# Ethical choice reversals

Hao, Chenxu; Lewis, Richard L.

**DOI**
10.1016/j.cogpsych.2024.101672

**Publication date**
2024

**Document Version**
Final published version

**Published in**
Cognitive Psychology

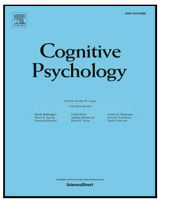
**Citation (APA)**
Hao, C., & Lewis, R. L. (2024). Ethical choice reversals. *Cognitive Psychology*, *153*, Article 101672. https://doi.org/10.1016/j.cogpsych.2024.101672

Cognitive Psychology 153 (2024) 101672

Contents lists available at ScienceDirect

Cognitive Psychology

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



# Ethical choice reversals

Chenxu Hao [a,b,*], Richard L. Lewis [b,c]

[a] Pattern Recognition & Bioinformatics, Department of Intelligent Systems, Delft University of Technology, The Netherlands
[b] Department of Psychology, University of Michigan, United States of America
[c] Weinberg Institute for Cognitive Science, University of Michigan, United States of America

ARTICLE INFO

*Keywords:*
Ethical decision
Preference reversals
Rationality

ABSTRACT

Understanding the systematic ways that human decision making departs from normative principles has been important in the development of cognitive theory across multiple decision domains. We focus here on whether such seemingly "irrational" decisions occur in *ethical* decisions that impose difficult tradeoffs between the welfare and interests of different individuals or groups. Across three sets of experiments and in multiple decision scenarios, we provide clear evidence that *contextual choice reversals* arise in multiples types of ethical choice settings, in just the way that they do in other domains ranging from economic gambles to perceptual judgments (Trueblood et al., 2013; Wedell, 1991). Specifically, we find within-participant evidence for *attraction effects* in which choices between two options systematically vary as a function of features of a third dominated and unchosen option—a *prima facie* violation of rational choice axioms that demand consistency. Unlike economic gambles and most domains in which such effects have been studied, many of our ethical scenarios involve features that are not presented numerically, and features for which there is no clear majority-endorsed ranking. We provide empirical evidence and a novel modeling analysis based on individual differences of feature rankings within attributes to show that such individual variations partly explains observed variation in the attraction effects. We conclude by discussing how recent computational analyses of attraction effects may provide a basis for understanding how the observed patterns of choices reflect boundedly rational decision processes.

## 1. Introduction

Understanding the systematic ways that human decision making departs from normative principles has been important in the development of cognitive theory across multiple decision domains (Chang & Cikara, 2018; Horvath & Wiegmann, 2022; Huber, Payne, & Puto, 1982; O'Curry & Pitts, 1995; Trueblood, Brown, Heathcote, & Busemeyer, 2013; Wedell, 1991). Our focus in this work is on multi-attribute *ethical* or *moral* decisions — decisions concerned with the welfare of others (Yu, Siegel, & Crockett, 2019). How individuals make ethical decisions not only affect the welfare of themselves and the others, but also have potential impact on law, politics, and other areas of public life (Sunstein, 2002). In addition, understanding how humans make ethical decisions is important to the development of ethical artificial intelligence (Borg, Sinnott-Armstrong, & Conitzer, 2024). The primary contribution of our work is to provide clear empirical evidence that systematic *contextual choice reversals*, or attraction effects, arise in ethical decisions across multiple scenarios. These scenarios all require the decision maker to make a tradeoff between ethically-involved attributes, some of which quantitative, and some qualitative. Such attraction effects, arising in many choice domains and in many decision-making organisms, are among the most striking apparent violations of axioms of rational choice theory, which demand consistency in

* Corresponding author at: Pattern Recognition & Bioinformatics, Department of Intelligent Systems, Delft University of Technology, The Netherlands.
*E-mail address:* c.hao-1@tudelft.nl (C. Hao).

https://doi.org/10.1016/j.cogpsych.2024.101672
Received 24 August 2022; Received in revised form 17 June 2024; Accepted 22 July 2024
Available online 7 August 2024

choice preference. A secondary contribution of our work is a computational method for modeling individual differences in attribute rankings and deriving their implications for attraction effects. We show with this method that individually varying rankings partly explain observed variation in choice reversals across individuals and decision scenarios.

Ethical or moral decision making has been studied in a variety of ways. These include incentivized choices that involve harm or reward (e.g. decisions involving a trade-off between some monetary reward and the number of painful electric shocks directed to either self or another agent; Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Rand, Greene, & Nowak, 2012; van Baar Jeroen, Chang, & Sanfey, 2019), and judgments about the moral appropriateness of actions involving the welfare of others (e.g., using hypothetical dilemmas such as the Trolley Problem; Awad, Dsouza, Shariff, Rahwan, & Bonnefon, 2020; Barak-Corren, Tsay, Cushman, & Bazerman, 2018; Foot, 1967; Kim et al., 2018; Merlhiot, Mermillod, Jean-Luc, Dutheil, & Mondillon, 2018; Thomson, 1976).

A common feature of these paradigms and much other work on ethical decisions is that the choices involve a *tradeoff* between the welfare or interests of different individuals or groups (often, but not always, including the decision maker). When these tradeoffs are particularly difficult, we refer to the choice problems as *dilemmas*.

Through careful experimental manipulation of features of these choice problems, behavioral scientists and moral philosophers have discovered many ways that ethical decisions depart from normative ethical theories or other normative decision and behavioral principles. A prominent example is *moral luck* (Nagel, 2012; Williams & Bernard, 1981), where judgments of moral blame or praise are based on consequences of actions where the consequences are out of the control of the actor (such judgments are thought to be inconsistent with a normative principle that morality should not be affected by luck; Kant, 1998; Williams & Bernard, 1981). Another example is considering only a subset of the ethically-relevant attributes of a decision, or basing judgments or choices on non-relevant attributes (Nadurak, 2018; Sinnott-Armstrong, Young, & Cushman, 2010).

One explanation for these and other departures from normative theory is that people use *moral heuristics* or "mental shortcuts" (Sunstein, 2005) to make decisions (see Table 14 for a summary). These heuristics may be understood as adaptive in that they strike a balance between cognitive effort and decision quality (Gigerenzer, 2010), but the practical concern is that they may also lead to undesirable consequences in law, politics, and other areas of public and private life (Nadurak, 2020; Sunstein, 2002).

Our concern in this work is whether ethical decisions also exhibit violations of one of the most fundamental principles of axiomatic rational choice theory— *independence of irrelevant alternatives*: the choice among options A and B should not depend on properties of a third irrelevant option D (e.g., Luce, 2012), a requirement for a kind of consistency. Past work has already shown that preferences for ethically-involved choices such as choosing among disease control programs that require a trade-off between number of lives and probability of saving are affected by the change of framing (Tversky & Kahneman, 1981). Expressed ethical preferences in classical ethical decision problems such as the trolley problem and medical dilemmas may also become inconsistent when the method of evaluation changes, e.g., separate evaluation, joint evaluation, or forced choice (Barak-Corren et al., 2018; Erlandsson, 2021; Erlandsson et al., 2020).

Specifically, our empirical question is whether *contextual preference reversals* arise in ethical choices, where a choice between two options or courses of action systematically varies as a function of properties of an unchosen option in the choice set. Such contextual preference reversals have been shown to arise in multiple choice domains, e.g. economic gambles, consumer goods, political candidates, perceptual size judgments (Huber et al., 1982; O'Curry & Pitts, 1995; Trueblood et al., 2013; Wedell, 1991) and in multiple decision making organisms, including humans, monkeys and slime molds (Huber et al., 1982; Latty & Beekman, 2011; Parrish, Evans, & Beran, 2015). We will refer to these phenomena as contextual *choice* reversals rather than preference reversals, using a theoretically more neutral term that refers to the behavioral data rather than an internal cognitive construct (preferences).

Some past work suggests that choice reversals can arise in classical ethical decisions, such as scenarios based on the trolley problem, regardless of the decision maker's expertise in ethics and philosophy (Horvath & Wiegmann, 2022). Reversals can also be observed in the domain of social choices, whose consequences affect the welfare of others — such as choosing political candidates (O'Curry & Pitts, 1995), hypothetical policy-making (Herne, 1997), making hiring decisions (Highhouse, 1996), and choosing job candidates while making a tradeoff between each candidate's stereotypical traits such as warmth and competence (Chang & Cikara, 2018; Chang, Gershman, & Cikara, 2019).

The work we present here differs from this past work in three important respects. First, we use a within-participant design that allows us to observe choice reversals at the *individual* level. Second, we always present options in choice sets of triplets, rather than comparing across two-option and three-option sets. This design — following Wedell (1991) and many others — permits direct tests of the effects of changing properties of the third irrelevant option (or *decoy*, as described below), where the number of options is not a confounding variable (Katsimpokis, Fontanesi, & Rieskamp, 2022).

A third important distinctive feature of our experiments is that our scenarios go beyond the classical scenarios seen in previous work (e.g., choosing rescue programs while making a tradeoff between saving the number of people and probability of saving). Our new scenarios pit interests of different groups and individuals against each other (or the decision maker themselves) and often include attributes that are not expressed numerically. Past investigations of the attraction effect in the domain of commercial decisions suggest that choice reversals often cannot be observed when the options are presented in different modalities (e.g., pictures, taste of beverages) or simply without numeric information (Frederick, Lee, & Baskin, 2014). Although our decision problems are still descriptive, by including attributes that are not numeric information, we aim to make our problems closer to real-life decisions and thus increase the ecologically validity of our experiments.

In the remainder of the paper we first review the formal structure of decision problems that give rise to choice reversals, and summarize how we create ethical dilemmas with this formal structure. Instead of using trolley problems, we aim to ground our dilemmas in more realistic scenarios. We then provide the details of our main experiments. Experiment 1 constructs ethical dilemmas

that are isomorphic to a seminal study demonstrating choice reversals in economic gambles (Wedell, 1991), using variants of a single scenario (a choice among rescue plans after a natural disaster). This study yielded ethical choice reversals with a pattern nearly identical to the original economic experiment. Experiments 2 and 3 use a set of multiple distinct scenarios, including some that involve attributes that are not presented numerically or without clear objective rankings. These scenarios also yield choice reversals, but to varying degrees. We show, with a simple computational model, that variation among these scenarios and among individual decision makers can be explained in part by variation in individuals' subjective rankings of the choice attributes.

We conclude with a summary and a discussion of limitations of the studies. We also reconsider the question of rationality in light of new computational and mathematical models of multi-attribute decision-making that predict contextual choice reversals as a consequence of boundedly rational utility maximization.

## 2. Contextual choice reversals and the attraction effect

Consider the following decision problem. You are choosing among three video games to purchase, and are weighing price against quality, here assessed as user experience of the game. The choices are a console-version video game that is very expensive and provides you with a great experience, a PC-version of the same video game that is cheaper but provides a lower-level experience, and a smartphone version that is the same price as the PC-version, but provides an even lower-level experience. Suppose that you opt here for the PC-version. But now consider a choice set with the same console-version and PC-version, but with a tablet version that provides the same great user experience as the console version but is more expensive. And suppose that faced with these options, you choose the console version.

You have exhibited a *contextual choice reversal*: you have switched your expressed preference between the console and PC-version of the game, dependent upon features of a third unchosen option. No matter how you tradeoff or weigh price and user experience, your choice behavior is inconsistent. Under a standard account, you either do not have stable preferences, or if you do, you did not choose rationally.

The structure of this example is a classic three-option two-attribute choice problem in which two options, termed *target* and *competitor*, are of roughly equal expected value but differ on both attributes: the target is superior to the competitor on one attribute but inferior on the other. The third option is a *decoy* and is *dominated* by the target, by being either inferior on both attributes, or equal on one and inferior on the other . By changing the position of the decoy in attribute space, it is possible to change which of the other two options is the dominating option (and therefore the target). The empirical finding that moving the decoy in this way systematically changes preferences is known as the *attraction* effect, because the decoy positioning "attracts" additional choices.

Past work suggests that the decoy option can be placed at various positions for the attraction effect to arise (Huber et al., 1982; Wedell, 1991). In order to test specific hypotheses, Huber et al. (1982) and Wedell (1991) position the decoys as the Range decoy (R), the Frequency decoy (F), the Range-Frequency decoy (RF), and the Rprime decoy (R′). R decoy increases the range of the dimension where the target is the weakest (R′ decoy strongly increases that range); F decoy increases the frequency of the dimension where the target is the strongest; RF decoy combines the features of both R and F decoys. We maintain the R, F and RF nomenclature for consistency with past work, but for the purpose of this work we are not interested in the distinction between R and F decoys and treat them together as a class of decoys that are dominated by the target option on a single dimension.

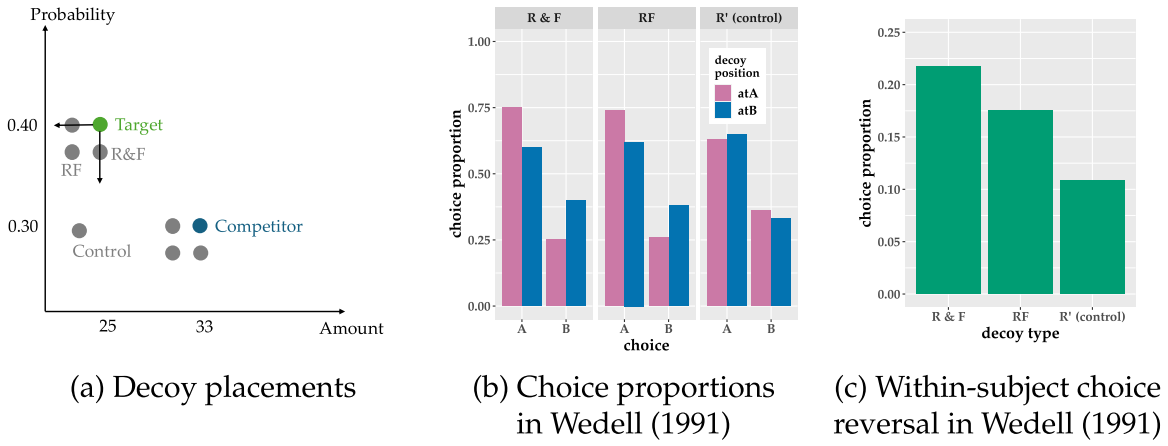
Fig. 1 (a, left) shows possible placements of decoy options using choice among economic gambles as an example (Wedell, 1991), where each option or gamble is a probability and value pair $\langle p, v\rangle$. Selecting option $\langle p, v\rangle$ means playing a gamble which pays out $v$ with probability $p$ and 0 with probability $1 - p$. The expected value of each option is thus $pv$. Fig. 1 (b, middle) also shows the choice proportions reported by Wedell (1991) for all variants of the decoy placement. All problems involve the same set of A and B options; there is an overall preference for the A options (corresponding in the Wedell (1991) stimuli to a risk-related preference for higher probability gambles). The decoy placement systematically yields the attraction effect for both A and B options, in both R & F and RF decoy placements. Fig. 1 (c, right) shows the within-subject choice reversal rates for pairs of decision problems reported by Wedell (1991) for R & F and RF decoy placements. A pair of decision problems consist of identical A, B options, with the decoy separately dominated by A and B. If a participant always selects the target option in a pair, then there is a within-subject choice reversal. In most trials, participants choose consistently (Fig. 17, Appendix B), and within-subject choice reversal rates are around 15% to 20%.

To investigate contextual choice reversals in ethical decisions, we use three experiments that use ethical dilemmas that have the same formal structure as that shown in Fig. 1. The challenge in designing such dilemmas is finding scenarios in which there is a trade-off between two attributes that impose the ethical dilemma, each of which admits three clearly distinct levels. Our first experiment addresses this challenge by creating ethical scenario isomorphs of the Wedell (1991) stimuli, using the same probabilities and values as those stimuli.

*Materials & data availability.*

*Experiment 1a & 1b.* All survey materials, data, and R analysis scripts are available from Open Science Framework: https://osf.io/9eqga/

*Experiment 2 & 3.* All survey materials, data, and R analysis scripts are available from Open Science Framework: https://osf.io/w8nrm/?view_only=7a1d4608190742c6a188ce036e43d29d. Experiment 2 is pre-registered at https://osf.io/4n9f7 and Experiment 3 is pre-registered at https://osf.io/7fdw8.




(a) Decoy placements (b) Choice proportions in Wedell (1991) (c) Within-subject choice reversal in Wedell (1991)

**Fig. 1.** (a) Decoy placements illustrated with an example from Wedell (1991). When choosing among sets of gambles, the decision maker makes a trade off between two attributes: probability and the amount of money to win. Two options with equal expected values vary in both attributes but have the same expected value. As shown above, both options have R & F and RF decoys. When the third option, the decoy, is dominated by A, then A is the target and B is the competitor, and vice versa. (b) Choice proportions showing the attraction effect in both R & F and RF dominance problems. The effect may be seen in the higher proportions of choices for the A options when the decoy is at A compared to when it is at B, and higher proportion of choices for the B options when the decoy is at B compared to when it is at A. There is an overall preference for option A. (c) Within-subject choice reversal rates in both R & F and RF dominance problems reported in Wedell (1991).

## 3. Experiment 1

In Experiment 1a, we present a replication study of Wedell (1991), where participants made choices among sets of economic gambles. In Experiment 1b, we present a ethical decision study where participants made choices among disaster rescue plans isomorphic to the economic gambles in Wedell (1991).

### 3.1. Experiment 1a: Wedell (1991) replication

#### 3.1.1. Method

*Participants.* One hundred and fifty-five participants were recruited from undergraduate psychology subject pool at the University of Michigan. Five participants were excluded due to survey incompleteness. In total, 150 participants (104 female; age M(SD) = 19(0.76) years) were included in the data analysis.

*Materials.* We constructed a questionnaire that contained 40 pairs of questions (80 in total) with the original stimuli from Wedell (1991; Table 15, Appendix B). There were 10 pairs of questions for each type of decoy. Each pair contains two questions with the same A and B targets but different decoys: one dominated by A, and another dominated by B. Each participant completed a survey that contained 10 random pairs drawn from the 40 pairs of questions and all questions were displayed in a random order.

Here is one example of the question presented to the participants:

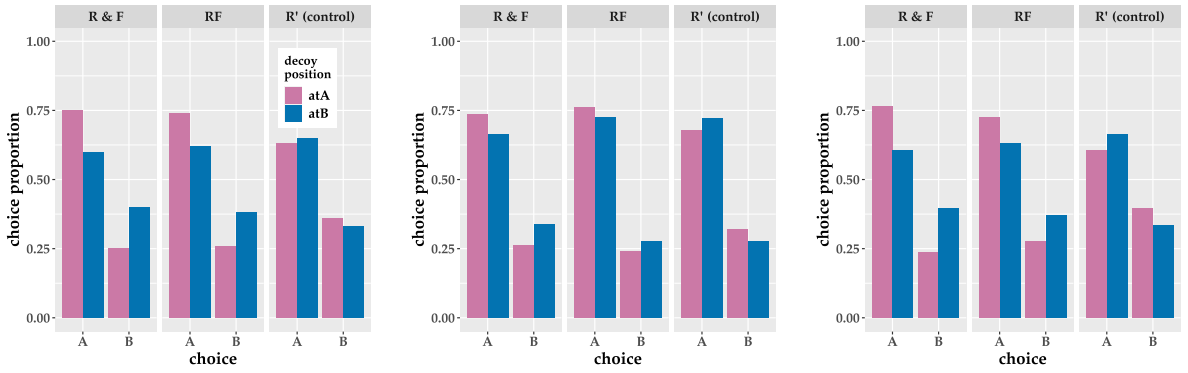
Imagine you are presented with these three bets. Choose the bet you would most prefer to take.

- *.40, $25*
- *.40, $22*
- *.30, $33*

All stimuli we used to construct the questionnaire are shown in Table 15, Appendix B, except for control (R′) decoys, which were constructed by altering the values of R decoys so that the control decoys were dominated by both targets (Wedell, 1991).

The goal of this experiment was to replicate (Wedell, 1991)'s finding on how the type and position of decoy influence participants' choices on each pair. Thus, we were specifically interested in having decoy type and decoy position as the two predictors. In Wedell (1991), each participant completed 10 pairs of questions. Decoy position (atA, atB) was a within-subject variable. Decoy type (R, F, RF in Wedell (1991)'s Experiment 1 and R and control/R′ in Wedell (1991)'s Experiment 2) and presentation order were between-subject variables. In our replication study, each person completed 10 pairs of questions. Both decoy position (atA, atB) and decoy type (R, F, RF, control/R′) were within-subject variables.

*Demographic information.* Participants answered a short demographic survey at the end. The questions include age, gender(male/female/other), age began to learn English, language used mostly at home, and highest grade completed.




(a) Wedell (1991), Economic gambles

(b) Experiment 1a, Economic gambles (N=150)

(c) Experiment 1b, Ethical decisions (N=51)

**Fig. 2.** Attraction effect across all trials, shown as choice proportions for A, B options in aggregate: (a). Attraction effect in Wedell (1991) with economic gambles; (b). Attraction effect in our (Wedell, 1991) replication study with economic gambles (Experiment 1a); (c). Attraction effect in Wedell-isomorphic ethical dilemmas (Experiment 1b).

### 3.1.2. Results

*Descriptive analysis.* We show the attraction effect[1] with choice proportions for the options across all trials below. Similar to what we see in Wedell (1991) results (Fig. 2a), we observe attraction effects across trials in this experiment (Fig. 2b).

To gain insights on how participants choose the alternatives and to visualize the results, we analyzed the proportion of within-subject choice reversal for R & F and RF decoy. To do this analysis, we categorized the response data into four types of response patterns for each pair of questions (e.g., Table 16, Appendix B): (1) choosing the same option for both questions in each pair (both As or both Bs), i.e., choosing consistently; (2) choosing targets for both questions in each pair (exhibiting a choice reversal); (3) choosing competitors for both questions in each pair (competitor reversal); (4) choosing decoy at least once.

As the majority participants had consistent choice within each pair regardless of decoy type, we focus on the target choice reversal rates and competitor reversal selection rates here (Fig. 3, full results see Appendix B).

In general, choice reversals occur around 20% of time within subjects (Table 17, Appendix B) and proportions of choice patterns (Fig. 3c) are consistent with Wedell (1991)'s original results (Fig. 3a) despite of the slightly high decoy selection rates (>10%) in our study. A clear choice reversal effect is observed for R & F decoys, whereas the choice reversal effect is weaker for the RF decoys. The control condition exhibited the least number choice reversals.

One common way to analyze choice data in context effects is analyzing the relative choice share among options, particularly, the relative choice share of target (RST; Berkowitsch, Scheibehenne, & Rieskamp, 2014), defined as $RST = \frac{Ntargets}{Ntargets+Ncompetitors}$, and an RST value larger than .5 indicates a systematic context effect (Berkowitsch et al., 2014).

We adopt Berkowitsch et al. (2014)'s method and further adapt it to our within-subject design. We focused on the target choice reversal rates and compared the relative choice shares of *target reversal* with that of *competitor reversal* by calculating a *relative choice share of the target reversal in a pair and define pairRST as*:

$$\text{pairRST} = \frac{N \text{ target reversals}}{N \text{ target reversals} + N \text{ competitor reversals}},$$

and an pairRST value larger than .5 indicates a systematic target choice reversal.

The choice shares of target reversal in a pair given different types of decoy are shown in Table 1. On the descriptive level, the pairRST of *R & F*, and *RF* decoy in both Wedell (1991) and experiment 1a exceeded .50, indicating a within-subject choice reversal effect. The pairRST of *R′ (control)* decoy was below .50 in both experiments.

*Bayesian statistical analysis.* Statistical analysis was conducted in R (R Core Team, 2013) using brms (Bürkner, 2017).

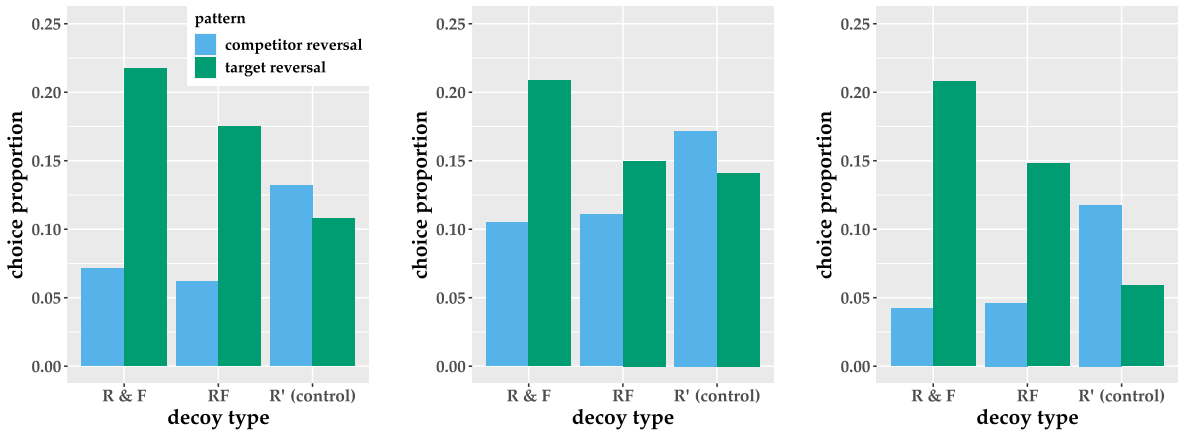
*Statistical models* We used Bayesian parameter estimation (Kruschke, 2011) to estimate the pairRST posterior 95% HDI with a hierarchical model with shrinkage towards a global mean:

$$\text{target reversal} \sim (1|\text{decoy type}),$$

while assuming a uniform prior.[2]

[1] As we focus on the contrast between R & F and RF decoys, we take the mean response rates for Wedell (1991)'s R and F decoys in our analyses.

[2] In pre-registration, we proposed a Bayesian multinomial logistic regression model to estimate the decoy type effect. We changed our analyses in response to reviewer suggestions. The pairRST analysis allows us to take into account choices of the competitor options and estimate the relative share of target reversals over both target reversals and reversed reversals. All previous versions of the preprint and scripts are available via OSF as well.




(a) Wedell (1991)

(b) Experiment 1a, Wedell (1991) replication (N=150)

(c) Experiment 1b, Ethical decisions (N=51)

**Fig. 3.** Within-subject response patterns (competitor reversals & target choice reversals) are shown in (a) for Wedell (1991) original study with economic gambles, (b) for Wedell (1991) replication study and (c) for ethical decisions.

**Table 1**
pairRST in Wedell (1991), Experiment 1a and 1b.

| Decoy | pairRST | Experiment |
|---|---|---|
| R & F | 0.75 | Wedell (1991) |
| RF | 0.74 | Wedell (1991) |
| R′(control) | 0.45 | Wedell (1991) |
| R & F | 0.66 | Replication |
| RF | 0.57 | Replication |
| R′(control) | 0.45 | Replication |
| R & F | 0.84 | Ethical choice |
| RF | 0.76 | Ethical choice |
| R′(control) | 0.33 | Ethical choice |

The statistical results show that within-subject choice reversals occurred with both *R & F* and *RF* decoys in the replication study (1a; Fig. 4a). *R & F* decoy condition had a strong within-subject choice reversal effect with an $\text{HDI}_{.95}$ of $(.59, .72)$, mean .66. *RF* decoy had a weaker effect with an $\text{HDI}_{.95}$ of $(.47, .67)$, mean .57. *R′ (control)* decoy had no effect, with an $\text{HDI}_{.95}$ of $(.37, .56)$, mean .47. Full posteriors of the model parameters are reported in Table 18, Appendix B.4.

### 3.2. Experiment 1b: Ethical choice

Following our Wedell (1991) replication, we created ethical dilemmas by transforming the tasks in Wedell (1991) into isomorphic problems. The isomorphs were created by preserving the numerical values of the $\langle p, v \rangle$ attributes of the original stimuli, but creating dilemmas from a forced choice among disaster rescue plans with different probabilistic outcomes for saving lives (see Fig. 5). The dilemma arises when one chooses between a plan with relatively moderately high probability of success but saving fewer lives, and a plan with a lower probability success but saving more lives. The tradeoff thus pits the lives of an imagined smaller group against the lives of a larger group.

#### 3.2.1. Method

*Participants.* Sixty participants were recruited from undergraduate psychology subject pool at the University of Michigan. Nine participants were excluded due to either: (1) not finishing the survey, or (2) failing the attention check question. In total, 51 participants (24 female; age M(SD) = 19(1.19) years) were included in the data analysis.

*Materials.* In our experiment, decoy position (dominated by A/target versus dominated by B/competitor) and decoy type R & F, RF, R′ (control) were manipulated as within-subject variables.

We constructed experimental materials that contains 40 pairs of ethical dilemmas (80 dilemmas in total), 10 pairs for each type of decoy. Each pair of dilemmas contained two questions with the same target and competitor choices but different decoys

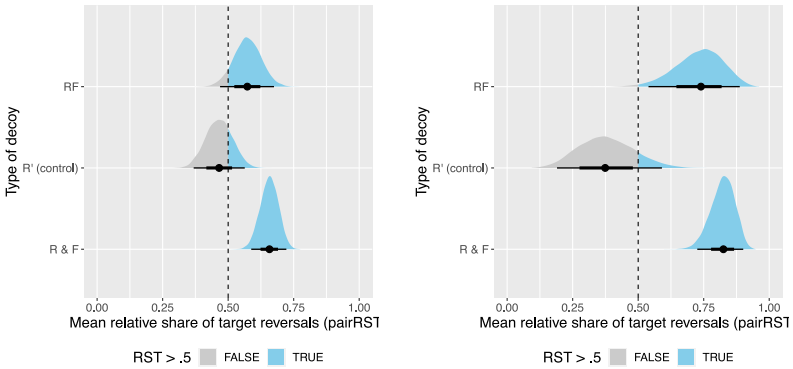



**Fig. 4.** Posterior pairRST in Experiment 1a and 1b given different types of decoy.

**Table 2**
An example of a question with the attributes *numbers of life to save* and *probability of saving*.

| A hurricane hits a small town causing most houses to be destroyed. Three emergency rescue plans have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows, which program would you choose? | |
|---|---|
| Options: | |
| *Decoy at A* | *Decoy at B* |
| A. A program that leads to a 40% chance of saving 25 people. | A. A program that leads to a 40% chance of saving 25 people. |
| D. A program that leads to a 40% chance of saving 22 people. | D. A program that leads to a 25% chance of saving 33 people. |
| B. A program that leads to a 30% chance of saving 33 people. | B. A program that leads to a 30% chance of saving 33 people. |

(dominated by A/at A and dominated by B/at B). Each participant completed a task that contains 10 random pairs of dilemmas and all questions were displayed in a random order. The task was implemented as a questionnaire using the Qualtrics software (Qualtrics, Provo, UT).

Since our ethical dilemmas are isomorphic to Wedell (1991) tasks, all numerical values of the attributes are identical to Wedell (1991)'s $\langle p, v\rangle$ attributes. The control decoys were created by altering the values of one of the 1D decoys so that the control decoys were dominated by both A and B.

An example of a pair of our task problems is shown in Table 2.

*Attention check.* We gave participants one attention check question. This question was always presented at the end of the survey in order to avoid interference with other scenarios. The question stated that the participant can only have time to save people from one out of the three rooms in a burning house and asks the participant to choose from saving 1, 3, or 5 people.

*Demographic survey.* Participants answered a short demographic survey at the end. The questions included age, gender (male/female/other), age began to learn English, language used mostly at home, and highest grade completed.

### 3.2.2. Results

*Descriptive analysis.* The descriptive analysis was the same as that in Experiment 1a.

Similar to the results in Wedell (1991) (Fig. 2a) and in our replication study (Fig. 2b), attraction effects appear across trials in this experiment (Fig. 2c). For the rest of the paper, we will include descriptive analyses of choice proportions across all trials in each experiment's corresponding Appendix.

Generally, within-subject choice reversals occur around 10% to 15% (see Table 17, Appendix B). The proportions of choice reversals and choosing the competitor in both questions in a pair ("competitor reversal") in our ethical decision making study are very similar to Wedell (1991) original results (Fig. 3a). A clear and strong choice reversal effect is observed for the R & F decoy type, whereas the choice reversal effect is less strong for the RF decoy. In the control condition, we observe the lowest choice reversal rates.

The choice shares of target reversal in a pair given different types of decoy are shown in Table 1. On the descriptive level, the pairRST of *R & F*, and *RF* decoy in both Wedell (1991) and experiment 1b exceeded .50, indicating a strong within-subject choice reversal effect. The pairRST of *R′ (control)* decoy was below .50 in both experiments.

*Bayesian statistical analysis.* We analyzed the data in the same way as we did in Experiment 1a. We used a Bayesian parameter estimation (Kruschke, 2011) to estimate the pairRST posterior 95% HDI with a hierarchical model with shrinkage towards a global mean:

target reversal ~ (1|decoy type),

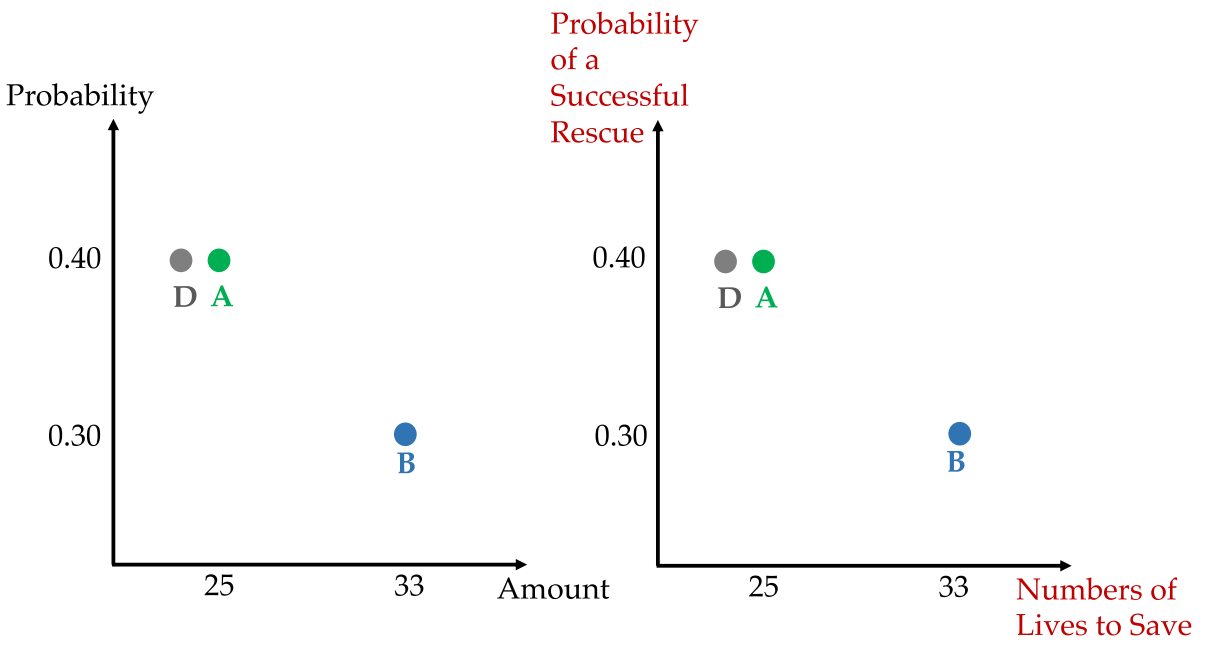



**Fig. 5.** The illustration of the task structures of Wedell (1991) decision problem (left) and our ethical dilemmas (right).

while assuming a uniform prior.

The statistical results are shown in Fig. 4b. In ethical choice study (1b), *R & F* decoy condition had a strong within-subject choice reversal effect with an $\text{HDI}_{.95}$ of $(.73, .90)$, mean .82. *RF* decoy had a slightly weaker effect with an $\text{HDI}_{.95}$ of $(.56, .90)$, mean .73. *R′ (control)* decoy had no effect, with an $\text{HDI}_{.95}$ of $(.18, .58)$, mean .38. Full posteriors of the model parameters are reported in Table 18, Appendix B.4.

### 3.3. A comparison of performance and choice reversal rates between economic gambles (1a) and ethical choices (1b)

The economic gambles in our Wedell replication and the ethical isomorphs in Experiment 1 both allow us to assess behavioral choices in terms of expected value as an objective baseline. More specifically, the mean expected value of each participant's performance can be used as a measure that allows us to compare performance in the two experiments. We are also interested in comparing the rate of choice reversals (specifically target reversals[3]) in the two experiments, and understanding how the reversal rate might be related to performance across participants.

We therefore conducted a *post-hoc, exploratory* Bayesian analysis with rstanarm (Goodrich, Gabry, Ali, & Brilleman, 2020) to compare participants' performances and target reversal rates in the economic gambles (Experiment 1a) and in the ethical decisions (Experiment 1b). The statistical models are:

$$\begin{aligned}\text{performance} &= \beta^0_{performance} + \beta^1_{performance}\text{economic gambles}\\ \text{target reversal} &= \beta^0_{targetreversal} + \beta^1_{targetreversal}\text{economic gambles},\end{aligned}$$

where the intercepts, $\beta^0$s, estimate the mean performance and mean target reversal rates in ethical decisions (reference group); $\beta^1$s estimate the mean difference in the performance and target reversal rates in ethical decisions and in economic gambles. All trials with control decoys were excluded from the analyses as these decoys were not expected to produce choice reversals in the first place. We used the default non-informative priors and ran four independent chains (4000 samples each, with the first 200 samples as the *warmup*).

In terms of performance, we found that participants in Experiment 1b (ethical decisions; $N = 51$, $M = 9.96$, $SD = 0.06$) performed better compared to participants in economic gambles (1a; $N = 150$, $M = 9.83$, $SD = 0.22$; $\beta^1_{performance} = -0.13$, 95% CI $= [-0.19, -0.07]$). The distributions of performances in two are displayed in Fig. 6, and the full posteriors are reported in Appendix B, Table 19.

In terms of target reversal rates, participants in Experiment 1b ($N = 51$, $M = 0.20$, $SD = 0.20$) have slightly higher reversal rates than participants in economic gambles (1a; $N = 150$, $M = 0.15$, $SD = 0.16$; $\beta^1_{targetreversal} = -0.04$, 95% CI $= [-0.10, 0.01]$). However,

[3] Given that each participant completed only 10 pairs of questions, and that the majority choices are consistent, the sample size for calculating pairRST for each individual becomes extremely small. Therefore, we choose to compare performance and target reversal rates directly instead of using pairRST in this exploratory analysis.

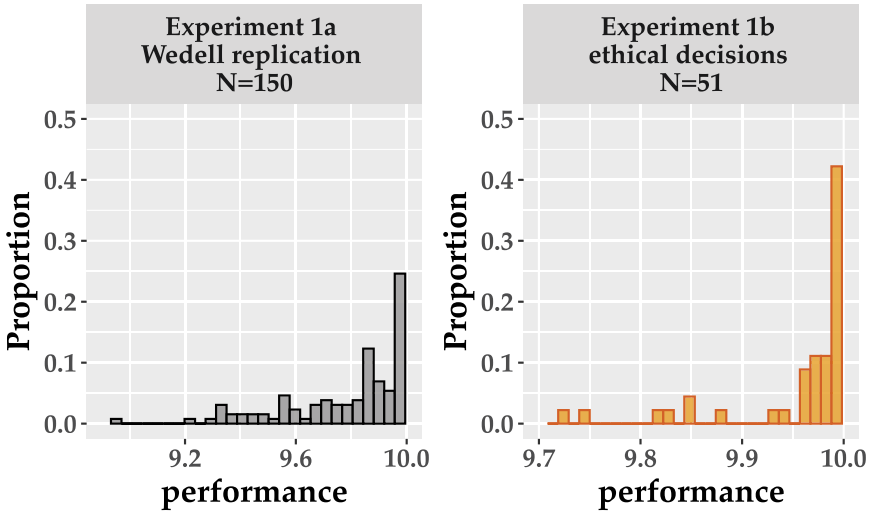



**Fig. 6.** Distributions of performances (mean EV) in economic gambles (Experiment 1a) and in ethical decisions (Experiment 1b).

note that the 95%CI includes 0. The distributions of target reversal rates in two are displayed in Fig. 7b, and the full posteriors are reported in Appendix B, Table 20.

We also investigated the relationship between participants' performances and their within-subject target reversal rates in economic gambles and ethical decisions with a simple Bayesian regression model using rstanarm (Goodrich et al., 2020):

$$\begin{aligned}\text{performance} =& \beta^0 + \beta^1 \text{target reversal} + \beta^2 \text{ethical decisions} \\ &+ \beta^3 \text{target reversal} * \text{ethical decisions},\end{aligned}$$

where "target reversal" is the centered values of an individual's target reversal rate and "ethical decisions" is a grouping variable, indicating whether the values correspond to Wedell(1991) replication study (reference group) or ethical decisions.

We used the default non-informative priors and ran four independent chains (4000 samples each, with the first 200 samples as the *warmup*). We found that target reversal rates predict higher performance in economic gambles (see Figure Fig. 7a; $\beta^1 = 0.60$, 95%CI = $[0.26, 0.93]$), but not in ethical decisions ($\beta^3 = -0.63$, 95%CI = $[-1.26, 0.01]$; $\beta^1 + \beta^3 < 0$)—plausibly because there is a ceiling effect in the ethical decisions (despite the higher reversal rate). The posterior estimates for model parameters and their 95% CIs are fully reported in Appendix B, Table 21. This relationship is interesting, because it suggests that higher reversal rates are not necessarily associated with lower performance — in fact the contrary, consistent with recent analyses suggesting that reversals are signatures of computationally rational choice strategies (Howes, Warren, Farmer, El-Deredy, & Lewis, 2016).

### *3.4. Discussion*

Our study investigates and observes contextual choice reversals in the domain of ethical decision making by using tasks with the same structure as classic contextual choice reversal studies. By providing participants with ethical dilemmas that are isomorphic to the economic gambles in Wedell (1991), we found evidence for choice reversals when the decoy was either R & F or RF, with the former having a slightly stronger reversal effect. Past work has already suggested that different decoy positions may produce different sizes of the attraction effect. While Wedell (1991)'s results did not suggest a significant difference in the main effect of decoy positions (R, F, and RF), both Huber and Puto (1983) and Trueblood et al. (2013) found that the R decoy has the largest effect, followed by RF decoy, and lastly F decoy. It is worth noting that While we found a stronger effect of R & F decoy compared with RF decoy, by not distinguishing between R and F decoys, we are not able to make further comparisons between these two types of decoys.

Through a post-hoc, exploratory comparison of the data from Experiment 1a and 1b, we also found differences in performance and the relationship between performance and target reversal rates in data from economic gambles and ethical decisions: (1) the performance in ethical decisions is better; (2) higher target reversal rates predict better performance in economic gambles but not in ethical decisions, possibly due to a ceiling effect in ethical decisions. While this comparison was an unplanned exploration (i.e., two experiments were conducted at different times and thus participants were not randomly assigned between the two experiments), we believe this result provides insights into the bounded rationality of utility maximization in choices. We will discuss this in more detail in the conclusion section.

Although we found evidence for contextual choice reversals in the ethical domain, we acknowledge several drawbacks of this Experiment. First, this task only involves two attributes (the probability of saving lives and numbers of lives that can be saved), both of which are measured on continuous scales, making it simple to compare the levels within each attribute (e.g., saving more lives

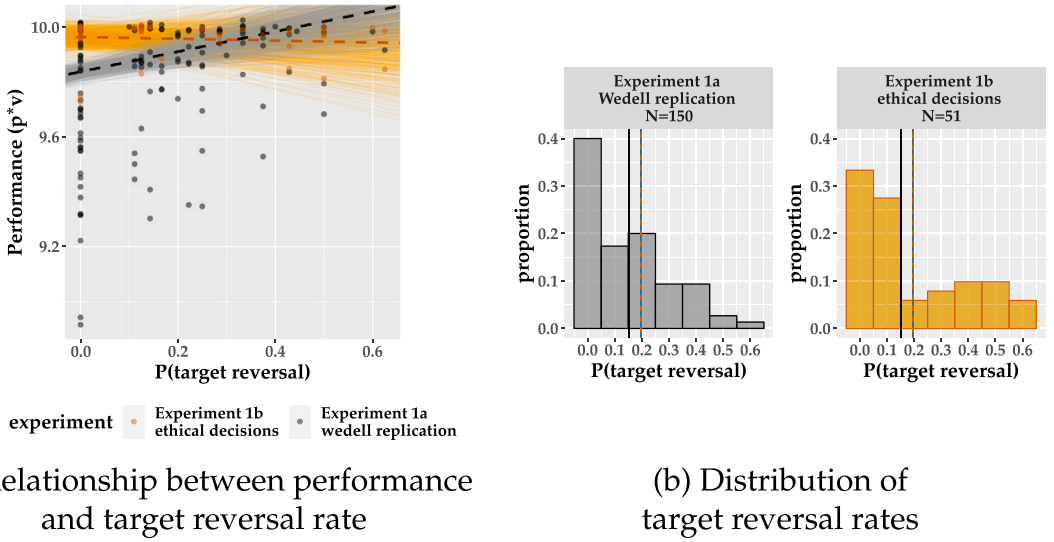



**Fig. 7.** (a) Relationship between performance and target reversal rate — each dot represents one subject. The light-color lines are constructed from 500 random samples from the total 15 200 posterior samples and the dashed lines are medians from the posterior distributions of model parameters; (b) Distributions of target reversal rates in economic gambles (Experiment 1a) and ethical decisions (Experiment 1b). The lines show overall target reversal rates in two experiments, and the blue dashed line shows the overall target reversal rates in Wedell (1991). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is preferred and higher probability is preferred). As such, some participants may view the task as a purely quantitative economic gamble and lose sight of the premise that the lives of two distinct groups are at stake and one must be abandoned. Second, this task only involves one scenario — choosing a natural disaster rescue plan. In reality, people face ethical dilemmas in all kinds of scenarios — some could be high-level decisions such as determining rescue plans, but some could also be more immediate, personal decisions such as whether one should spend more money on a product that is fair-trade and environmental friendly. Based on our current task, we cannot directly generalize our findings to other ethical domains without further investigations, as ethical choices in other domains involve very different scenarios and attributes.

In the following experiments, we explore contextual choice reversals in ethical decisions further by creating tasks spanning various ethical domains, from personal choices involving weighing one's financial interest against fair trade practices, to choices connected to policies that influence the lives of groups of people.

## 4. Experiment 2

In our previous experiment, we found empirical evidence for contextual choice reversals in a domain that involves ethical decision making. However, the previous study only involved repeated tasks that contained the same two numerically-expressed attributes (probability of saving lives and numbers of lives saved) in a rescue-plan-selection scenario. Therefore, our new study aims to:

1. replicate our existing finding that contextual choice reversals can be observed in ethical decisions.
2. follow the design of the previous study and create ethical decisions in various domains that go beyond "ethical gambles".

One of the most challenging aspects of designing this experiment is to map the structure of contextual choice reversal tasks to a greater variety of ethical decisions (rather than just *the probability of saving lives* and *numbers of lives saved*). To create the structure of contextual choice reversal tasks with a dominating target, a competitor, and a decoy option, we need three options with two attributes, and two to four levels in each attribute. This allows us to have the structure where the target and the competitor each is a dominating option on only one attribute, and the target dominates the decoy on both attributes (Fig. 8). If one of the two attributes have only two levels, then the target can only dominate the decoy on one attribute. The target and competitor options also need to be approximately equally attractive.

Constructing stimuli with a desired dominance structure requires us to make assumptions about the ranking of attribute levels. While *number of lives saved* has a clear objective rank that is likely to be consistent across individuals (i.e., larger numbers of lives saved are better), some of our ethical scenarios require us to use attributes whose levels do not have such clear rankings. Our approach is to find a ranking of the levels that is likely to be endorsed by the majority of participants (we explicitly address the possibility of individual variation in ranking in Section 6). For example, in a dilemma, participants need to decide which prisoner to release while making a trade-off between the age of the victim and crime motivation. All prisoners in this dilemma robbed victims of different age and their crime motivations range from "to pay off gambling debt" to "to buy medication for their sick parent" or "to buy medication for their sick child". Although we may assume that it is more permissible to commit a robbery to save a loved

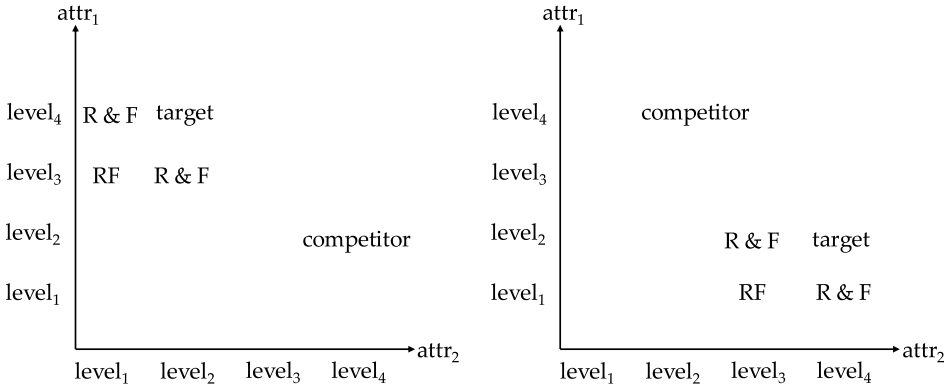



**Fig. 8.** The structure for a pair of dilemmas that have the structure of a classic contextual choice reversal task. If both attributes have four levels, then we can create RF decoys. If one of the attributes have two levels, we can create a pair of dilemmas with R or F decoys. For example, if $attr_1$ only has level 2 and level 4, the required relationships among target, competitor, and decoy still hold, with the decoy being R or F.

**Table 3**
Experiment materials for constructing ethical dilemmas: attributes and their four levels.

| Attribute | Levels |
|---|---|
| Car crash victim age | Baby, child, teenager, adult |
| Crime | Stealing a laptop from an unattended room, physical assault without a weapon, physical assault with a gun, rob a person at gun point |
| Crime (theft) location | Warehouse, local pharmacy, elementary school, someone's home |
| Crime (theft) victim's age | Child, teenager, middle-aged person, old person |
| Crime (theft) motivation | For a sick child, for a sick parent, for a friend's sick pet, to payoff gambling debt |
| Pollution | Low, medium, high, very high |
| Emergency delivery speed | Overnight, 3 days, 5 days, 7 days |
| Responsibility | Often miss work but finish most tasks, often late to work but finish most tasks, finish tasks, very loyal and do an excellent job |
| Employment duration | 6 months, a year, 2 years, 3 years |
| Car crash injury level | Lose a leg, lose both legs, total paralysis, death |
| Shoe cost | Low ($55), medium ($68), high ($86), highest ($113) |
| Computer cost | Low ($192), medium ($258), high ($532), highest ($1131) |
| Shoe salary | Use child labor, pay the workers poorly, pay the workers fairly, pay the workers well and provide health benefits |
| Computer salary | Use child labor, pay the workers poorly, pay the workers fairly, pay the workers well and provide health benefits |

one than to pay off gambling debt, there is no objective standard on which motivations are more permissible, especially considering that each decision maker has their own individual experience and ethical values.

To find the majority-preferred rankings of levels in each attribute, we conducted a small study with 57 participants from the psychology undergraduate subject pool at the University of Michigan.

In this study, we investigated individuals' preferences among four levels of each attribute that we would use to construct ethical dilemmas subsequently. The attributes and each attribute's four levels are shown in Table 3.

We constructed six pairwise comparisons among the four levels of each attributes and all subjects were asked to make a choice in each pairwise comparison. An example of a set of questions that are pairwise comparisons among the four levels of the attribute "crime motivation" is included in Appendix C.1.

We found that there were consistent preferences among participants in some attributes (such as speed of delivery) but not in all attributes. There were individual variations in the preferences for ranking of the levels within an attribute. To create materials for our tasks, we kept the attribute for which more than half of the participants provided the same ranking orders. For example, for the "crime motivation" attribute, >50% participants ranked the levels as "stealing prescription drugs for a sick child" > "stealing prescription drugs for a sick parent" > "stealing prescription drugs for a friend's sick pet" > "stealing prescription drugs to pay off gambling debt". With such information, we were able to construct tasks for Experiment 2.

Experiment 2 involves three parts. Part 1 aims to check that participants' preferences are consistent with the rankings from the pilot study. Part 2 and Part 3 are the 9 ethical dilemmas with Wedell (1991)-like structures. Details of the three parts are provided below in the Method section.

**Table 4**
The nine scenarios we created for Experiment 2 and their two attributes/dimensions.

| Scenario | Attribute 1 | Attribute 2 |
|---|---|---|
| Emergency delivery | Speed of an emergency drug delivery | The amount of pollutant produced by the vehicle |
| Jail overcrowding | Motivation for committing a robbery | Probability of re-committing the same crime |
| Jail overcrowding 2 | Motivation for committing a robbery | Age of the victim |
| Inevitable injury | Type of injury in an inevitable car accident | Probability of the injury |
| Rescue plan | Number of lives to save in a rescue | Probability of saving the lives successfully |
| Rescue a survivor | Age of the survivors in a natural disaster | Probability of saving each survivor |
| Firing an employee | How much sense of responsibility an employee has | How many years an employee has worked at the company |
| Worker welfare | Price of the laptop | How well the company that sells the laptop treats its workers |
| Worker welfare 2 | Price of a pair of boots | How well the company that sells the boots treats its workers |

**Table 5**
An example of a question with the attributes *motivation for committing a robbery* and *probability of recommitting the same crime*. The latter is the same across options.

| Decision problem: | |
|---|---|
| You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner (who committed robbery). Which prisoner would you release? | |
| Options: | |
| *Motive* | *Probability of recommitting* |
| To buy medication for his friend's sick pet | 40% |
| To buy medication for his sick child | 40% |
| To payoff his gambling debt | 40% |

## 4.1. Method

### 4.1.1. Participants

We recruited 502 U.S. participants (256 female; age M(SD) = 33(12.26) years) from Prolific (www.prolific.co) to complete this study in three sessions.[4] We included the 475 participants (242 female; age M(SD) = 33(12.19) years) who completed all three parts in the data analyses.

### 4.1.2. Design and materials

This experiment follows a 2 × 2 mixed design with one between-subject variable and one within-subject variable. The between-subject variable is decoy type (R & F vs. RF) and the within-subject variable is decoy position (atA, atB). We have 9 items/scenarios in total (Table 4). For the complete descriptions of the scenarios, see Appendix C.2.

Among all scenarios, 7 scenarios (*emergency delivery, jail overcrowding, rescue plan, rescue a survivor, firing an employee, worker welfare, worker welfare 2*) have both R or F (labeled as R & F) and RF decoys whereas 2 (*jail overcrowding 2, inevitable injury*) have only R & F decoys. Thus, as each participant sees all 9 items, decoy type is only between-subject for the 7 scenarios that have both R & F and RF decoys. Each item is a pair of questions: in one question, decoy is at A, and in another, at B.

Each scenario poses a dilemma where the welfare of different sides is at stake. For a brief ethical content analysis of the scenarios explaining how each scenario poses a dilemma, see Appendix C.3.

*Demographic survey.* At the end of part 1 of this experiment, all participants also answered a short demographic survey at the end. The questions included age, gender (male/female/other), age they began to learn English, language used mostly at home, and highest grade completed.

### 4.1.3. Procedures

Participants completed part 1, part 2, and part 3 of the experiment in three separate sessions, each session activated on Prolific (www.prolific.co) the day after the previous session.

In Part 1, each participant completed 16 ethical decision tasks. In each task, each participant makes a choice among three multi-attribute options: one attribute is the same among the options and another attribute varies on three different levels. The 16 decision tasks correspond to the attributes in our ethical scenarios. An example of a question is in Table 5. In this example, the attribute *probability of recommitting the same crime* is the same among the options whereas the attribute *motivation* varies. An example of the full set of questions corresponding to the *speed* and *pollution* attributes is given in Appendix C.4.

Each participant completes 9 decision tasks in each part of Part 2 and Part 3. The 9 decision tasks in each part correspond to the 9 ethical scenarios. To manipulate decoy position and decoy type of the decision tasks in each part, we have created four different versions of the tasks for Part 2 and 3, presented in Table 6.

[4] Sample size was determined based on a power analysis based on our pre-registered data analysis method (see SI).

**Table 6**
The four task versions in part 2 and 3 of Experiment 2.

| Part | Version | Decoy position | Decoy type |
|---|---|---|---|
| 2 | 1 | atA | R & F |
| | 2 | atA | RF |
| | 3 | atB | R & F |
| | 4 | atB | RF |
| 3 | 1 | atB | R & F |
| | 2 | atB | RF |
| | 3 | atA | R & F |
| | 4 | atA | RF |

Each participant was randomly assigned one of the four versions of tasks. If a participant completed version 1 in Part 2, then they would also complete version 1 in Part 3. For each task version, Part 2 and Part 3 differ in decoy position, but not in decoy type.

All decision tasks were implemented in Qualtrics software (Qualtrics, Provo, UT). Four versions of the tasks were setup on Prolific (www.prolific.co) as four separate studies, thus the random assignment of the task version was done as participants signed up for the studies on Prolific.

### 4.2. Results

#### 4.2.1. Descriptive analysis

Part 1 results showed that participants mostly made consistent choices when they were asked to choose among three options representing three of four levels in an attribute. If a participant did not choose the assumed best option in a decision problem corresponding to an attribute, we marked that participant-attribute pair as inconsistent. In SI, we included descriptive results after excluding trials corresponding to inconsistent participant-attribute pairs (452 pairs, and the results did not change).

In the following analyses, we exclude the *firing an employee* item due to an experimental error and the *rescue a survivor* item due to its extremely high decoy selection rates (R & F decoy: .45; RF decoy: .42). The exclusion of the *rescue a survivor* item in statistical analyses does not change any of the following conclusions, and the full results *with* the *rescue a survivor* item are included in SI. We also focus on the aggregated data across items. For the complete descriptive results by each item, see Appendix C.5.

We present the choice proportions across all trials and all dilemmas in Appendix C.5. Here, we focus on analyzing the within-subject choice reversals. The response patterns within each pair of dilemmas are coded the same way as in Experiment 1. As expected, participants most frequently selected the option consistently (although less compared with the results in Wedell, 1991). We present only response rates for decoy selection ("decoy selected"), choosing the competitors for both questions in a pair ("competitor reversal"), and within-subject choice reversals ("target reversal"). For complete proportions of choice patterns in Experiment 2, see Fig. 19, Appendix C.5.

As the aggregated data (see Fig. 9) show, choice reversal rates (R & F decoy: 0.22; RF decoy: 0.23) in Experiment 2 are higher than those in Wedell (1991). However, the rates for decoy selection (R & F decoy: 0.07; RF decoy: 0.04) and competitor reversals (R & F decoy: 0.15; RF decoy: 0.23) are slightly higher than those in Wedell (1991) as well. We do not observe any difference between R or F and RF decoys.

We also show the pairRST by scenario in Table 7, which suggested that choice reversals occurred in some scenarios with both types of decoy, but not all scenarios.

#### 4.2.2. Bayesian statistical analysis

Data analysis was conducted in R (R Core Team, 2013) using brms (Bürkner, 2017).

*Statistical models.* Similar to Experiment 1, we estimated the relative choice share of within-subject choice reversals (pairRST) with a hierarchical model

$$\text{target reversal} \sim \text{decoy type} + (1|\text{scenario}),$$

assuming a uniform prior. In our analysis, we set the *R & F decoy* as the reference category.

The pairRST posterior 95% HDI suggested (Fig. 10) that within-subject choice reversal occurred in four items given *R & F* decoy: *inevitable injury* ($\text{HDI}_{.95}$ : $(.73, .84)$, mean .78), *jail overcrowding* ($\text{HDI}_{.95}$ : $(.60, .73)$, mean .67), *rescue plan* ($\text{HDI}_{.95}$ : $(.53, .67)$, mean .60), and *emergency delivery* ($\text{HDI}_{.95}$ : $(.53, .67)$, mean .60). There was also a weak effect in *jail overcrowding 2*, but the range of the 95% HDI indicated high uncertainty of the estimation ($\text{HDI}_{.95}$ : $(.40, .67)$, mean .54). With *RF* decoy, within-subject choice reversals occurred in three items: *jail overcrowding* ($\text{HDI}_{.95}$ : $(.59, .73)$, mean .66), *rescue plan* ($\text{HDI}_{.95}$ : $(.52, .67)$, mean .60), and *emergency delivery* ($\text{HDI}_{.95}$ : $(.52, .67)$, mean .60). The full posteriors of model parameters are reported in Table 23, Appendix C.5. We did not see a difference between R& F and RF decoy (for decoy type parameter (baseline R & F), $\text{CI}_{.95}$ : $(-0.1, 0.10)$, mean 0.00).

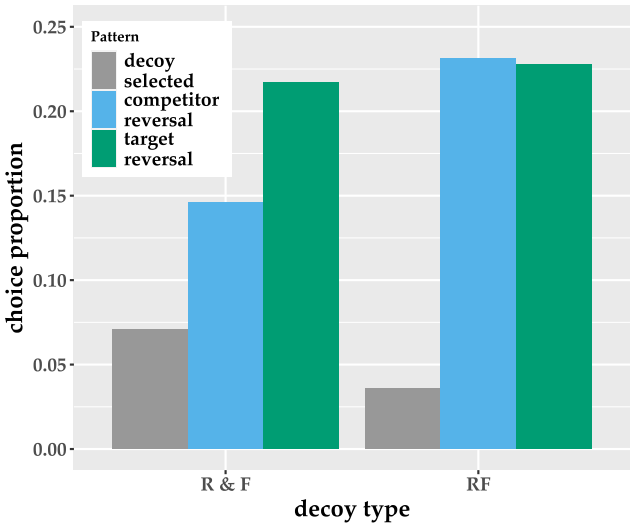



**Fig. 9.** Response patterns (competitor reversals & target/choice reversals) aggregated over all items in Experiment 2 (N = 475).

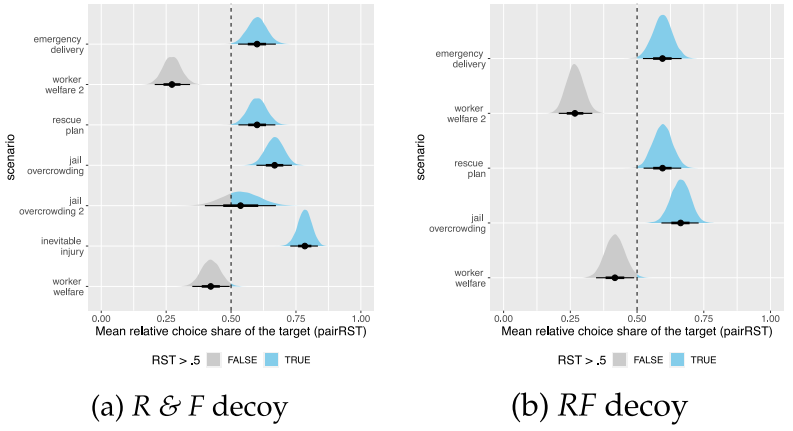



**Fig. 10.** Posteriors of pairRST in each scenario in Experiment 2.

**Table 7**
pairRST in Experiment 2.

| Scenario | Decoy | pairRST |
|---|---|---|
| Worker welfare | R & F | 0.43 |
| Inevitable injury | | 0.79 |
| Jail overcrowding 2 | | 0.54 |
| Jail overcrowding | | 0.63 |
| Rescue plan | | 0.59 |
| Worker welfare 2 | | 0.24 |
| Emergency delivery | | 0.68 |
| Worker welfare | RF | 0.40 |
| Jail overcrowding | | 0.71 |
| Rescue plan | | 0.61 |
| Worker welfare 2 | | 0.28 |
| Emergency delivery | | 0.52 |

### *4.3. Discussion*

This study focuses on exploring contextual choice reversals in decisions in various ethical domains. Similar to Experiment 1, our experiment follows the same structural design for the ethical dilemmas to create a choice environment commonly used in tasks

that investigate contextual choice reversals. Different from Experiment 1, this study extends the ethical domains beyond where the decision maker needs to choose a rescue plan. We created various scenarios that touch on environmental concerns (e.g., *emergency delivery*) and human rights concerns (e.g., *worker welfare*). We also moved beyond using task stimuli that are isomorphic to Wedell (1991) tasks. Instead, the attributes in these dilemmas are often qualitative — not all attributes are on a continuous scale — thus, the decision maker cannot calculate expected values directly by multiplying the attribute values together.

We found evidence for contextual choice reversals in aggregate, and we observed variations among individual items — we found very clear contextual choice reversals in some items, but not in all. First, we found clear choice reversals in the *rescue plan* item, an item taken directly from Experiment 1 and isomorphic to Wedell (1991) tasks. This results replicates Experiment 1. Other items that produced clear contextual choice reversals include *inevitable injury, jail overcrowding, emergency delivery*. In the following discussion, we put focus on the items that did not produce contextual choice reversals and consider why they did not. The choice proportions for each option in each item (Fig. 21, Appendix C.5) can provide us with insights.

The item *jail overcrowding 2* had slightly high decoy selection rates and very consistent selection rates. In this item, the decision makers have to choose which prisoner to release — all prisoners committed robbery but for different reasons. There was an overwhelming preference for option B (one who robbed a teenager to buy medication for his sick child; Fig. 21d, Appendix C.5). In other words, the two options are not equally attractive.

The unbalance could be explained on two levels. First, it is possible that one attributed is weighed more than the other during the decision making process — in this case, people could be generally more forgiving for those who try to save their child. Second, such unbalance is closely related to the challenges of mapping the structure of contextual choice reversal tasks to ethical decisions with attributes that have discrete levels. Although we did a pilot study to construct attributes with levels that have a majority-preferred ranking, we did not investigate how each participant weighs the attributes or their risk preferences. Therefore, we still do not know if the difference between two levels are equal across attributes. This could suggest that the differences between target's and competitor's levels for victim age attribute (middle aged person, teenager) are not distinctive enough. Thus, the options may seem to be closer together on the victim age attribute, pushing people to focus more on the motivation attribute.

Similar unbalance is observed in *worker welfare* and *worker welfare 2* (Fig. 21c & g, Appendix C.5). It is possible that the prices of the products are not distinguished enough, pushing people to focus more on the employee payment attribute.

Finding contextual choice reversals in some items in Experiment 2 suggests that there is potential to generalize our finding of contextual choice reversals to more ethical domains. However, we also face many challenges. The main challenges are: (1) despite finding attributes in various scenarios that have levels with a majority-preferred ranking, not everyone have the same ranking in the given context; (2) we only have ordinal information on the discrete levels of an attribute, but we cannot know to what extent the levels differ from each other. Lastly, the items we used in our scenarios could be memorable. It is possible that people could remember the scenario even as they completed Part 2 and Part 3 one day apart.

In the following study, we revise the items accordingly to address these potential issues.

## 5. Experiment 3

Our Experiment 2 found empirical evidence for contextual choice reversals in a variety of ethical domains, providing us with potential to generalize our results in the ethical domain. However, we also found that item/scenario variations where some items produced contextual choice reversals and some did not.

In this experiment, we made four main changes to Experiment 2:

1. **Item revisions.** We excluded the previous *rescue a survivor* item due to its extremely high decoy selection rates. We included the *firing an employee* item correctly. We modified the three items *worker welfare, worker welfare 2,* and *motivation & victim age* separately to make the differences among price attributes and victim age attribute larger.
2. **Randomization for decoy types.** Instead of having three sessions, we put previous Part 2 and Part 3 together into one session. We manipulated decoy type as a within-subject variable. This allowed us to randomly present the R & F or RF version of each item to each participant.
3. **Manipulation of task instructions.** Past studies have suggested that context effects such as the attraction effect could be weakened with additional manipulations that shift the decision maker's attention (e.g., highlighting the category of options in commercial product decisions; Ha, Park, & Ahn, 2009). We explored whether the knowledge of the existence of the dominance relationship between the target option and the decoy would increase contextual choice reversals by pushing the participants to look for the direct comparison between a dominating option (i.e., the target) and a dominated option (i.e., the decoy). Thus, we added the new between-subject manipulation of instruction, where participants will be randomly given an instruction that introduces the dominance of the target over the decoy.
4. **Fillers.** We introduced 16 fillers to make the critical items that we constructed less distinguishable.

This experiment involves two parts. As in Experiment 2, Part 1 aims to check that participants' preferences are consistent with ranks of various levels of the attributes that we used to construct the dilemmas. Part 2 contains the 8 critical items with Wedell (1991)-like structures. Details of the two parts are provided below in the Method section.

**Table 8**
The eight critical scenarios in Experiment 3 and their two attributes/dimensions.

| Scenario | Attribute 1 | Attribute 2 |
|---|---|---|
| Emergency delivery | Speed of an emergency drug delivery | The amount of pollutant produced by the vehicle |
| Jail overcrowding | Motivation for committing a robbery | Probability of re-committing the same crime |
| Jail overcrowding 2 | Motivation for committing a robbery | Age of the victim |
| Inevitable injury | Type of injury in an inevitable car accident | Probability of the injury |
| Rescue plan | Number of lives to save in a rescue | Probability of saving the lives successfully |
| Firing an employee | How much sense of responsibility an employee has | How many years an employee has worked at the company |
| Worker welfare | Price of the laptop | How well the company that sells the laptop treats its workers |
| Worker welfare 2 | Price of a pair of boots | How well the company that sells the boots treats its workers |

**Table 9**
The two task versions in part 2 of Experiment 3.

| Block | Version | Decoy position | Fillers |
|---|---|---|---|
| 1 | 1 | atA | Filler version 1 |
| | 2 | atB | Filler version 2 |
| 2 | 1 | atB | Filler version 2 |
| | 2 | atA | Filler version 1 |

### 5.1. Method

#### 5.1.1. Participants

We recruited 500 U.S. participants from Prolific (www.prolific.co) to complete this study in two sessions (demographic data were collected during the second session) and 480 participants (260 female; age M(SD) = 32(11.32) years) finished both sessions. We included the 456 participants (251 female; age M(SD) = 32(11.38) years) who passed the attention check in the data analyses.

#### 5.1.2. Design and materials

This experiment follows a $2 \times 2 \times 2$ mixed design with 1 between-subject variable and 2 within-subject variables. The between-subject variable is whether the participant receives the instruction explaining dominance or not. The within-subject variables are decoy type (R & F vs. RF) and decoy position (atA, atB). We have 8 critical items/scenarios (Table 8) and 16 filler items/scenarios. The specific descriptions of the critical items can be found in Table 22, Appendix C.2. Among the 8 critical items, 6 items (*emergency delivery, jail overcrowding, rescue plan, firing an employee, worker welfare, worker welfare 2*) have both R & F and RF decoys whereas 2 items (*jail overcrowding 2, inevitable injury*) have only R & F decoys. Fillers do not have decoys, but they have a structure that imitates that of critical items. In other words, each filler also has two versions where two out of the three options are the same and the third option varies in the two versions. All participants saw see all 24 items (48 questions).

*Attention check.* To check whether participants read the scenarios carefully, we created eight multiple-choice questions asking about various details in the scenarios, such as "which of the following is not a motivation for committing a robbery in the decision problems". Each participant was randomly presented with five out of eight questions. If a participant answers at least three questions correctly, they pass the attention check.

*Demographic survey.* Between the two blocks in part 2 of the study, participants answered a short demographic survey at the end. The questions included age, gender (male/female/other), age they began to learn English, language used mostly at home, and highest grade completed.

#### 5.1.3. Procedures

Participants completed Part 1 and Part 2 of the experiment in two separate sessions, each session activated on Prolific (www.prolific.co) the day after the previous session.

Part 1 of the experiment is the same as the Part 1 of Experiment 2. Each participant completed 30 ethical decision tasks. The tasks correspond to attributes that appear in critical items and attributes that appear in fillers.

Before participants started Part 2 of the experiment, they received an instruction about the decision tasks. The instruction is either a task instruction asking the participants to follow the instructions in the question carefully and choose the action they would be most likely to take in the given scenario or an instruction with additional explanations on what a dominating option and a dominated option is using an example from Huber et al. (1982). In the latter instruction, participants were also be told that some of the scenarios they see in the task will have a dominating option and a dominated option. Each participant was randomly presented with either a task instruction only or with the instruction that has explanations on dominance.

Part 2 of the experiment contains 24 pairs of multiple-choice questions (48 questions in total). 8 pairs are critical items and the other 16 pairs are fillers. All 24 pairs of questions were presented in 2 blocks, with a demographic survey separating them. To manipulate decoy position, we have created two different versions of the tasks in part 2, presented in Table 9.

Each participant was randomly presented with one of the two versions of tasks. If a participant completed version 1 in block 1, then they would also complete version 1 in block 2. Each block contains 24 questions, presented in a random order. Block 1

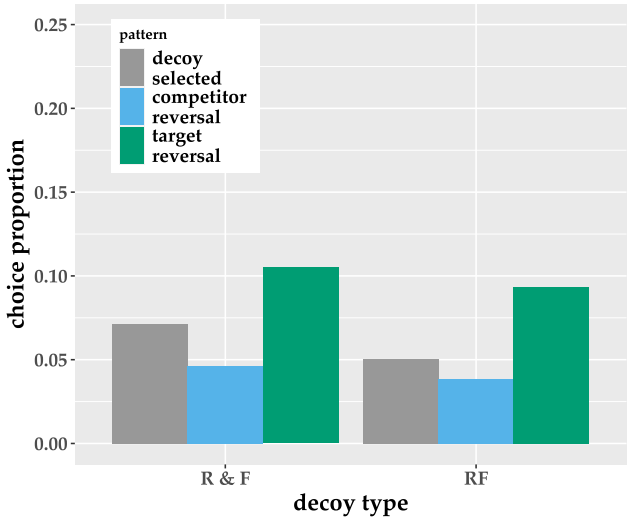



**Fig. 11.** Aggregated response patterns for all items (excluding *firing an employee*) in Experiment 3 (N = 456).

and 2 differ in decoy position and filler versions. When each item is presented to a participant *for the first time*, its R & F or RF decoy version is presented randomly. If an item's R & F decoy version is presented in block 1, then its R & F decoy version will be presented in block 2 as well.

At the end of Part 2, all participants completed a section of attention check questions, asking them details about the scenarios in the decision tasks.

All decision tasks were implemented and randomized in Qualtrics software (Qualtrics, Provo, UT).

### 5.2. Results

#### 5.2.1. Descriptive analysis

Part 1 results showed that participants were mostly consistent when they were asked to make a choice among three options representing three of four levels in an attribute. If a participant did not choose the assumed best option in a decision problem corresponding to an attribute, we marked that participant-attribute pair as inconsistent. In SI, we included descriptive results after excluding trials corresponding to inconsistent participant-attribute pairs (391 pairs, and the results did not change).

In the following analyses, we exclude the *firing an employee* item due to its distinctively high decoy selection rate (R & F decoy: .26; RF decoy: .18). The exclusion of the *firing an employee* item does not change any of the following conclusions, and the full results *with* the *firing an employee* item are included in SI. We also focus on the aggregated data across items. The complete descriptive results by each item are in Appendix C.6.

Same as in previous experiments, we present the proportions of choice across all trials and all dilemmas in Appendix C.6. We focus on the proportions of choice patterns in Experiment 3 in Fig. 11, excluding rates for consistent choices (see Fig. 23, Appendix C.6 for full results). The consistent-selection rates are extremely high (R & F decoy: .78; RF decoy: .82), indicating that participants mostly chose consistently between the pair of questions corresponding to the same item. The rates for contextual choice reversals (R & F decoy: .10; RF decoy: .09) and competitor reversal selections (R & F decoy: .05 ; RF decoy: .04) are fairly low from Experiment 3.

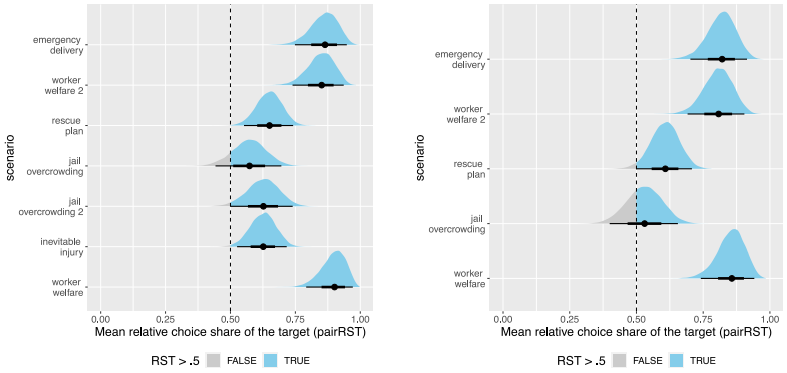
We show the pairRST by each item and scenario in Table 10.

#### 5.2.2. Bayesian statistical analysis

Data analysis was conducted in R (R Core Team, 2013) using brms (Bürkner, 2017).

*Statistical models, & priors: see Experiment 2.* We did not find any effect of whether subjects received instruction or not (SI). Thus, we collapsed the with-instruction and no-instruction group together.

We show the posterior 95% HDI of pairRST in Fig. 12. With *R & F* decoy, within-subject choice reversal occurred in five items: *emergency delivery* ($HDI_{.95}$ : (.76, .95), mean .86), *worker welfare* ($HDI_{.95}$ : (.80, .98), mean .90), *worker welfare 2* ($HDI_{.95}$ : (.75, .94), mean .85), *rescue plan* ($HDI_{.95}$ : (.55, .74), mean .65), *inevitable injury* ($HDI_{.95}$ : (.53, .72), mean .63). Items *jail overcrowding* ($HDI_{.95}$ : (.45, .70), mean .57) and *jail overcrowding 2* ($HDI_{.95}$ : (.50, .74), mean .63) had a slightly weaker effect. With *RF* decoy, a strong within-subject choice reversal occurred in three items: *emergency delivery* ($HDI_{.95}$ : (.72, .95), mean .82), *worker welfare* ($HDI_{.95}$ : (.75, .95), mean .85), *worker welfare 2* ($HDI_{.95}$ : (.70, .91), mean .81). The choice reversal effect was weaker in items *rescue plan* ($HDI_{.95}$ : (.50, .71), mean .61) and *jail overcrowding* ($HDI_{.95}$ : (.40, .66), mean .53). In addition, the full posteriors of model parameters are in Table 24, Appendix C.6. We found no difference between R & F and RF decoy (for decoy type parameter (baseline R & F), $CI_{.95}$ : (−0.1, 0.00), mean 0.00).




(a) *R & F* decoy (b) *RF* decoy

**Fig. 12.** Posteriors of pairRST in each scenario in Experiment 3.

**Table 10**
pairRST in Experiment 3.

| Scenario | Decoy | pairRST |
|---|---|---|
| Worker welfare | R & F | 0.93 |
| Inevitable injury | | 0.62 |
| Jail overcrowding 2 | | 0.61 |
| Jail overcrowding | | 0.54 |
| Rescue plan | | 0.70 |
| Worker welfare 2 | | 0.82 |
| Emergency delivery | | 0.89 |
| Worker welfare | RF | 0.88 |
| Jail overcrowding | | 0.52 |
| Rescue plan | | 0.51 |
| Worker welfare 2 | | 0.88 |
| Emergency delivery | | 0.85 |

**Table 11**
pairRST for shared items in Experiment 2 & 3.

| Scenario | Decoy | pairRST |
|---|---|---|
| Inevitable injury | R & F | 0.57 |
| Jail overcrowding | | 0.51 |
| Rescue plan | | 0.53 |
| Emergency delivery | | 0.54 |
| Jail overcrowding | RF | 0.53 |
| Rescue plan | | 0.52 |
| Emergency delivery | | 0.51 |

### *5.3. Combined analyses for Experiments 2 and 3*

To appropriately combine the items and compare results from Experiment 2 and Experiment 3, we present the following combined exploratory analyses. The results of choice proportions across all trials in identical dilemmas shared between Experiments 2 and 3, and the choice proportions across all trials in revised items in Experiments 2 and 3 can both be found in Appendix D.

Below we show the choice patterns (excluding consistent-selection rate) of the four identical items in aggregate (N = 931) below in Fig. 13, where the rates of choice reversals in R & F and RF decoy conditions are quite high. In addition, the pairRST for the four identical items with R& F and RF decoy conditions are in Table 11, suggesting slight choice reversals in all four items.

The estimated pairRST posterior 95% HDI for *R & F* decoy and *RF* decoy are shown in Fig. 14. With *R & F* decoy, we can observe an attraction effect in all four items: *emergency delivery* ($HDI_{.95}$ : (.62, .74), mean .73), *rescue plan* ($HDI_{.95}$ : (.58, .71), mean .65), *inevitable injury* ($HDI_{.95}$ : (.68, .78), mean .73), and *jail overcrowding* ($HDI_{.95}$ : (.60, .72), mean .67). With *RF* decoy, an attraction effect can be observed in three items: *emergency delivery* ($HDI_{.95}$ : (.57, .69), mean .63), *rescue plan* ($HDI_{.95}$ : (.53, .65), mean .59), and *jail overcrowding* ($HDI_{.95}$ : (.55, .68), mean .62).

We have also investigated how much time on average participants spent on a question during each session in Experiment 2 and Experiment 3 (Table 12). On average, participants spent less time on a question (at least for critical items) in Experiment 3 than either session in Experiment 2. In both experiments, participants spent less time when they see the same item for the second time.

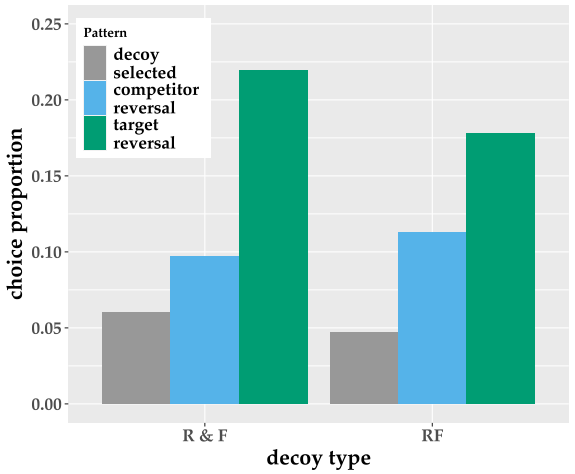



**Fig. 13.** Response patterns for shared items in Experiment 2 & 3.

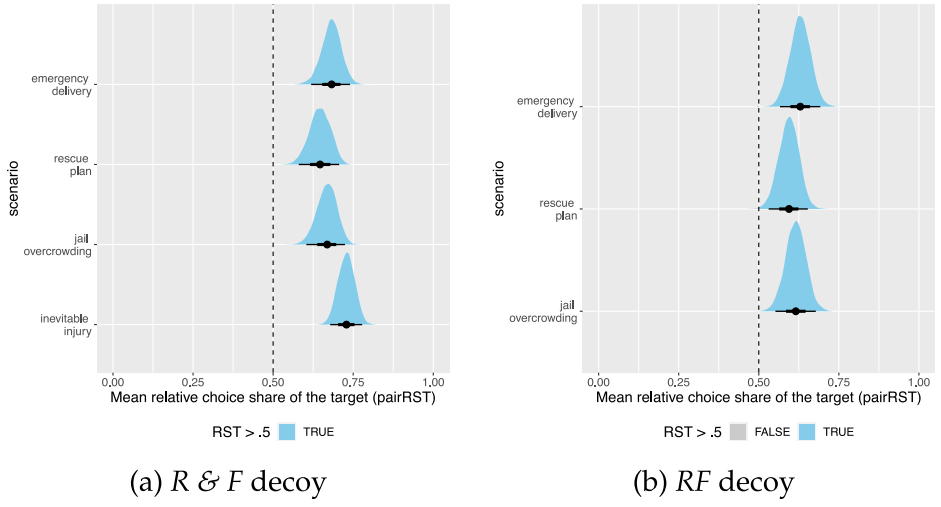



**Fig. 14.** Posteriors of pairRST in each scenario for shared items in Experiment 2 & 3.

**Table 12**
Average time spent per item.

| | | |
|---|---|---|
| Experiment 2 | Session 1 | 23.47 s |
| | Session 2 | 20.97 s |
| Experiment 3 | Critical items | 20.59 s |
| | First-seen critical | 24.55 s |
| | Second-seen critical | 16.58 s |

### 5.4. Discussion

In this section, we briefly discuss the results of Experiment 3. We will discuss the comparison between Experiment 2 and Experiment 3 in the conclusion section.

Experiment 3 builds upon Experiment 2 to investigate whether contextual choice reversals arise in ethical domains. We also explored whether we observe stronger contextual choice reversal effects by providing participants additional information about the potential existence of the dominating and dominated options in the tasks. We improved the materials from Experiment 2 by revising the items, improving our randomization for decoy types, and adding fillers as an attempt to make the critical items less distinctive and memorable.

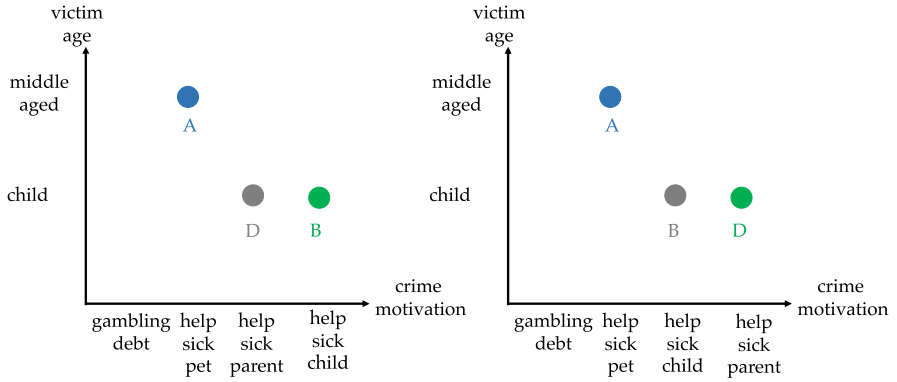



**Fig. 15.** Task structures for the *jail overcrowding 2* dilemma according to our assumed ranking for levels in the crime motivation attribute (left) and according to a different yet possible ranking for levels in the crime motivation attribute (right).

Similar to Experiment 2, we found evidence for contextual choice reversals in some, but not all items. While past work suggests that presenting non-numeric information as an attribute may reduce the choice reversals (Frederick et al., 2014), we did not observe the absence of choice reversals in dilemmas that only contain or partly contain non-numeric attribute levels consistently. Our conjecture is that the variations across scenarios may be due to individual differences in ranking of attribute levels (which we will discuss in detail in the next section). In addition, we also found very high consistent selection rates. This could be due to the fact that the participants did all pairs of questions within one session, making them more likely to remember their choices when they encountered the item for the first time in block 1. We also did not find any difference in the effects of R & F and RF decoys on contextual choice reversals. Nor did we find any increased or decreased contextual choice reversals rates when participants were given additional information about the relationship between the dominating option and the dominated option. While it is possible that the introduction did not leave a strong impression and was followed by a long set of decision making problems, we were not able to glean more information from this result.

We next explore how variation in reversal rates across items might be affected by individual participants' ranking of levels in the attributes.

## 6. Explaining variation in contextual choice reversals across ethical dilemmas

For attributes with qualitative rather than quantitative levels, different individuals may have different subjective rankings of the attribute levels. This poses challenges to the assumptions made for the task structure of the classical paradigm used to investigate choice reversals, which is that all decision makers have the same ranking of the features involved in the attributes. This assumption is not problematic for probability and value in economic gambles. But for the ethical dilemmas, this is not always the case. Consider the example below (illustrated in Fig. 15).

In a *jail overcrowding 2* dilemma, the decision maker needs to make a trade off between victim's age and the crime motivation for committing a robbery. The victim age attribute has two levels: "child" and "middle-aged", and the crime motivation attribute has four levels: "to pay off gambling debt", "to help a friend's sick pet", "to help a sick parent", and "to help a sick child". Our original construction of the task assumes the ranking "robbing middle-aged person" > (i.e., is more permissible than) "robbing a child" on the victim age attribute, and the ranking "help sick child" > "help sick parent" > "help friend's sick pet" > "pay off gambling debt" on the crime motivation attribute. This assumption is based on our data investigating people's ranking for levels in various ethically-involved attributes and yields three options: a competitor, A ("robbing a middle-aged person to help a friend's sick pet"), a target, B ("robbing a child to help a sick child"), and a decoy, D ("robbing a child to help a sick parent").

However, as our data on individuals' ranking on levels in various attributes only indicate a majority preference, it is possible that some participants may have a different ranking for the same levels. One different yet possible ranking on the crime motivation attribute could be "help sick parent" > "help sick child" > "help friend's sick pet" > "pay off gambling debt". If the participant has such a ranking, then our originally intended target, B ("robbing a child to help a sick child") becomes a decoy, and our originally intended decoy, D ("robbing a child to help a sick parent"), becomes the target for this participant. Thus, if this participant chooses the target option in their perspective, the choice would be reflected as a choice of decoy based on our original task structure. Thus, we believe that variation in choice reversal rates across scenarios could be systematically related to variations in people's ranking of the levels in the attributes.[5]

[5] Recent work by Bergner, Oppenheimer, and Detre (2019) has suggested that context effects such as preference choice reversals can be explained with models based on the ranking of options and principles of voting geometry. We use a simple generative model that is based on ranking to show how the individual differences in ranking systematically affect the variation in choice reversal rates. Our model is agnostic about the mechanistic or rational basis for the reversal, and so our model has a different aim than Bergner et al..

### 6.1. A generative model of choices in ethical dilemmas

The model makes three assumptions: (1) individuals choose consistently within a pair (regardless of which option dominates the decoy) most of time; (2) there is noise in individual decisions, which causes individuals to choose the inferior options occasionally; (3) the distribution of the decision makers' feature rankings is based on the results from our study where we discovered the attributes for constructing ethical dilemmas.

The model takes as input an individual's rankings for levels in all attributes and then simulate choices for all 16 ethical dilemmas from Experiment 3. Recall that the 16 dilemmas correspond to the 8 critical scenarios from Experiment 3 (Table 8). Each scenario includes two questions. The difference between these two questions is decoy position.

Given the individual's ranking for levels in an attribute, the dominating and dominated relationship among the three options in a question may be different from our initial assumptions when we constructed the questions. An individual's ranking of attribute levels induces dominance relations that may differ from the assumptions we made when constructing the items. Consequently, what is intended to be a target may become a competitor or even a decoy given some individual's rankings. Thus, given an input of individual's rankings for levels in all attributes, the model re-creates the structural relationships among the options in each dilemma — in other words, the model identifies the target, competitor, and decoy in each dilemma given the individual's rankings. Then, the model generates choices based on that individual's rankings and the re-created structures of the decision problems.

The possible structures given all possible rankings of levels in attributes are summarized in Table 13. Given the structures of dilemmas based on individual's rankings, the model generates choices for all ethical dilemmas in Experiment 3.

The model requires two parameters: an error rate, $\epsilon \in [0, 1]$, and a rate of choosing consistently, $\mathbf{p}_{\mathbf{consistent}} \in [0, 1]$. When a question has a best option given some individual's rankings, the model selects randomly between the other two non-dominating options with $p = \epsilon$. When a question has a worst option given some individual's rankings, the model selects the worst option with $p = \epsilon$. When either question in a pair does not have an *Attraction Configuration*, the model selects the same options in the questions in that pair with $\mathbf{p}_{\mathbf{consistent}}$.

After generating choices for all questions, for each question, we map the choice back to the option in the original *Attraction Configuration* of that same question. The original *Attraction Configuration* contains the target, the competitor, and the decoy that we initially constructed given our assumptions on the feature rankings.

We present the general algorithm for generating the decision problem's structure/configuration given a possible ranking and the method to generate choices given the problem structures in Appendix E.

### 6.2. Explaining variation in attraction effects across scenarios

We simulated 500 subjects in eight scenarios (with 1000 runs per subject). Each simulated subject had rankings for attributes that match a randomly sampled empirical subject in our study ($N = 57$) where we discovered the attributes for constructing ethical dilemmas. For each simulated subject, all 16 questions were generated in random order.

In the simulation, we set the decoy selection rate for questions with an *Attraction Configuration* as the same from our empirical data (mean decoy rate: .08). The error rate ($\epsilon$) was set to a low and reasonable value (0.05) and the probability of choosing consistently ($\mathbf{p}_{\mathbf{consistent}}$) was set to 0.7, as people choose consistently in about 70% trials with an *Attraction Configuration* empirically.

The simulation reproduced the pattern of choice patterns from the empirical data fully(SI). However, our simulation generally predicted consistent choice rates lower than empirical data and competitor reversal choice rates higher than empirical data (Fig. 29, Appendix E). We focus on the results of simulated and empirical decoy selection rates and choice reversal rates in Fig. 16.

We observe that the simulations of choice patterns based on individual ranking for levels in attributes predict choice reversals not perfectly but quite well. The model predicts reversal rates better in dilemmas with R & F decoys than in dilemmas with RF decoys. Besides reversal rates, the model also predicts decoy selection rates well for both dilemmas with R & F decoys and dilemmas with RF decoys. This suggests that some item variations can indeed be accounted for by the variations of people's individually varying rankings.

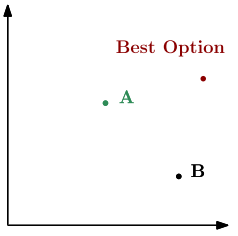
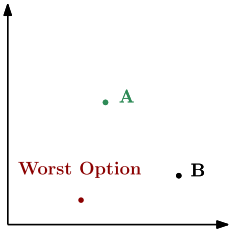
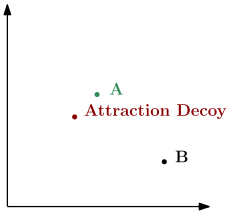
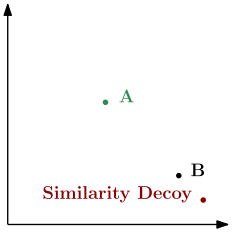
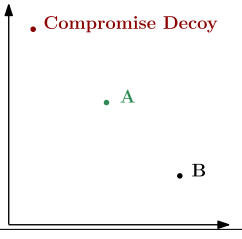
### 6.3. Discussion

Our model explores the implications of individual differences in attribute rankings.

To make our model complete, we have made a few assumptions such as the distribution of the feature rankings among decision makers, and the inclusion of error rate and consistent-choice rate. The distribution is based on our study with a somewhat small sample size, and the other two parameters are reasonable values. These decisions could potentially contribute to the reason why our model results only partly account for the observed choice reversal pattern. One possible future direction is to implement higher-level subjective utility functions that capture how different individuals make tradeoffs among options that involve qualitative attributes. This, of course, requires deeper investigations of subjective utility functions in value-based decisions. Another future direction is to test our model predictions further empirically, and investigate whether the manipulation of ranking of attribute levels would indeed change the configuration of the context effects and thus affect individual choices.

Finally, this generative model provides the foundation of a method to understand better context effects in domains with wide variation in attribute rankings, especially domains that involve qualitative attributes.

**Table 13**
Possible structures/configurations given all possible feature rankings. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

| Structure/Configuration | Illustrated example |
|---|---|
| *Best Option Configuration*: a best option dominates the other two options in both attributes |   |
| *Worst Option Configuration*: a worst option is dominated by the other two options in both attributes |   |
| *Attraction Configuration*: a target, competitor, and decoy can be clearly identified |   |
| *Similarity Configuration*: this configuration commonly produces a similarity effect |   |
| *Compromise Configuration*: this configuration commonly produces a compromise effect |   |

## 7. Conclusion

This work provides a clear and consistent evidence of within-participant contextual choice reversals in ethical domains, first with tasks isomorphic to economic gamble tasks, then with tasks spanning multiple ethical scenarios with qualitative attributes. By comparing choice reversals and performance — as measured by obtained expected value — between the economic (Experiment 1a) and isomorphic ethical decisions (Experiment 1b), we also found that performance in the ethical decisions was better than in the economic gambles, and that higher target reversal rates was associated with better performance in the economic gambles but not in the ethical decisions (we discuss this more below). In Experiment 2 and 3, we found choice reversals in ethical dilemmas across various domains, some of which involve qualitative attributes without the clear objective rankings that come with quantitative

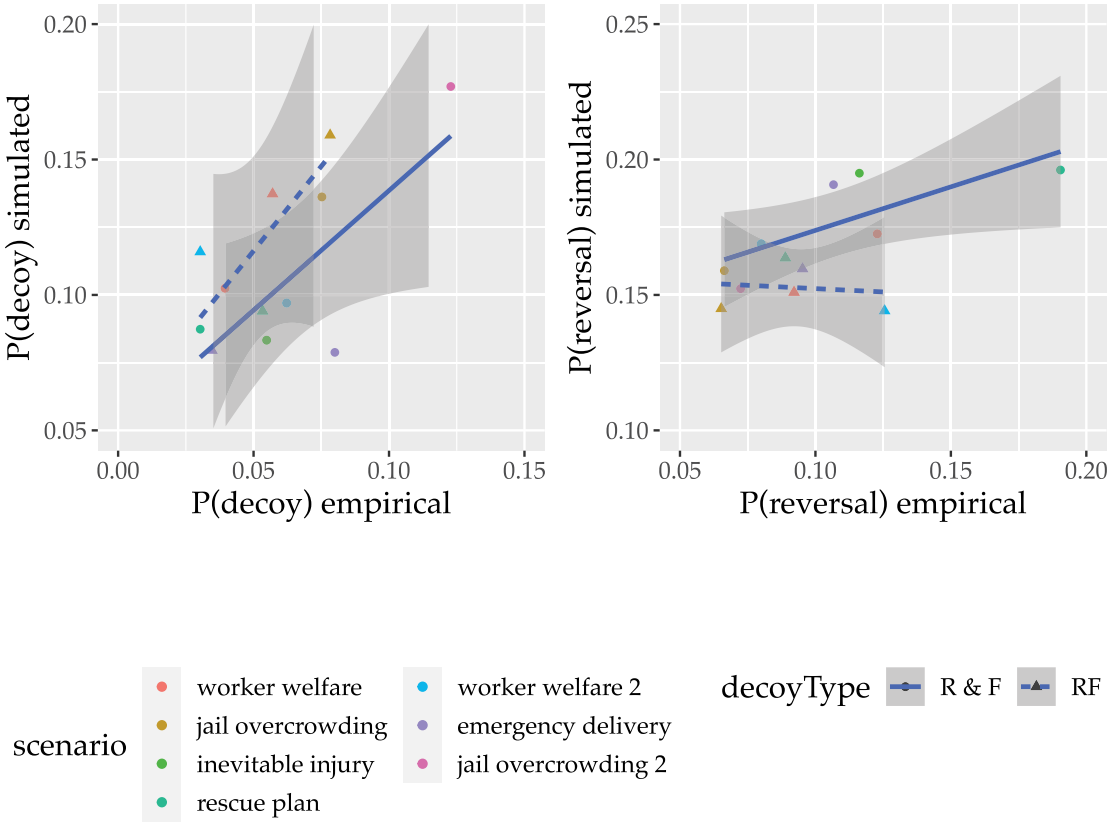



**Fig. 16.** Simulated and Experiment 3's empirical decoy selection rates (left) and choice reversal rates (right) for each scenario and decoy type.

attributes. In the further analyses comparing the results from Experiment 2 and 3, we observed smaller effects of choice reversals in Experiment 3 in aggregate, especially among the 4 identical items shared between the two experiments. This could largely be due to the fact that the participants in Experiment 3 completed the tasks in a single session and thus remembered the items well—an explanation consistent with the result that the rates of choosing consistently within a pair were extremely high. The time participants spent on these decisions may also provide some insight into the lowered choice reversal rates in Experiment 3, as the average time participants spent on critical items was shorter than that in Experiment 2, consistent with previous findings that choice reversals tend to diminish as time pressure increases (Pettibone, 2012).

Why did not choice reversals appear in all scenarios? This suggests a variation among items. We conjectured that at least some of this variation may be due to variation in individual subjective rankings of the attribute values in the scenarios — such variation could in effect change the location of the decoy so that it is no longer dominated by the intended target. We began to address this methodological challenge by developing a generative model that takes individual differences in feature rankings into account. We showed that this can indeed partly account for the variation of reversal rates among our scenarios. Another possibility is that, for decision problems that have non-numeric attributes, the space of attribute can be affected by individual weighting and individual risk preference. We hope to address this factor along with the individual variations in ranking of attribute levels in future work.

Although we found that both R & F and RF decoys can produce choice reversals. The effects of the two decoy types appear inconsistently across three experiments. We found that compared with RF decoy, R & F decoy had a stronger effect in Experiment 1, but not in Experiment 2 or 3. Such variations of different decoy types have been observed in past work as well (i.e., the effect sizes of the decoy types are not always consistent). However, as mentioned previously (see Section 3.4), such inconclusive results in our experiment could partly due to our experimental setup, where we did not distinguish further between R and F decoys, when both Huber and Puto (1983) and Trueblood et al. (2013) found that R decoy, had a larger effect than RF decoy, which had a larger effect than F decoy.

Our primary aim in this work is not to demonstrate that human ethical decision making is irrational. Rather, it is to establish further empirical bridges between ethical decision making and general domain-independent theories of decision making, such as the multi-attribute choice models that can account for the contextual attraction effects we observed (Bhatia, Loomes, & Read, 2021). Our own preferred theoretical account does not treat contextual choice reversals as evidence for irrationality but rather as evidence for boundedly rational (Simon, 1955) decision strategies (Howes et al., 2016)—more specifically, decision strategies that are well adapted to the cognitive and perceptual bounds of the decision maker. The Howes et al. (2016) model combines two

sets of noisy observations: noisy ordinal observations comparing attribute values across pairs of options, and noisy calculations of subjective expected value. The model chooses the option with the highest expected value given these noisy observations, and the result is systematic contextual choice reversals. But under this account, such choice reversals are a signature of boundedly rational utility maximization. This account of contextual choice reversals is consistent with our exploratory finding in Experiment 1a where higher target reversal rates predicted better task performance in the economic gambles — a counter-intuitive finding under the assumption that choice reversals are irrational. Although we did not find this relationship between target reversals and performance in isomorphic ethical decisions, we observed that the overall reversal rates in ethical decisions are just as high — in fact, higher — as those in choices in economic gambles.

As with any study of moral or ethical decision making involving descriptions of fictional scenarios, our study is limited by the fact that our participants understood that no lives or welfare were actually at stake, and the decisions were made based on descriptions rather than actual experience with the choice setting. But modern ethical choices, especially in cases of policy-making, are also often marked by decisions based on descriptions and by a considerable psychological (and sometimes geographic) distance between the decision maker and the people affected by the decisions. We therefore think it is likely that such real-world decisions will also be subject to the kind of context effects that we demonstrated here. For future work, we believe it is still worth investigating whether choice reversals will arise when we present the decision problem perceptually or through different modalities that can even increase the ecological validity further (i.e., like the studies of real-life commercial product decisions by Frederick et al., 2014).

Finally, we also caution against the view that the presence of such effects necessarily indicates irrational choice strategies; on the contrary, they may be signatures of adaptive, bounded rational decision making.

## CRediT authorship contribution statement

**Chenxu Hao:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Project administration. **Richard L. Lewis:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data and materials are available via links included in the article.

## Acknowledgments

We thank Tyler Adkins, Logan Bickel, Pyeong Whan Cho, Ian Cook, Hannah Foster, Nicole Hamilton, Steven Langsford, Sarah Marks, Connor McMann, Logan Walls for helpful discussions and feedback over the many years.

We would also like to thank the reviewers and the editor for their insightful suggestions.

This research was supported by University of Michigan College of Literature, Science and the Arts, and Templeton World Charity Foundation Grant #TWCF0205. Sponsors played no role in the conduct of the research or the preparation of the article or the decision to submit the article for publication.

## Appendix A. Background

See Table 14.

## Appendix B. Experiment 1

### *B.1. Wedell (1991) stimuli*

See Table 15.

### *B.2. Data structure and descriptive analysis*

The Table 16 shows an example of the data structure. Empirical data were organized as choice patterns for all subjects.

The Table 17 shows the proportions of choice reversals in Experiment 1a and 1b, calculated from total number of pairs that exhibit a choice reversal (defined as subject choosing targets for this pair for both decoy positions) over total number of pairs done by all subjects for each decoy type.

**Table 14**
Summary of inconsistent behaviors and their underlying moral heuristics.

| Behaviors | Heuristics/Effect | Type |
|---|---|---|
| Preference over government's policy to save more lives or more life-years depends on the framing when the expected utility is the same (Sunstein, 2004). | Framing effect | Substitution |
| People are willing to punish companies' ethical decisions that are based on cost–benefit analysis when the companies' liability is unclear under the law (Viscusi, 2000) | Rejecting cost–benefit analysis in decisions affecting lives | Rules-of-thumb |
| Objections to emission trading led to the delay and reduction of the use of a pollution reduction tool that is, in many contexts, the best available (Sunstein, 2002). | Do not allow moral wrongdoing for a fee | Rules-of-thumb |
| People are averse to risks of death from products that are designed to promote safety, e.g., airbags (Koehler & Gershoff, 2003). | Betrayal risk aversion | Substitution |
| Punishment judgments towards corporations are a product of outrage and leads to decreased wages, increased prices, lost jobs (Kahneman, Schkade, & Sunstein, 1998) or less beneficial products such as vaccines and birth control pills on the market (Baron & Ritov, 1993). | Outrage heuristic (Kahneman & Frederick, 2002) | Substitution |
| People overestimate the carcinogenic risk from pesticides and underestimate the risks of natural carcinogens (Rozin, 2001). | Do not tamper with nature | Rules-of-thumb |
| Harmful acts are generally seen worse than harmful omissions (Baron & Ritov, 2004; Rodriguez-Arias, Rodriguez Lopez, Monasterio-Astobiza, & Hannikainen, 2020). | Omission bias | Substitution |
| Most U.S. citizens say that they approve of postmortem organ donation, yet relatively few sign a donor card (Johnson & Goldstein, 2003). | If there is a default, do nothing | Substitution |
| Do what the majority of one's peers do (Gigerenzer, 2010) | Imitate your peers | Substitution |

**Table 15**
Gambles used in Wedell (1991)'s original Experiment 1. R = range decoy; F = frequency decoy; RF = range-frequency decoy (Wedell, 1991).

| Target bets | Decoy bets | | |
|---|---|---|---|
| | R | F | RF |
| Target *A* | | | |
| .40, $25 | .40, $20 | .35, $25 | .35, $20 |
| .50, $20 | .50, $18 | .45, $20 | .45, $18 |
| .67, $15 | .67, $13 | .62, $15 | .62, $13 |
| .83, $12 | .83, $10 | .78, $12 | .78, $10 |
| Target *B* | | | |
| .30, $33 | .25, $33 | .30, $30 | .25, $30 |
| .40, $25 | .35, $25 | .40, $20 | .35, $20 |
| .50, $20 | .45, $20 | .50, $18 | .45, $18 |
| .67, $15 | .62, $15 | .67, $13 | .62, $13 |

**Table 16**
An example of the data coded as choice patterns for J subjects.

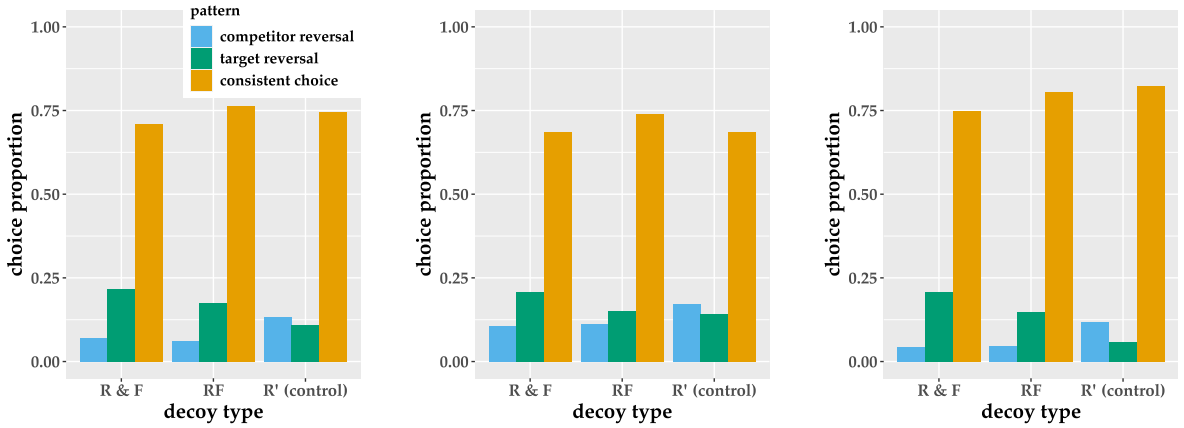
| Subject | Pair | Decoy type | Pair choice pattern |
|---|---|---|---|
| 1 | 1 | R & F | Consistent choice |
| ... | ... | ... | ... |
| 1 | 21 | RF | Target reversal |
| 2 | 11 | R & F | Target reversal |
| ... | ... | ... | ... |
| 2 | 31 | R & F | Decoy selected |
| ... | ... | ... | ... |
| J | 11 | R & F | Target reversal |
| ... | ... | ... | ... |
| J | 21 | RF | Competitor reversal |

### *B.3. Full description results*

See Fig. 17.

### *B.4. Full statistical results*

See Tables 18–21.




(a) Wedell (1991) (b) Wedell (1991) replication (N=150) (c) Ethical decisions (N=51)

**Fig. 17.** Descriptive results — full response patterns in: (a). Wedell (1991); (b). Wedell (1991) replication; (c). Ethical decisions.

**Table 17**
Proportions of within subject choice reversals (PR) occurrences in Experiment 1a and 1b.

| Decoy type | Subjects | Total pairs done | # of PR | Proportions of PR |
|---|---|---|---|---|
| Experiment 1a | | | | |
| R | 142 | 382 | 82 | 0.21 |
| F | 137 | 367 | 83 | 0.23 |
| RF | 142 | 371 | 63 | 0.17 |
| R′ (control) | 145 | 380 | 63 | 0.17 |
| Experiment 1b | | | | |
| R | 47 | 129 | 14 | 0.11 |
| F | 49 | 150 | 20 | 0.13 |
| RF | 47 | 112 | 16 | 0.14 |
| R′ (control) | 49 | 109 | 19 | 0.16 |

**Table 18**
Full model posteriors of experiment 1a and 1b.

| Parameter | Exp | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|---|
| b_Intercept | 1a | 1.00 | 2393 | 0.57 | 0.23 | 0.11 | 0.57 | 1.07 |
| sd_decoy_Intercept | 1a | 1.00 | 1478 | 0.32 | 0.38 | 0.05 | 0.19 | 1.38 |
| r_decoy[R&F,] | 1a | 1.00 | 2418 | 0.09 | 0.23 | −0.41 | 0.08 | 0.55 |
| r_decoy[R_prime,] | 1a | 1.00 | 2449 | −0.10 | 0.23 | −0.62 | −0.09 | 0.34 |
| r_decoy[RF,] | 1a | 1.00 | 2420 | 0.00 | 0.23 | −0.49 | 0.00 | 0.46 |
| b_Intercept | 1b | 1.00 | 1977 | 0.65 | 0.47 | −0.31 | 0.65 | 1.57 |
| sd_decoy_Intercept | 1b | 1.00 | 1466 | 0.63 | 0.65 | 0.12 | 0.42 | 2.44 |
| r_decoy[R&F,] | 1b | 1.00 | 1981 | 0.17 | 0.47 | −0.76 | 0.17 | 1.13 |
| r_decoy[R_prime,] | 1b | 1.00 | 2023 | −0.27 | 0.48 | −1.21 | −0.25 | 0.66 |
| r_decoy[RF,] | 1b | 1.00 | 1949 | 0.08 | 0.47 | −0.84 | 0.08 | 1.04 |

**Table 19**
Experiment 1 — posterior statistics for parameters in the model comparing performance in ethical decisions (reference group) and economic gambles.

| Parameter | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|
| $\beta^0_{performance}$(Intercept) | 1.00 | 12 471 | 9.96 | 0.03 | 9.91 | 9.96 | 10.02 |
| $\beta^1_{performance}$ | 1.00 | 11 590 | −0.13 | 0.03 | −0.19 | −0.13 | −0.07 |
| sigma | 1.00 | 11 455 | 0.20 | 0.01 | 0.18 | 0.20 | 0.22 |
| mean_PPD | 1.00 | 14 840 | 9.87 | 0.02 | 9.83 | 9.87 | 9.90 |
| log-posterior | 1.00 | 7 418 | 38.07 | 1.20 | 35.00 | 38.38 | 39.46 |

**Table 20**
Experiment 1 — posterior statistics for parameters in the model comparing target reversal rates in ethical decisions (reference group) and economic gambles.

| Parameter | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|
| $\beta^0_{targetreversal}$(Intercept) | 1.00 | 12 055 | 0.20 | 0.02 | 0.15 | 0.20 | 0.24 |
| $\beta^1_{targetreversal}$ | 1.00 | 11 214 | −0.04 | 0.03 | −0.10 | −0.04 | 0.01 |
| sigma | 1.00 | 12 505 | 0.17 | 0.01 | 0.15 | 0.17 | 0.19 |
| mean_PPD | 1.00 | 14 261 | 0.16 | 0.02 | 0.13 | 0.16 | 0.20 |
| log-posterior | 1.00 | 6 708 | 69.24 | 1.25 | 66.01 | 69.56 | 70.67 |

**Table 21**
Experiment 1 — posterior statistics for parameters in the regression model in (3.3). The reference group is Wedell (1991) replication (i.e., economic gambles).

| Parameter | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|
| $\beta^0$ (Intercept) | 1.00 | 17 139 | 9.78 | 0.02 | 9.73 | 9.78 | 9.82 |
| $\beta^1$ | 1.00 | 11 098 | 0.60 | 0.17 | 0.26 | 0.60 | 0.93 |
| $\beta^2$ | 1.00 | 17 954 | 0.20 | 0.05 | 0.10 | 0.19 | 0.29 |
| $\beta^3$ (interaction) | 1.00 | 12 372 | −0.63 | 0.32 | −1.26 | −0.63 | 0.01 |
| mean_PPD | 1.00 | 17 054 | 9.83 | 0.03 | 9.77 | 9.82 | 9.88 |
| log-posterior | 1.00 | 6 523 | −42.62 | 1.61 | −46.59 | −42.29 | −40.49 |
| $\sigma$ | 1.00 | 16 642 | 0.29 | 0.01 | 0.27 | 0.29 | 0.32 |

## Appendix C. Experiment 2 & 3

### *C.1. Example of questions for finding attributes to construct materials for Experiment 2 & 3*

Below we show a set of questions that are presented to participants for determining how participants' rank the levels in an attribute. The attribute is crime motivation and the levels are: stealing prescription drugs for a sick child, stealing prescription drugs for a sick parent, stealing prescription drugs for a friend's sick pet, and stealing prescription drugs to pay off gambling debt.

1. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?
    - A man who stole prescription drugs for his sick child.
    - A man who stole prescription drugs for his friend's sick pet.
2. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?
    - A man who stole prescription drugs for his sick child.
    - A man who stole prescription drugs for his sick parent.
3. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?
    - A man who stole prescription drugs for his sick parent.
    - A man who stole prescription drugs for his friend's sick pet.
4. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?
    - A man who stole prescription drugs for his sick child.
    - A man who stole prescription drugs to pay off his gambling debt.
5. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?
    - A man who stole prescription drugs for his sick parent.
    - A man who stole prescription drugs to pay off his gambling debt.
6. You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release?

**Table 22**
Items (scenarios) appeared in Experiment 2 and Experiment 3. The *rescue a survivor* item only appeared in Experiment 2.

| Item | Scenario |
|---|---|
| Emergency delivery | You are responsible for an emergency delivery of medical supplies to a small village to prevent some serious illness. There are different vehicles that you may choose from. They all cost the same but vary in the delivery speed and the amount of pollutants produced. Which of the following vehicles do you choose? |
| Jail overcrowding | You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release? (The prisoners' crime motivation and recidivism vary.) |
| Jail overcrowding 2 | You are the sheriff of the jail in a small town. The jail is overcrowded and you have to release a prisoner. Which prisoner would you release? (The prisoners' crime motivation and victim's ages vary.) |
| Inevitable injury | You work for a shipping company and your job is to monitor autonomous cars and control them in the case of an emergency. One day when you are working, one of the autonomous cars experiences a sudden brake failure. The car is approaching a busy intersection where there are pedestrians crossing the street. If you do nothing, the car will hit the nearby vehicle, killing all passengers inside the car and the nearby vehicle. By taking control of the car, you can navigate it to crash into one of the pedestrians crossing the street, but doing so may result in the injury of the pedestrians. Which of the following outcomes would you choose? |
| Rescue plan | A hurricane hits a small town causing most houses to be destroyed. Three emergency rescue plans have been proposed. Assuming that the exact scientific estimates of the consequences of the plans are as follows, which plan would you choose? |
| Rescue a survivor[a] | A devastating hurricane that destroys most homes hits a small island. You are the lead expert on the emergency rescue team and you find three severely injured survivors buried underneath the rubble. A member of your team has evaluated the likelihood of successfully rescuing each survivor. After carefully examining the situation, you realize that this confined space is very fragile and you can try to rescue only one person before it collapses. The survivors you do not try to rescue will certainly die. Who would you try to rescue? (Likelihood of rescuing and age of survivors vary.) |
| Firing an employee | You are the manager of a small group of people in a company. Due to low sales, you have to fire an employee. Who would you fire? (The employees' years of working experiences and their sense of responsibility vary.) |
| Worker welfare | You are buying a laptop that is produced by different companies. Assuming that the products all have the same quality, which of the following companies would you choose to buy it from? (Price and how well the companies pay their workers vary.) |
| Worker welfare 2 | You are buying a pair of boots that is produced by different companies. Assuming that the products all have the same quality, which of the following companies would you choose to buy it from? (Price and how well the companies pay their workers vary.) |

[a] The *rescue a survivor* item only appeared in Experiment 2.

- A man who stole prescription drugs for his friend's sick pet.
- A man who stole prescription drugs to pay off his gambling debt.

### C.2. Scenarios in Experiment 2 & 3

See Table 22.

### C.3. Ethical content analysis of the scenarios

Recall that we follow the definition in Yu et al. (2019) that ethical decisions are decisions that affect others' welfare (Yu et al., 2019). In this section, we provide a content analysis of the scenario by explaining how each scenario poses a dilemma in which the welfare of different parties is at stake.

1. *emergency delivery*: In this scenario, the decision maker needs to select a vehicle to complete an emergency drug delivery to a remote village while making a trade-off between the speed of the vehicle and the amount of pollutant that the vehicle produces. The speed of the vehicle directly affects the welfare of the villagers who are in need of the emergency medication, however, the faster vehicle also produces more pollutants, which would further threat the environment, posing a long-term threat to all people.
2. *jail overcrowding*: In this scenario, the decision maker needs to decide which prisoner to release due to overcrowding issue in a small town. The trade-off is between deciding based on the original motivation of the crime and the probability of the prisoner to recommit the same crime after being release. This poses a dilemma because even though it is more permissible for someone to commit robbery to buy drugs for their sick child (compared to the motivation of paying off gambling debt), this act is associated with higher probability of recommitting the same crime — which damages the welfare of the robbery victim. Essentially, the welfare of the released prisoner or their family is pitted against the welfare of potential victims and the society in general.

3. *jail overcrowding 2*: In this scenario, the decision maker needs to decide which prisoner to release due to overcrowding issue in a small town. The trade-off is between deciding based on the original motivation of the crime and deciding based on the age of the victim. This poses a dilemma because even though it could be permissible for someone to commit robbery to buy drugs for their sick child, it could be less permissible to rob an old person or a child at the same time. Essentially, the welfare of the released prisoner or their family is pitted against the welfare of potential victims.
4. *inevitable injury*: In this scenario, the decision maker takes over the automatic car that is lost control and must make a decision of which pedestrian to run into — otherwise all passengers in the car dies. Running into different pedestrians cause different injury to them with different probabilities — and weaker injury is associated with higher probabilities. This decision directly affects the welfare of the pedestrians involved.
5. *rescue plan*: This scenario is taken from Experiment 1, where the decision maker must decide on a rescue plan after a hurricane — each plan leads to saving different numbers of people, but saving more people is associated with lower probability of a successful rescue. In this decision, the survival of the few people is pitted against the survival of many people.
6. *rescue a survivor*: In this scenario, the decision maker also needs to decide on who to rescue after a hurricane — but this scenario focuses on the welfare of single survivors. The younger survivor is less likely to be successfully rescued. Here, the survival of individuals are pitted against the survival of each other.
7. *firing an employee*: In this scenario, the decision maker needs to decide which employee in a company to fire due to low sales. The employees involved in the decisions have worked at the company for different numbers of years (i.e., some employees have more experience), but the ones who have worked at the company for longer may have less sense of responsibility (e.g., they may often miss work or come to work late). This decision also directly impacts the welfare of the involved employees.
8. *worker welfare* and *worker welfare 2*: In these two scenarios, the decision maker decides on which product to buy. The scenarios involve different types of products. However, in both scenarios, some products are cheaper, but they may be produced by companies that do not treat their employees well (or use child labor); some products are more expensive, but they are produced by companies that pay their employees well and provide health benefits. In these scenarios, the welfare of the decision maker is directly involved — and it is pitted against the welfare of the companies' employees.

### *C.4. An example of a set of questions in part 1 of Experiment 2 and Experiment 3*

Below is a set of questions corresponding to the item speed-pollution. In this set, pollution attribute says constant and speed attribute varies in each choice. Each participant was randomly presented one out of the four questions and three out of the four choices within in the question.

1. You are responsible for an emergency delivery of medical supplies to a small village to prevent some serious illness. There are different vehicles that you may choose from. They all cost the same and produce the same amount of pollutants but vary in the delivery speed (overnight, 3 days, 5 days, 7 days). Which of the following vehicles do you choose?

    - A car that produces a low amount of pollutants and makes the delivery overnight.
    - A car that produces a low amount of pollutants and makes the delivery in 3 days.
    - A car that produces a low amount of pollutants and makes the delivery in 5 days.
    - A car that produces a low amount of pollutants and makes the delivery in 7 days.

2. You are responsible for an emergency delivery of medical supplies to a small village to prevent some serious illness. There are different vehicles that you may choose from. They all cost the same and produce the same amount of pollutants but vary in the delivery speed (overnight, 3 days, 5 days, 7 days). Which of the following vehicles do you choose?

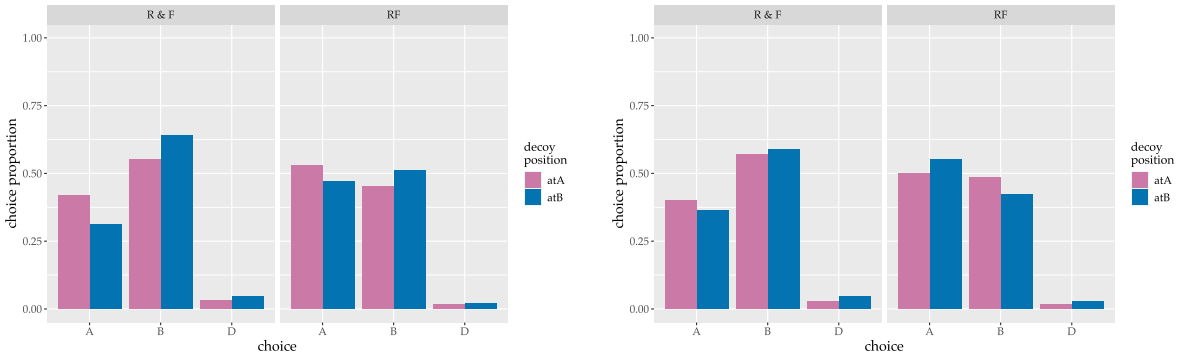
    - A car that produces a medium amount of pollutants and makes the delivery overnight.
    - A car that produces a medium amount of pollutants and makes the delivery in 3 days.
    - A car that produces a medium amount of pollutants and makes the delivery in 5 days.
    - A car that produces a medium amount of pollutants and makes the delivery in 7 days.

3. You are responsible for an emergency delivery of medical supplies to a small village to prevent some serious illness. There are different vehicles that you may choose from. They all cost the same and produce the same amount of pollutants but vary in the delivery speed (overnight, 3 days, 5 days, 7 days). Which of the following vehicles do you choose?

    - A car that produces a high amount of pollutants and makes the delivery overnight.
    - A car that produces a high amount of pollutants and makes the delivery in 3 days.
    - A car that produces a high amount of pollutants and makes the delivery in 5 days.
    - A car that produces a high amount of pollutants and makes the delivery in 7 days.

4. You are responsible for an emergency delivery of medical supplies to a small village to prevent some serious illness. There are different vehicles that you may choose from. They all cost the same and produce the same amount of pollutants but vary in the delivery speed (overnight, 3 days, 5 days, 7 days). Which of the following vehicles do you choose?

    - A car that produces a very high amount of pollutants and makes the delivery overnight.




(a). Attraction effect from data in the first session

(b). No attraction effect from data in the second session

**Fig. 18.** Aggregated choice proportions during the first and second session in Experiment 2 (N = 475). We observe a clear attraction effect across subjects in the first session, but the effect is not present in the second session.

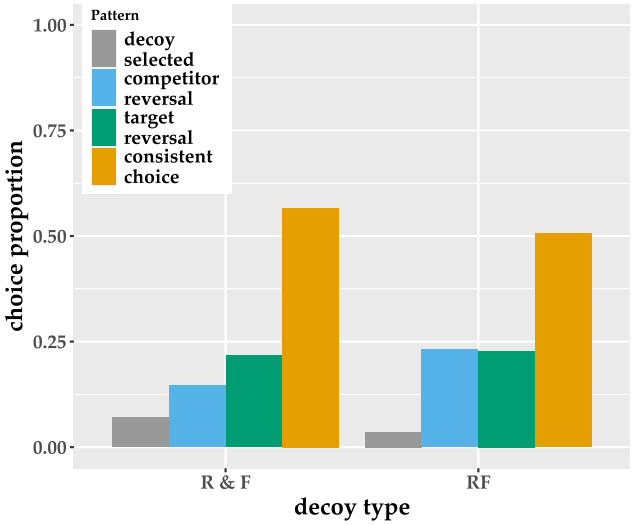



**Fig. 19.** Full response patterns for data aggregated over all items in Experiment 2 (N = 475).

- A car that produces a very high amount of pollutants and makes the delivery in 3 days.
- A car that produces a very high amount of pollutants and makes the delivery in 5 days.
- A car that produces a very high amount of pollutants and makes the delivery in 7 days.

### *C.5. Experiment 2 results*

*Choice proportions across trials.* During the first session, participants saw all the scenarios for the first time. The choice proportions in the first occurrences of the scenarios allow us to explore the attraction effect exhibited in the data. If we see that the target is preferred over the competitor under the presence of the decoy, then we observe an attraction effect across participants. In this experiment, we observe an attraction effect at the first occurrences of scenarios (Fig. 18). However, the effect is not present in the second session, potentially due to memories of the scenarios from the first session, considering all scenarios are fairly distinctive.

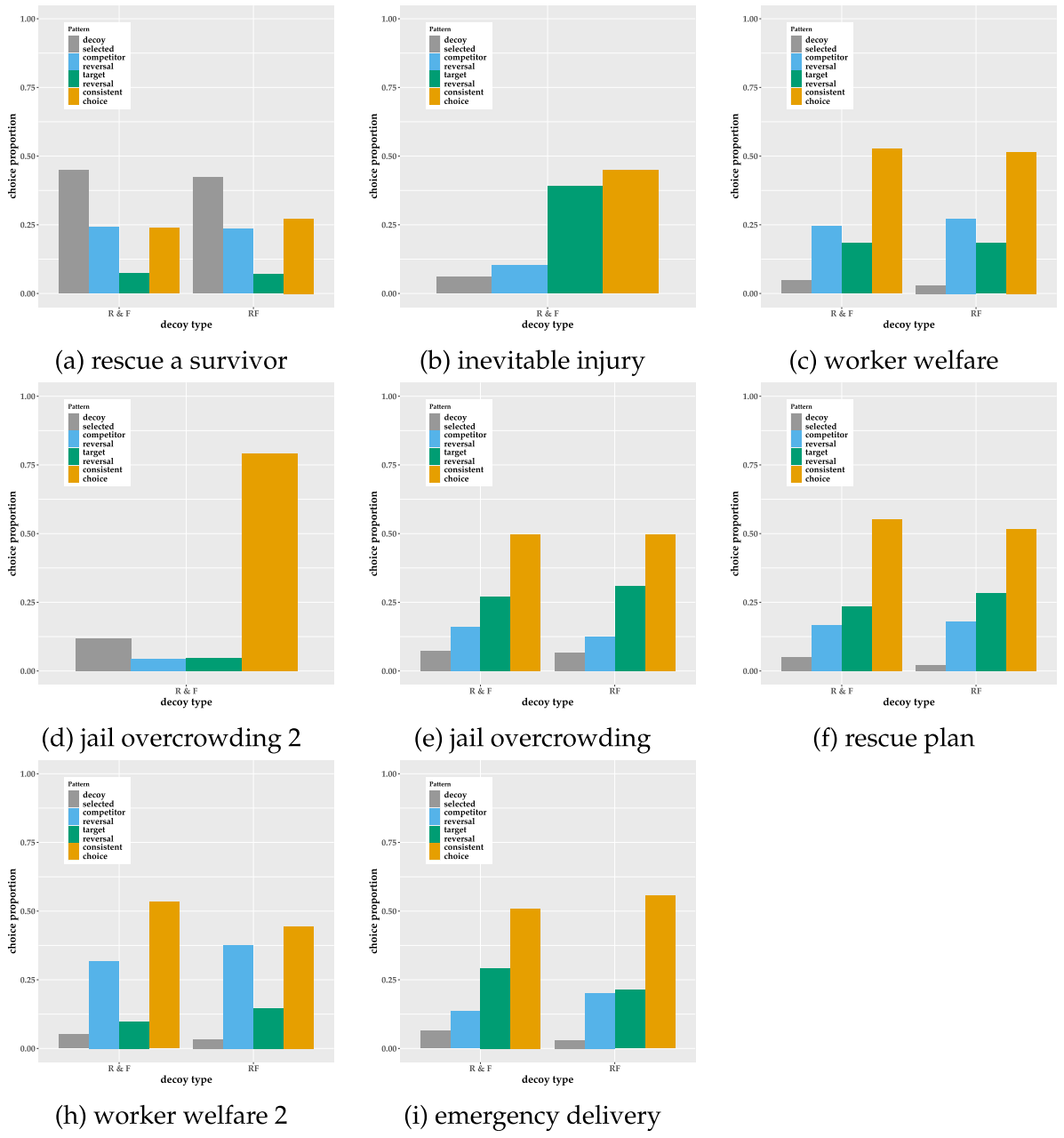



**Fig. 20.** Choice patterns for all 8 ethical dilemmas.

### *C.5.1. Descriptive analysis by item*

Here we present the descriptive results of choice patterns and choice proportions by item (see Fig. 20 and Fig. 21).

### *C.5.2. Full model results*

See Table 23.

## *C.6. Experiment 3 results*

*Choice proportions across trials.* We report below the attraction effect across all trials (Fig. 22), i.e., the choice proportions for each option at each decoy position across all trials and all scenarios. Although participants completed the pairs of items in one session, they still saw the different versions of items (decoy-at-A and decoy-at-B for critical items) in two separate blocks. Similar to Experiment

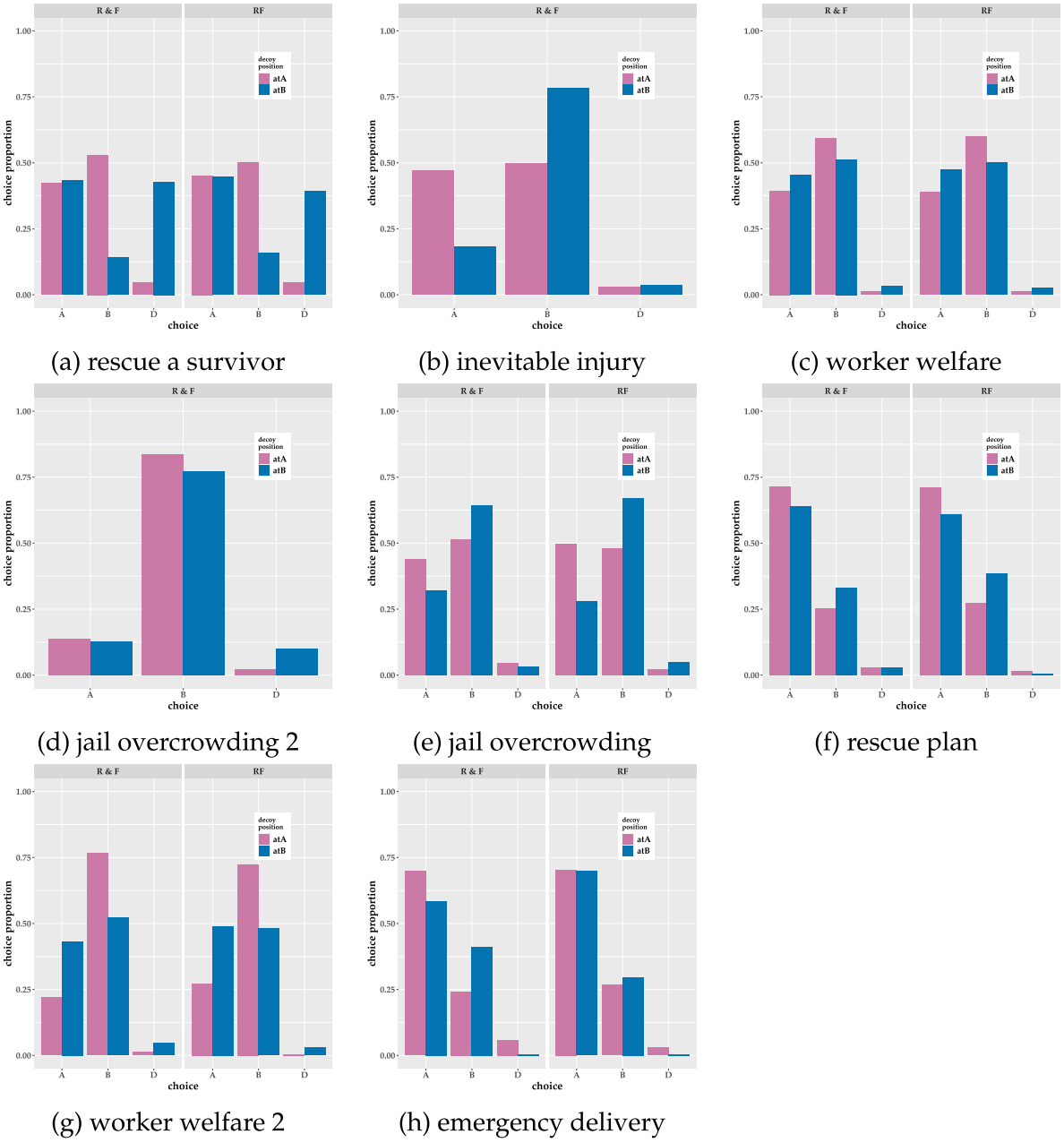



**Fig. 21.** Choice proportions across trials for each ethical dilemma.

2, we observe an attraction effect across subjects at the first occurrences of scenarios. The attraction effect is not present in the second block, potentially due to memories of the scenarios from the first block.

### *C.6.1. Descriptive analysis by item*

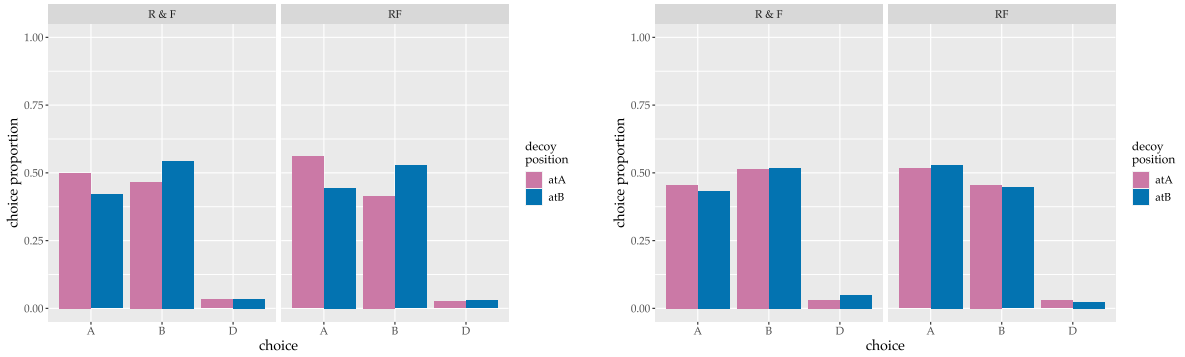
See Figs. 24 and 25.

### *C.6.2. Full model results*

See Table 24.

**Table 23**
Full posteriors of model parameters in Experiment 2.

| Parameter | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|
| b_Intercept | 1.00 | 2896 | 0.55 | 0.09 | 0.37 | 0.55 | 0.74 |
| b_decoy_type_baseline (R&F) | 1.00 | 8591 | −0.00 | 0.03 | −0.06 | −0.01 | 0.05 |
| sd_scenario_Intercept | 1.00 | 2434 | 0.22 | 0.09 | 0.11 | 0.20 | 0.46 |
| r_scenario[worker welfare,] | 1.00 | 3105 | −0.13 | 0.09 | −0.32 | −0.13 | 0.05 |
| r_scenario[inevitable injury,] | 1.00 | 3087 | 0.23 | 0.09 | 0.05 | 0.23 | 0.42 |
| r_scenario[jail overcrowding 2,] | 1.00 | 3737 | −0.02 | 0.11 | −0.23 | −0.02 | 0.19 |
| r_scenario[jail overcrowding,] | 1.00 | 3054 | 0.11 | 0.09 | −0.07 | 0.11 | 0.30 |
| r_scenario[rescue plan,] | 1.00 | 3085 | 0.05 | 0.10 | −0.14 | 0.05 | 0.24 |
| r_scenario[worker welfare,] | 1.00 | 3065 | −0.28 | 0.09 | −0.47 | −0.28 | −0.09 |
| r_scenario[emergency delivery,] | 1.00 | 3113 | 0.05 | 0.09 | −0.14 | 0.05 | 0.23 |




(a). Attraction effect from data in the first block

(b). No attraction effect from data in the second block

**Fig. 22.** Aggregated choice proportions during the first and second blocks in Experiment 3 (N = 456). We observe a clear attraction effect across subjects in the first block, but not in the second block.

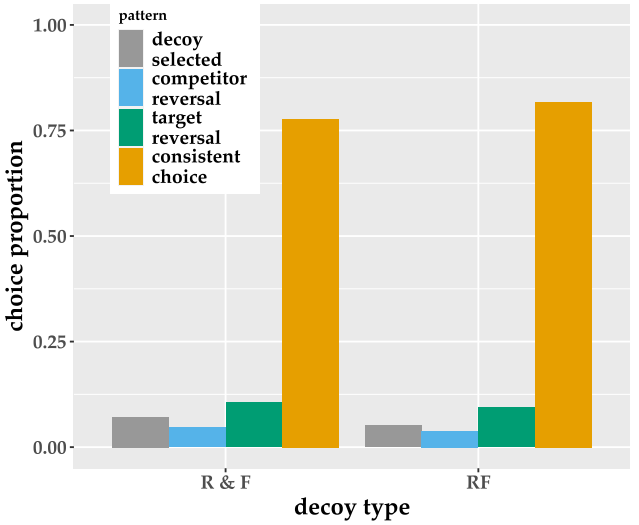



**Fig. 23.** Aggregated response patterns for all items (excluding *firing an employee*) in Experiment 3 (N = 456).

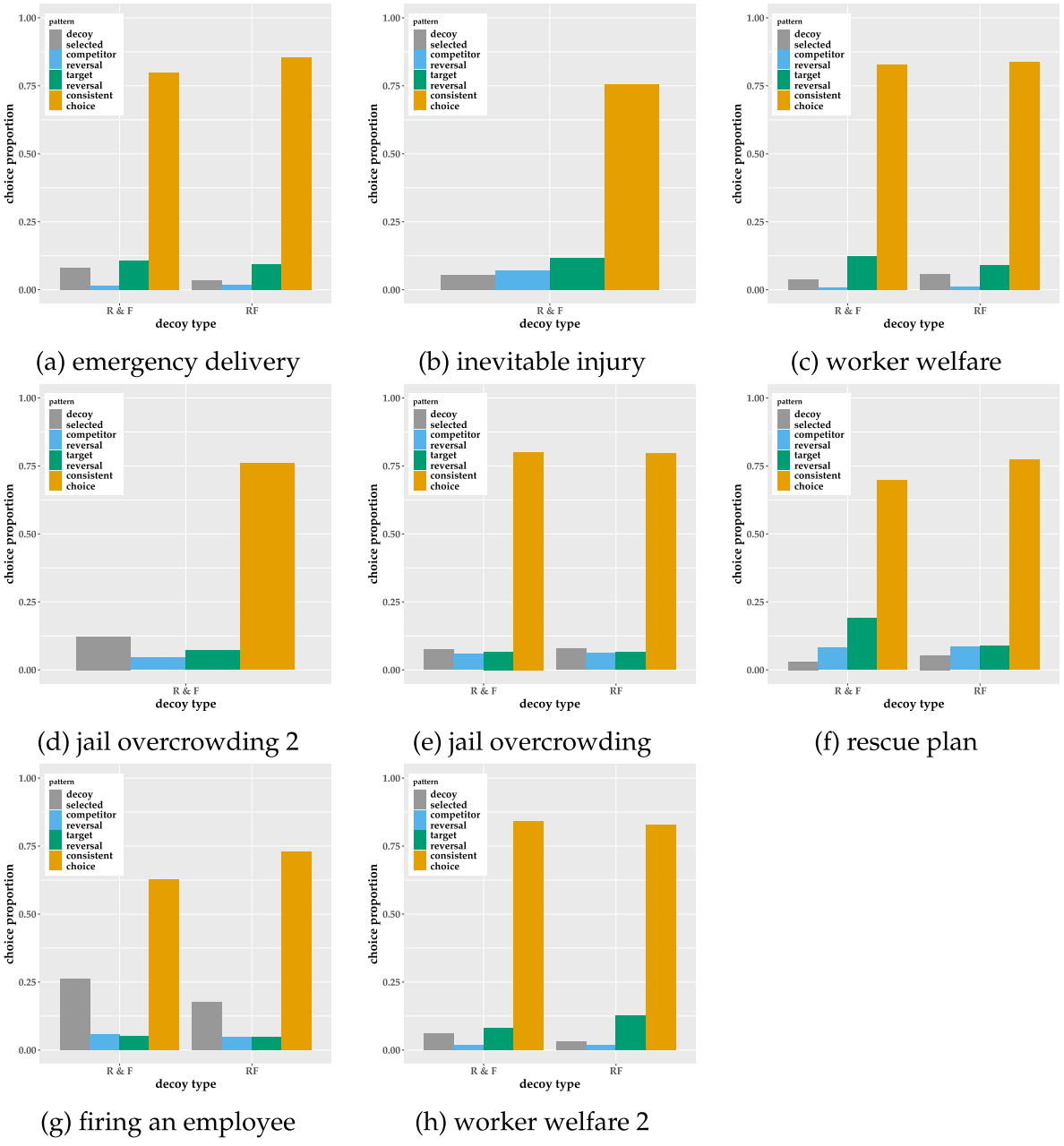



**Fig. 24.** Choice patterns for all 8 ethical dilemmas.

## Appendix D. Combined analyses for Experiments 2 and 3

*Choice proportions across trials in identical dilemmas.* Fig. 26 contains the aggregated choice proportions for the first and second occurrences of the scenarios that are *identical* in experiment 2 & 3 (*emergency delivery, jail overcrowding, inevitable injury, rescue plan*; the choice proportions for each scenario are provided in SI). Experiment 2 shows a stronger attraction effect across subjects among these items compared to Experiment 3 in both first and second occurrences. During the second occurrence, the effect is not present in Experiment 3.

*Choice proportions across trials in revised dilemmas.* Below we include the results of attraction effects across subjects for the revised scenarios (*worker welfare 2, worker welfare, jail overcrowding 2*, Fig. 27). We observe an attraction effect in Experiment 3 after the revision of items, especially in the first occurrences of items.

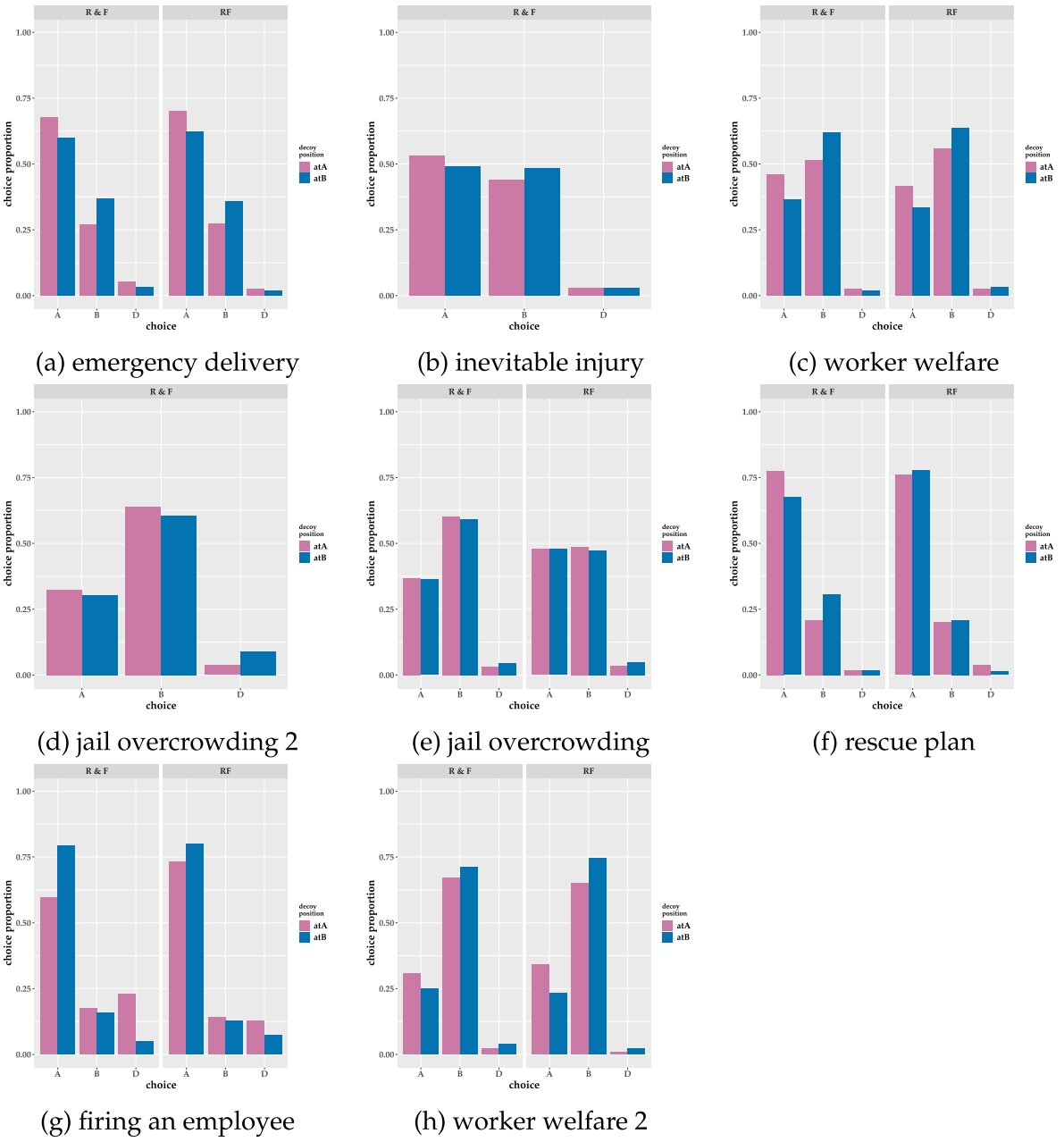



**Fig. 25.** Choice proportions for each ethical dilemma in Experiment 3.

Aggregated response patterns for identical items in Experiment 2 and 3 combined data are shown in Fig. 28.

## Appendix E. Explaining variation in contextual choice reversals across ethical dilemmas

### *E.1. Algorithm for generating decision problems' structures given feature rankings*

Here we describe a general algorithm to generate the decision problem's structure given any possible ranking.

When we constructed decision problems in Experiment 3, we constructed options ($A$, $B$, $D$) whose rankings of levels in each attribute were our assumed rankings. However, in reality, decision makers do not have the same rankings of levels in each attribute.

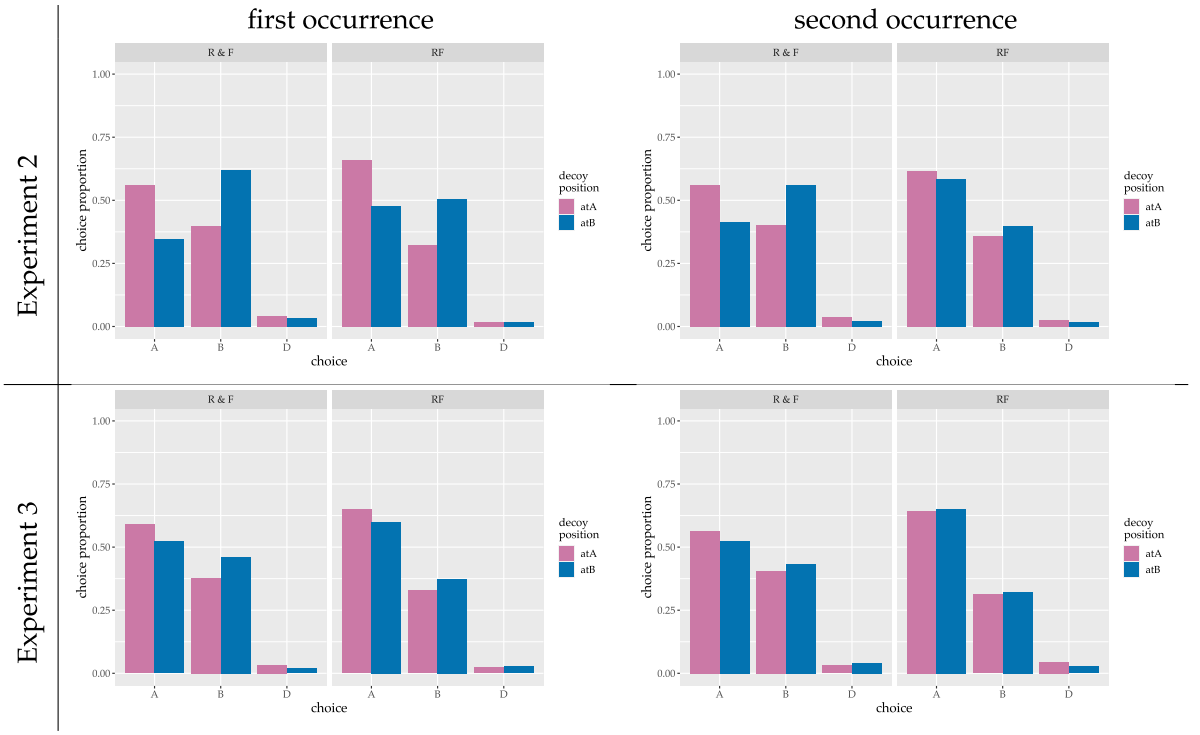



**Fig. 26.** Attraction effect across subjects shown as aggregated choice proportions for first and second occurrences of identical items (*emergency delivery, jail overcrowding, inevitable injury, rescue plan*) in Experiment 2 (N = 475) and 3 (N = 456).

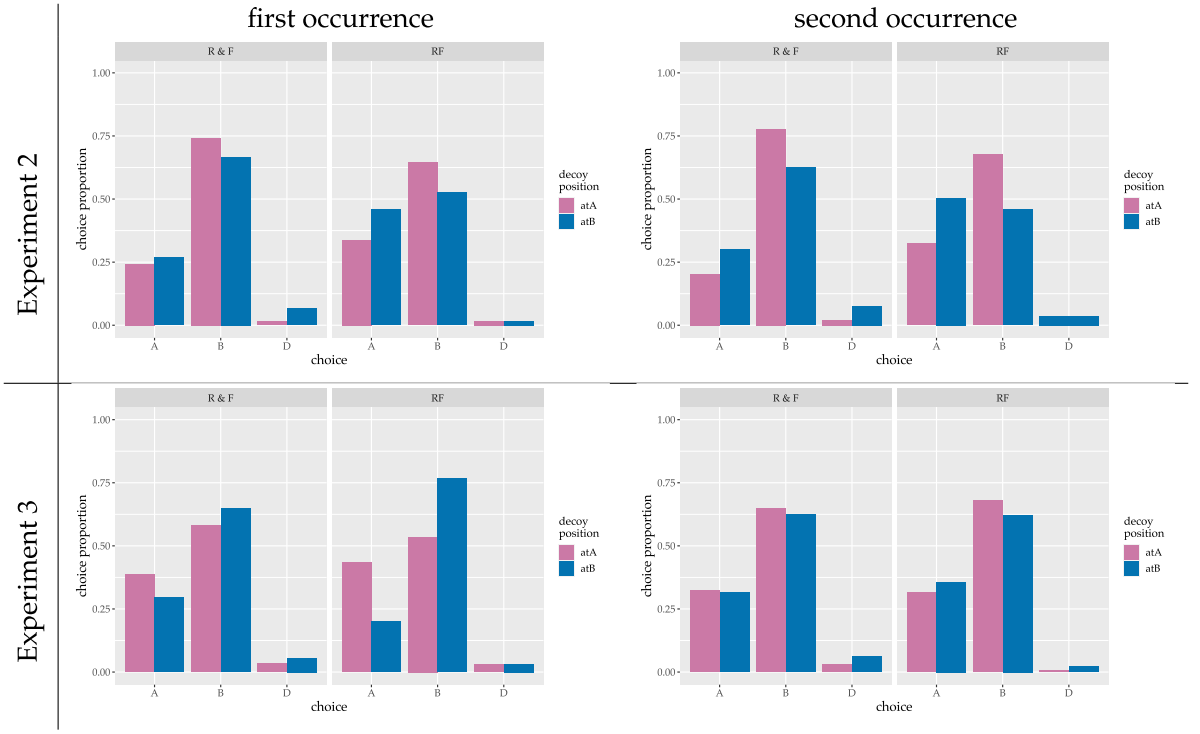



**Fig. 27.** Attraction effect shown as aggregated choice proportions for first and second occurrences of revised items (*worker welfare 2, worker welfare, jail overcrowding 2*) in Experiment 2 (N = 475) and 3 (N = 456).

**Table 24**
Full posteriors of model parameters in Experiment 3.

| Parameter | Rhat | n_eff | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|---|---|
| b_Intercept | 1.00 | 2975 | 0.72 | 0.08 | 0.56 | 0.73 | 0.88 |
| b_decoy_type_baseline (R&F) | 1.00 | 8414 | −0.04 | 0.04 | −0.13 | −0.04 | 0.04 |
| sd_scenario_Intercept | 1.00 | 2555 | 0.19 | 0.08 | 0.09 | 0.17 | 0.40 |
| r_scenario[worker welfare,] | 1.00 | 3662 | 0.17 | 0.09 | 0.00 | 0.17 | 0.35 |
| r_scenario[inevitable injury,] | 1.00 | 3695 | −0.10 | 0.09 | −0.28 | −0.10 | 0.08 |
| r_scenario[jail overcrowding 2,] | 1.00 | 3999 | −0.10 | 0.09 | −0.29 | −0.10 | 0.08 |
| r_scenario[jail overcrowding,] | 1.00 | 4034 | −0.15 | 0.09 | −0.34 | −0.15 | 0.03 |
| r_scenario[rescue plan,] | 1.00 | 3657 | −0.07 | 0.09 | −0.25 | −0.07 | 0.10 |
| r_scenario[worker welfare,] | 1.00 | 3757 | 0.12 | 0.09 | −0.05 | 0.12 | 0.31 |
| r_scenario[emergency delivery,] | 1.00 | 3839 | 0.14 | 0.09 | −0.03 | 0.13 | 0.32 |

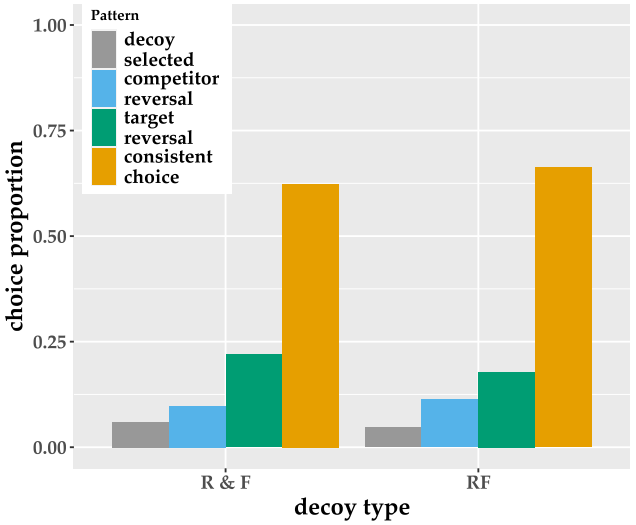



**Fig. 28.** Descriptive results for identical items aggregated in Experiment 2 and 3 combined data (N = 931).

Therefore, our first step is to generate all 24 possible rankings of four levels in an attribute ($a_1a_2a_3a_4$, $a_1a_2a_4a_3$, $a_1a_3a_2a_4$, $a_1a_3a_4a_2$, $a_1a_4a_2a_3$, $a_1a_4a_3a_2$, $a_2a_1a_4a_3$, $a_2a_1a_3a_4$, $a_2a_3a_4a_1$, $a_2a_3a_1a_4$, $a_2a_4a_1a_3$, $a_2a_4a_3a_1$, $a_3a_1a_4a_2$, $a_3a_1a_2a_4$, $a_3a_2a_4a_1$, $a_3a_2a_1a_4$, $a_3a_4a_1a_2$, $a_3a_4a_2a_1$, $a_4a_1a_3a_2$, $a_4a_1a_2a_3$, $a_4a_2a_1a_3$, $a_4a_2a_3a_1$, $a_4a_3a_1a_2$, $a_4a_3a_2a_1$). These rankings are strict total order sets. We do not exclude the possibility that there are intransitive orders.

Second, given a ranking for one attribute and a ranking for another attribute, we reconstruct which of the original options ($A$, $B$, $D$) is the target, competitor, and decoy in the current decision problem. However, not all rankings allow us to yield a mapping between original options ($A$, $B$, $D$) and an *Attraction Configuration* (i.e., a target, a competitor, and a decoy). We describe how we discover all possible structures below.

For each question, we assume 2 attributes with four levels in each attributes and we are able to construct options with 2 attributes (16 possible options). To choose three options out of 16 possibilities without repetition, there are 560 sets given $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. We plot each set of the three options in a $2D$ space and match them to the 5 possible configurations (Table 13):

1. *Best Option Configuration*: 316 out of 560 sets have this configuration.
2. *Worst Option Configuration*: 100 out of 560 sets have this configuration.
3. Among the rest 144 pairs, 128 pairs have an *Attraction Configuration*.
4. The rest pairs have either a *Similarity Configuration* (10 pairs) or a *Compromise Configuration* (6 pairs).

### *E.2. Generating choices given decision problems' structures*

Recall that in Experiment 3, each subject sees all 8 scenarios, among which 6 scenarios have both R & F decoy and RF decoy, and 2 scenarios only have R & F decoy. Thus, each subject randomly sees a total of 16 questions — the R & F-decoy or RF-decoy version of the 6 scenarios and the other 2 scenarios, and the manipulation of decoy position (i.e., dominance) is within subject.

We use the following choice function to generate choices for 16 ethical dilemmas in Experiment 3:

1. Each subject does 8 pairs of questions, corresponding to 8 scenarios. For each scenario, we look at the structure of the 2 questions in a pair.


**Fig. 29.** Simulated and empirical "consistent choice" selection rates (left) and "competitor reversal" rates (right) for each scenario and decoy type. Generally, simulated same selection rates were lower than those in empirical data and simulated "competitor reversal" selection rates were higher than those in empirical data.

2. If both questions have an *Attraction Configuration*, then we sample a choice pattern given the Wedell (1991) data. We have chosen to use the distribution of response patterns (Table 25) given Wedell (1991) data as Wedell (1991) tasks use gambles as decision problems, where both attributes involved in the decision — probability of winning and the amount of money to win have a clear ranking across individuals (i.e., higher probability and higher amount are desirable) — resulting in less noisy responses. However, we have adjusted the decoy selection rate to match the mean decoy selection rate from our empirical results (0.08) in Experiment 3 and re-normalized the distributions.
3. If only one question has an *Attraction Configuration*, then for the single question with an *Attraction Configuration*, we sample a choice given the Wedell data (also adjusted the decoy selection rate, see Table 26).
4. For the questions that do not have an *Attraction Configuration*, we sample 2 choices sequentially (given the randomly assigned order in which subject sees the question).

    - If seeing a scenario for the first time:

        a. this question has the *Best Option Configuration*: select best option with 1-error ($\epsilon$)
        b. this question has the *Worst Option Configuration*: select worst option with error $\epsilon$
        c. this question has the *Similarity Configuration*: sample a choice given the pattern in Trueblood (2012) data (Table 27).
        d. this question has the *Compromise Configuration*: sample a choice given the pattern in Trueblood (2012) data (Table 28). Since we do not have decoy options in this case as defined in Trueblood (2012), both options other than the compromising option are considered as extreme options.

    - If seeing a scenario for the second time: with $\mathbf{p_{consistent}}$, select the same option as before; with $(1 - \mathbf{p_{consistent}})$, select an option that is not the same option as before.

Lastly, we map the choices back onto the original $A, B, D$ in the decision problem.

**Table 25**
Distribution of choice patterns in Wedell (1991) after re-normalizing.

| Decoy type | Pattern | prob |
|---|---|---|
| R & F | Consistent choice | .64 |
| | Target reversal | .19 |
| | Competitor reversal | .06 |
| | Decoy selected | .08 |
| RF | Consistent choice | .69 |
| | Target reversal | .16 |
| | Competitor reversal | .06 |
| | Decoy selected | .08 |

**Table 26**
Marginal distributions of choices (Wedell, 1991) after re-normalizing.

| Decoy position | Decoy type | Option | prob |
|---|---|---|---|
| atA | R & F | A | .65 |
| | | B | .25 |
| | | D | .08 |
| atB | R & F | A | .52 |
| | | B | .38 |
| | | D | .08 |
| atA | RF | A | .66 |
| | | B | .24 |
| | | D | .08 |
| atB | RF | A | .56 |
| | | B | .34 |
| | | D | .08 |

**Table 27**
Distribution of choices in Trueblood (2012) — similarity effect. Focal option refers to the option that is enhanced by the decoy.

| Option | prob |
|---|---|
| Decoy | .2 |
| Focal option | .5 |
| Non-focal option | .3 |

**Table 28**
Distribution of choices in Trueblood (2012) — compromise effect.

| Option | prob |
|---|---|
| Compromise | .48 |
| Extreme options | .26 |

### *E.3. Simulation results*

See Fig. 29.

## Appendix F. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cogpsych.2024.101672.

## References

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337. http://dx.doi.org/10.1073/pnas.1911517117.

Barak-Corren, N., Tsay, C.-J., Cushman, F., & Bazerman, M. H. (2018). If you're going to do wrong, at least do it right: Considering two moral dilemmas at the same time promotes moral consistency. *Management Science*, *64*(4), 1528–1540. http://dx.doi.org/10.1287/mnsc.2016.2659.

Baron, J., & Ritov, I. (1993). Intuitions about penalties and compensation in the context of tort law. In C. Camerer, & H. Kunreuther (Eds.), *Making decisions about liability and insurance: a special issue of the journal of risk and uncertainty* (pp. 17–33). Dordrecht: Springer Netherlands, http://dx.doi.org/10.1007/978-94-011-2192-7_2.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*, 74–85. http://dx.doi.org/10.1016/j.obhdp.2004.03.003.

Bergner, A. S., Oppenheimer, D. M., & Detre, G. (2019). VAMP (voting agent model of preferences): A computational model of individual multi-attribute choice. *Cognition*, *192*, Article 103971.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models.. *Journal of Experimental Psychology: General*, *143*(3), 1331.
Bhatia, S., Loomes, G., & Read, D. (2021). Establishing the laws of preferential choice behavior. *Judgment & Decision Making*, *16*(6).
Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). *Moral AI: And how we get there*. Random House.
Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. http://dx.doi.org/10.18637/jss.v080.i01.
Chang, L. W., & Cikara, M. (2018). Social decoys: Leveraging choice architecture to alter social preferences. *Journal of Personality and Social Psychology*, *115*(2), 206.
Chang, L. W., Gershman, S. J., & Cikara, M. (2019). Comparing value coding models of context-dependence in social choice. *Journal of Experimental Social Psychology*, *85*, Article 103847.
Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences*, *111*(48), 17320–17325. http://dx.doi.org/10.1073/pnas.1408988111.
Erlandsson, A. (2021). Seven (weak and strong) helping effects systematically tested in separate evaluation, joint evaluation and forced choice. *Judgment and Decision Making*, *16*(5), 1113–1154.
Erlandsson, A., Lindkvist, A., Lundqvist, K., Andersson, P. A., Dickert, S., Slovic, P., et al. (2020). Moral preferences in helping dilemmas expressed by matching and forced choice. *Judgment and Decision Making*.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, *5*, 5–15.
Frederick, S., Lee, L., & Baskin, E. (2014). The limits of attraction. *Journal of Marketing Research*, *51*(4), 487–507.
Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528–554. http://dx.doi.org/10.1111/j.1756-8765.2010.01094.x.
Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. Retrieved from https://mc-stan.org/rstanarm. R package version 2.21.1.
Ha, Y.-W., Park, S., & Ahn, H.-K. (2009). The influence of categorical attributes on choice context effects. *Journal of Consumer Research*, *36*(3), 463–477.
Herne, K. (1997). Decoy alternatives in policy choices: Asymmetric domination and compromise effects. *European Journal of Political Economy*, *13*(3), 575–589.
Highhouse, S. (1996). Context-dependent selection: The effects of decoy and phantom job candidates. *Organizational Behavior and Human Decision Processes*, *65*(1), 68–76.
Horvath, J., & Wiegmann, A. (2022). Intuitive expertise in moral judgments. *Australasian Journal of Philosophy*, *100*(2), 342–359.
Howes, A., Warren, P. A., Farmer, G. D., El-Deredy, W., & Lewis, R. L. (2016). Why contextual preference reversals in humans maximize expected value. *Psychological Review*, http://dx.doi.org/10.1037/a0039996.
Huber, J., Payne, J. J., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*(1), 90–98.
Huber, J., & Puto, C. (1983). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of Consumer Research*, *10*(1), 31–44. http://dx.doi.org/10.1086/208943, Retrieved from https://doi.org/10.1086/208943.
Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. http://dx.doi.org/10.1126/science.1091721.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Vol. 49*, In *Heuristics and biases: the psychology of intuitive judgment* (pp. 49–81). http://dx.doi.org/10.1017/CBO9780511808098.004.
Kahneman, D., Schkade, D., & Sunstein, C. (1998). Shared outrage and erratic awards: The psychology of punitive damages. *Journal of Risk and Uncertainty*, *16*(1), 49–86. http://dx.doi.org/10.1023/A:1007710408413.
Kant, I. (1998). *Kant: Religion within the boundaries of mere reason: And other writings*. Cambridge University Press.
Katsimpokis, D., Fontanesi, L., & Rieskamp, J. (2022). A robust bayesian test for identifying context effects in multiattribute decision-making. *Psychonomic Bulletin & Review*, 1–18.
Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., et al. (2018). A computational model of commonsense moral decision making. CoRR abs/1801.04346.
Koehler, J. J., & Gershoff, A. D. (2003). Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes*, *90*(2), 244–261. http://dx.doi.org/10.1016/S0749-5978(02)00518-6.
Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.
Latty, T., & Beekman, M. (2011). Irrational decision-making in an amoeboid organism: transitivity and context-dependent preferences. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1703), 307–312.
Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.
Merlhiot, G., Mermillod, M., Jean-Luc, L. P., Dutheil, F., & Mondillon, L. (2018). Influence of uncertainty on framed decision-making with moral dilemma. *PLoS One*, *13*(5).
Nadurak, V. (2018). Two types of heuristics in moral decision making. *Filosifija Sociologija*, *29*(3), 141–149.
Nadurak, V. (2020). Why moral heuristics can lead to mistaken moral judgments. *Kriterion (Austria)*, *34*, 99–113.
Nagel, T. (2012). *Canto classics: Vol. Canto edition, Mortal questions*. Cambridge University Press.
O'Curry, Y. P., & Pitts, R. (1995). The attraction effect and political choice in two elections. *Journal of Consumer Psychology*, *4*(1), 85–101. http://dx.doi.org/10.1207/s15327663jcp0401_04, Retrieved from https://www.sciencedirect.com/science/article/pii/S1057740895704247.
Parrish, A. E., Evans, T. A., & Beran, M. J. (2015). Rhesus macaques (macaca mulatta) exhibit the decoy effect in a perceptual discrimination task. *Attention, Perception, & Psychophysics*, *77*(5), 1715–1725.
Pettibone, J. C. (2012). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and Decision Making*, *7*(4), 513.
R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.
Rodriguez-Arias, D., Rodriguez Lopez, B., Monasterio-Astobiza, A., & Hannikainen, I. R. (2020). How do people use 'killing', 'letting die' and related bioethical concepts? Contrasting descriptive and normative hypotheses. *Bioethics*, *34*(5), 509–518. http://dx.doi.org/10.1111/bioe.12707.
Rozin, P. (2001). Technological stigma: Some perspectives from the study of contagion. In J. Flynn, P. Slovic, & H. Kunreuther (Eds.), *Risk, media, and stigma: understanding public challenges to modern science and technology* (pp. 31–40). London: Earthscan, (Chapter 3).
Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118. http://dx.doi.org/10.2307/1884852.
Sinnott-Armstrong, W., Young, L., & Cushman, F. (2010). Moral intuitions. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 246–272). Oxford University Press.
Sunstein, C. R. (2002). *Risk and reason: safety, law, and the environment*. Cambridge University Press.
Sunstein, C. R. (2004). Lives, life-years, and willingness to pay. *Columbia Law Review*, *104*(1), 205–252.
Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, *28*(4), 531–542. http://dx.doi.org/10.1017/S0140525X05000099.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.
Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic bulletin & review*, *19*, 962–968.
Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological Science*, *24*(6), 901–908. http://dx.doi.org/10.1177/0956797612464241.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.
van Baar Jeroen, M., Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1).
Viscusi, W. K. (2000). Corporate risk analysis: A reckless act? *Stanford Law Review*, *52*(3), 547–597.
Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 767–778.
Williams, B., & Bernard, W. (1981). *Cambridge paperback library*, *Moral luck: Philosophical papers 1973–1980*. Cambridge University Press.
Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modeling morality in 3-d: Decision-making, judgment, and inference. *Topics in Cognitive Science*, *11*(2), 409–432. http://dx.doi.org/10.1111/tops.12382.