas involve larger numbers of saved people than the "20 vs. 10" condition or even the "100 vs. 10" condition (i.e., disaster cases) (see Wiegmann & Waldmann, 2014, Experiment 5). Since we did not measure moral judgments in such disaster scenarios, we take the proportion of 17% to be an estimate of the upper bound of the true proportion of "deontologists" in our sample.

6. General discussion

It is generally undisputed that the foreseen outcomes of an action matter for its moral evaluation. Psychological theories of moral judgment acknowledge this, but how people reason about outcomes in morally charged situations has received little attention in the literature. Initially, one might be tempted to speculate that people do simple ordinal comparisons. When acting in a life-and-death dilemma saves more lives than not acting, the outcome trade-off may be registered as favourable, and it will factor into the action's global evaluation as a "pro" reason. However, such a simple notion of outcome comparisons quickly runs into problems, for example when different kinds of entities are compared, say, the life of one person against the lives of two fish, or against inanimate objects. A "common currency" is required. Subjective utility is a standard concept in decision theory, which has only recently been brought to bear on morally charged judgments and decisions (Cohen & Ahn, 2016).

In the present research we have shown that the contrast between subjective utilities of outcomes of an action, in comparison to inaction, predicts people's judgments of moral permissibility in different types of moral scenarios involving trade-offs between multiple outcomes. The contrasts also explain the different moral evaluation of dilemmas compared to cases in which one group's state is merely improved at another's expense. We observed a relatively high tendency to make trade-offs in life-and-death dilemmas, whereas the trade-off curves were flatter in improving situations. In these cases we discovered that subjects were more reluctant to trade-off a mere improvement against death when humans were involved compared to animals. For plants the willingness to make trade-offs was strongest.

While previous studies only assessed judgments about what participants would personally do, we demonstrated that our generalized subjective-utilitarian model (GSUM) can predict moral judgments about other people's actions in classic life-and-death dilemmas as well as for other multiple-outcome scenarios. In classical life-and-death dilemmas, GSUM's predictions converge with the predictions of earlier models (Cohen & Ahn, 2016). It apparently makes little difference whether the model considers negative valuations of dead states or assigns them a value of zero. However, as demonstrated in our improving scenarios, moral dilemmas do not only arise when life versus death is at stake, they may also require the considerations of different states of entities who remain alive. While the model of Cohen and Ahn (2016) is only applicable to situations that can be reduced to a comparison between the positive values of different entities (implicitly assuming that death or destruction can be represented by a constant, for example, zero), GSUM,

[11] We thank an anonymous reviewer for suggesting these additional analyses.

Table 3
Overview of fit measures for GSUM predictions based on the utility estimates of all participants ("full set", N = 123), and based on subgroups of participants who used the valuation scale more similarly to each other ("subsets", d1 = utilities of N = 66 participants who valued dead people at − 1000 in any numerical condition, d2 = utilities of N = 42 participants who valued dead people at − 1000 and improved people = 1000 in any numerical condition).

			GSUM full set			GSUM subsets			
Exp.	Condition	Criterion	Model	(Pseudo-)R^2	NRMSE	data		(Pseudo-)R^2	NRMSE
Exp1	Saving	Means	Exponential	0.77	0.16	d2	Linear	0.79	0.14
		Medians	Linear	0.76	0.16	d1	Sigmoid	0.82	0.14
Exp2	Saving	Means	Exponential	0.76	0.17	d1	Exponential	0.79	0.16
		Medians	Exponential	0.82	0.13	d2	Linear	0.83	0.13
	Improving	Means	Exponential	0.90	0.11	d1	Linear	0.79	0.13
		Medians	Exponential	0.85	0.12	d1	Sigmoid	0.86	0.12

due to its sensitivity to all relevant actual, hypothetical or counterfactual states of entities, can also analyse other dilemmas. In our improving scenarios, for example, these states were dead, normal, and improved, but other cases can be construed. Such situations are beyond the scope of Cohen and Ahn's (2016) model.

6.1. Can deontological response patterns be explained by differences in subjective utilities?

We have seen that for improving scenarios, GSUM correctly predicts lower permissibility ratings for trade-offs involving the lives of higher-valued entities, such as people or monkeys compared to trade-offs involving lower-valued entities, such as trees, fish, or roses. This pattern makes intuitive sense, and indeed it was apparent in participants' moral judgments. These evaluations can also be predicted by psychological variants of deontological ethics, which ascribe special rights to humans and not to other forms of live – with some deontological positions even claiming that human lives may not be traded off at all (see Alexander & Moore, 2016, for an overview of variants of deontological ethics). However, we have seen that these evaluations do not necessarily require positing gradually weakening deontological constraints on harming. GSUM can explain them without resorting to deontological ethics. Specifically, a person being dead is considered to be much worse than animals or plants being dead, while the difference between normal and improved states is more similar for all groups. This explains why scenarios in which the improvement of one group is traded off against the death of others received constantly low permissibility predictions by GSUM when people were concerned, compared to other types of entities. Thus, our findings show that psychologically the valuation of outcomes alone can account for some intuitions that otherwise might be interpreted as supporting deontological ethics.

6.2. The moral status of animals and plants

Treating animals and plants as moral entities is a quite recent development in Western philosophy. Kant (1974) made a sharp distinction between humans who have rights and must not be treated as means and animals who are largely outside the realm of morality (Korsgaard, 2018). In the meantime, both consequentialist (Singer, 1975) and nonconsequentialist (Korsgaard, 2018) philosophers have acknowledged the moral worth of animals. This development seems to have been partly triggered by an increasing awareness that animals are sentient beings who have emotions and can feel pain. In psychology, there has been increased interest in the psychological foundations of speciesism in the past years (Caviola et al., 2021; Caviola, Everett, & Faber, 2019; Crimston, Bain, Hornsey, & Bastian, 2016; Goodwin & Benforado, 2015; Horta, 2010).

One way to explain people's greater readiness to approve of trade-offs between animals compared to human lives could be that in both cases observers realize that acting leads to a gain compared to inaction, for example, because a larger number of lives are saved or because some lives are vastly improved. In the case of humans however, additional deontological considerations are activated, for example the intuition that humans have special rights not to be sacrificed in such a way. These deontological constraints then reduce people's willingness to morally approve of the action. Our results, in contrast, suggest that effects of speciesism may manifest much earlier in the assessment of values. Specifically, our results show that especially in improving scenarios, losses are not considered as equally outweighed by gains in trade-offs among members of different species, even when the objective numbers are constant. An interesting avenue for future research will be to investigate why people value the lives of different species so differently. Recent research asking subjects to assess the cognitive and suffering capacity of humans versus animals found that even when these features were matched, people still granted special consideration to human lives that were not extended to other species (Caviola et al., 2021).

Less is known about where the moral value of plants comes from. Our utility study shows that they are valued less than animals but still show intuitively plausible value differences. Although some people believe in the sentience of trees (e.g., Wohlleben, 2017), we believe that a more plausible source of the valuation of plants is their relation to human interests. Roses, for example, are aesthetically pleasing and it pains us if we see a bulldozer running over them. Moreover, there is an increasing awareness that our well-being is connected to nature and the climate, and we realize that destroying the rain forest, for example, has widespread consequences for our lives.

While species differences may to some extent be explained by differences of the associated subjective utilities, there are also established patterns in moral judgment that GSUM cannot capture in its current form. For example, when keeping outcomes constant, it is generally seen as morally worse to cause harm intentionally, by action rather than omission, as a means rather than a side effect, and by so-called personal force rather than more indirectly (for overviews see May, 2018; Waldmann et al., 2012). It has been suggested that these factors should be subsumed under the concept of "agential involvement" (May, 2018). The more involved an agent is in bringing about a harm, by any of the ways listed above and possibly others, the more severe our moral judgment tends to be.

6.3. Conclusion

In its current version, we regard GSUM as the formalization of one important component in a larger network of factors that jointly produce moral judgment. It constitutes an outcome formalism that can be implemented within different psychological theories of moral judgment. Even though we have not focused on these larger issues in this article, we think that the causal model framework may be best suited for this task. Causal models connect outcomes of different valences to the actions that produce them, and these actions can in turn be connected to mental states and character dispositions (Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Sloman, Fernbach, & Ewing, 2009; Waldmann, 2017; Waldmann, Wiegmann, & Nagel, 2017). Given that a central component of causal models are outcomes generated by actions, a mechanism computing trade-offs between outcomes is central. GSUM

could serve as the mechanism that compares the alternative outcomes when different causal paths are instantiated. In sum, we have demonstrated that the subjective utilities of outcomes predict genuinely moral judgments about multiple-outcome structures, not just personal preferences between possible courses of action. A central future goal will be to embed the trade-off component in a more complex theory that is sensitive to other relevant factors of moral judgments, such as intentionality and causality.

CRediT authorship contribution statement

Neele Engelmann: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Michael R. Waldmann:** Conceptualization, Writing – review & editing, Supervision.

Acknowledgements

We have no known conflicts of interests to disclose. Findings related to this project were presented at the 2019 Annual Meeting of the Cognitive Science Society in Montreal, Canada (Engelmann & Waldmann, 2019).

We would like to thank Alex Wiegmann for helpful comments. We also thank Michał Białek and three anonymous reviewers.

References

Alexander, L., & Moore, M. (2016). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554.

Białek, M., Turpin, M. H., & Fugelsang, J. A. (2019). What is the right question for moral psychology to answer? Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science, 30*(9), 1383–1385.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084–1093.

Caviola, L., Everett, J. A., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology, 116*(6), 1011.

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

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

Crimston, D., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology, 111*(4), 636.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Cromley, A. R., & Cohen, D. (2019). *Subjective values theory: The psychophysics of psychological value*. https://doi.org/10.31234/osf.io/wfd5s

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*(6), 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

DeBruine, L. (2020). *faux: simulation for factorial designs*. Retrieved from. https://doi.org/10.5281/zenodo.2669586.

Engelmann, N., & Waldmann, M. R. (2019). Moral reasoning with multiple effects: Justification and moral responsibility for side effects. In *Proceedings of the 41st meeting of the cognitive science society* (pp. 1703–1709). Austin, TX: Cognitive Science Society.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Feltz, A., & May, J. (2017). The means/side-effect distinction in moral cognition: A meta-analysis. *Cognition, 166*, 314–327.

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

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks: Sage.

Goodwin, G. P., & Benforado, A. (2015). Judging the goring ox: Retribution directed toward animals. *Cognitive Science, 39*(3), 619–646.

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin.

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.

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.

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.

Grothendieck, G. (2013). nls2: Non-linear regression with brute force. Retrieved from https://CRAN.R-project.org/package=nls2.

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

Horta, O. (2010). What is speciesism? *Journal of Agricultural and Environmental Ethics, 23*(3), 243–266.

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgement. *Mind & Language, 25*(5), 561–582.

Kant, I. (1974). *Anthropology from a pragmatic point of view. Translated by Mary Gregor*. The Hague: Martinus Nijhoff.

Kassambara, A. (2019). ggpubr: ggplot2 based publication ready plots. Retrieved from https://CRAN.R-project.org/package=ggpubr.

Kelley, K. (2019). MBESS: the MBESS R package. Retrieved from https://CRAN.R-project.org/package=MBESS.

Korsgaard, C. M. (2018). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4*(1).

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology, 129*, 101412.

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments. Retrieved from https://CRAN.R-project.org/package=ez.

Mangiafico, S. (2016). Summary and analysis of extension program evaluation in R (version 1.18.8). Retrieved from https://rcompanion.org/handbook/F_08.html.

Mangiafico, S. (2019). rcompanion: Functions to support extension education program evaluation. Retrieved from https://CRAN.R-project.org/package=rcompanion.

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

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

Nadelhoffer, T., & Feltz, A. (2008). The actor–observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics, 1*(2), 133–144.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). nlme: Linear and nonlinear mixed effect models. Retrieved from https://CRAN.R-project.org/package=nlme.

Plunkett, D., & Greene, J. D. (2019). Overlooked evidence and a misunderstanding of what trolley dilemmas do best: Commentary on Bostyn, Sevenhant, and Roets (2018). *Psychological Science, 30*(9), 1389–1391.

R Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356.

Royzman, E., & Hagan, J. P. (2017). The shadow and the tree: Inference and transformation of cognitive content in psychology of moral judgment. In J.-F. Bonnefon, & B. Trémolière (Eds.), *Moral inferences* (pp. 64–82). London: Routledge.

RStudio Team. (2016). RStudio: Integrated development environment for R. Boston, MA. Retrieved from http://www.rstudio.com/.

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science, 15*(2), 207–215.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.

Singer, P. (1975). *Animal liberation: A new ethic for our treatment of animals*. New York: HarperCollins.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. *Psychology of Learning and Motivation, 50*, 1–26.

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision making, 2*(2), 1–17.

Soter, L. K., Berg, M. K., Gelman, S. A., & Kross, E. (2021). What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions. *Cognition, 217*, 104886.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.

Steele, K., & Stefánsson, H. O. (2020). Decision theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164.
Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 250.
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415.
Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs, 36*(4), 359–374.
Torchiano, M. (2020). effsize: Efficient effect size computation. Retrieved from https://CRAN.R-project.org/package=effsize.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.
Waldmann, M. R. (Ed.). (2017). *The Oxford handbook of causal reasoning*. New York: Oxford University Press.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.
Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon, & B. Trémolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge.
Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/.
Wickham, et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686.
Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition, 131*(1), 28–43.
Willemsen, P., & Reuter, K. (2016). Is there really an omission effect? *Philosophical Psychology, 29*(8), 1142–1159.
Wohlleben, P. (2017). *The hidden life of trees: What they feel, how they communicate*. Glasgow: William Collins.
Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214.
Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News, 2*(3), 7–10. Retrieved from https://CRAN.R-project.org/doc/Rnews/.