Context Effects, Value Learning, and Individual Differences in Value-Based Decisions

by

Chenxu Hao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Psychology)
in The University of Michigan
2021

Doctoral Committee:

Professor Richard L. Lewis, Chair
Dr. Mara Bollard
Professor Thad Polk
Professor Colleen Seifert
Professor Chandra Sekhar Sripada

Chenxu Hao

chenxuh@umich.edu

ORCID iD: 0000-0001-7498-547X

To my dear parents,

Lei Chen and *Lei Hao*

ACKNOWLEDGEMENTS

Throughout graduate school and the writing of this dissertation, I have received countless support.

First and foremost, I would like to express my deepest gratitude to my advisor, Rick Lewis. His continuous trust, understanding, guidance, patience, and encouragement always keep me inspired to never stop learning and never give up. I remember that when I started my graduate school research work, I was very afraid of making mistakes, and I once asked Rick:"what if I had made a mistake by accident and would have to retract a paper?" And he told me:" that would always be better than having a published mistake unnoticed." Because of him (and Richard Feynman's stories, to be fully honest), I will always keep in mind the value of integrity and always dare to ask out loud all the "why" and "how" questions about my own work and the world around me.

I would like to thank the members of my committee, Mara Bollard, Thad Polk, Colleen Seifert, and Chandra Sripada for their great insights, support, and valuable feedback.

I would also like to thank my amazing collaborators and colleagues, especially the past and current members of Lewis lab: Tyler Adkins, Logan Bickel, Pyeong Whan Cho, Ian Cook, Hannah Foster, Nicole Hamilton, Steven Langsford, Sarah Marks, Connor McMann, Soo Hyun Ryu, and Logan Walls. I'd like to thank them for the many helpful discussions and contributions over the many years.

I am grateful to my peers and friends for their support and company. Thank you

to Julia Liao, Ziyong Lin, Lilian Cabrera-Haro, and Colleen Frank for always believing in me and caring for me.

Finally, my deepest gratitude to my family from near and afar: my mom and dad — Lei Chen and Lei Hao, my partner Nikolas Eptaminitakis, and my dearest, most hungry, most hilarious, most mischievous feline friend Lucky.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

In this dissertation, we present two lines of research that investigate value-based decisions. The first focuses on value-based decisions in multi-attribute choices. Specifically, we investigate contextual preference reversals in multi-attribute decisions. These reversals occur when the choice preference between two options changes in the presence of a third unchosen decoy. We demonstrate for the first time that these reversals also occur in ethical dilemmas involving both quantitative and qualitative attributes. However, these reversals do not arise to the same extent across ethical dilemmas. We use a generative computational model to show that the variation of reversals across dilemmas can be partly explained by individual differences in rankings of ethical features.

The second line of work focuses on value-based decisions in the Value Learning Task (VLT; Raymond & O'Brien, 2009), a paradigm where people learn values associated with options in win and loss conditions through trial-and-error while trying to maximize accumulated reward. The VLT has a symmetric outcome structure for wins and losses. However, people consistently learn wins better than losses (Lin et al., 2020). We investigate the nature of this asymmetry with a simple reinforcement learning model. The model predicts the learning asymmetry observed in empirical data regardless of whether the parameters are set to maximize empirical fit or total payoff in the task. This asymmetry arises as a result of the interaction between a neutral initial value estimate and a choice policy that exploits while exploring, leading to more poorly discriminated value estimates for loss stimuli. We also illustrate that the final

value estimates produced by the model can provide a simple account of a post-learning explicit value categorization task. Lastly, we also investigate how differences in estimated individual learning rates help to explain individual differences in the observed win-loss asymmetries.

Together, these two lines of research investigate some complicated aspects of value-based decisions such as value learning through experience and attribute-value integration and evaluation in multi-attribute choices. However, beyond the complex phenomena, our two lines of work integrate the same simple theory — the bounded rationality framework. In this dissertation, we also discuss how our research connect to the bounded rationality framework.

CHAPTER I

Introduction

Suppose that you drive an old used car. One day while driving on the highway, you notice that the gear has suddenly slipped. You know that this indicates a transmission issue and takes your car to a car mechanic. The car mechanic checks your car, tells you that the car is safe to drive and gives you three options: a partial fix of the troubled parts that leads to a good probability of fixing the issue and costs you a moderate amount of money, a full fix that guarantees fixing the issue and costs you an extremely high amount of money, or not fixing the issue — which saves your money. Presented with these alternatives and their attributes (or dimensions), you consider which option yields the highest (subjective) value for you — for example, if you want to save money, then you might go with the third option. Such a decision is an example of a *value-based decision*, where the decision maker (in this case, you) processes the information, learns the value of the alternatives, evaluates the options, and makes a choice on the basis of subjective value (Rangel, Camerer, & Montague, 2008). Everyone encounters situations that require making *value-based decisions* in their daily life — such decisions could be as small as buying a product in the store or as significant as choosing a career or a life partner.

Value as a specific concept could be defined in multiple ways. We distinguish two main definitions in order to present our work: a typical understanding of value is the

economic value of an option, which refers to the amount of benefit one could gain from choosing that option (Brosch & Sander, 2013). One example is the expected value (EV) — in a gamble where one is presented with options that have known probability of winning and the amount of money to win, the decision maker could decide whether to play the gamble by simply calculating the EV (probability $*$ amount); another example is the subjective expected utility (Savage, 1972) — in the car problem above, the decision maker may combine the two attributes in the options with some subjective utility function, allowing all options to be on the same scale for evaluation. Economic value could also be linked together with personal values concerning personal belief related to social relations, cultural background, etc (Brosch & Sander, 2013). Another definition of value is specified in the reinforcement learning (RL) theory (Sutton & Barto, 2018), where the value or value function of a state-action pair is its expected cumulative future rewards (Gershman & Daw, 2017; Sutton & Barto, 2018).

In this thesis, we present our empirical and computational modeling work investigating *value-based decisions* in two domains: *ethical dilemmas* where individuals make choices among multi-attribute options while making some tradeoff among the attributes and the *Value Learning Task* (VLT; Raymond & O'Brien, 2009; Lin et al., 2020) — a common paradigm where individuals make choices between a pair of neutral stimuli associated with either win or loss while trying to maximize total earned reward. We also discuss the connection between our work and the framework of bounded rationality — specifically how context effects in multi-attribute ethical decisions can potentially be explained in the bounded rationality framework given recent work, and how we apply the framework of computational rationality to analyze human performances in the VLT.

In this introduction, we first provide some relevant background on value learning, multi-attribute choice, and the bounded rationality framework. We then provide an overview of this dissertation.

1.1 Learning Values from Experience

Value-based decision making provides a common framework for analyzing decisions in both humans and other animals across various domains (e.g., how humans choose consumer goods, trade stocks, make plans, and animals' foraging behaviors, etc; Rangel et al., 2008). One important process in making a value-based decision is to identify the available actions/options and learn the values of them (Rangel et al., 2008), and a prominent way to learn the values is from experience (Kahneman, 2003; Gershman & Daw, 2017). In the car example above, you may decide which fixing option to choose based on previous experience with how you have used your car and communicated with car mechanics.

Learning values from experience can be modeled within the reinforcement learning (RL; Sutton & Barto, 2018) framework, which not only provides explanations for the processes of value-based decisions with a formal computational theory, but also provide theoretical foundations for understanding the underlying neural mechanisms of value learning (Montague, Hyman, & Cohen, 2004; Daw, Niv, & Dayan, 2005; Brosch & Sander, 2013). The RL theory formally defines the problem where the agent learns the association between state-action/observation pairs and rewards through trial and error while aiming to maximize cumulative rewards. In the RL problem, an agent starts with some initial value for each option or state-action pair. The initial value could be zero, random, or estimates based on past experience. At each time step, given a state or observation, the agent chooses an action/option to take. Consequently, the agent receives a reward at the following time step and updates the value of the chosen option with some value-updating function. As the agent gains more experience through trials and error, the value estimates for the options approach the true values, which would yield better actions (Sutton & Barto, 2018).

The RL theory could be applied to both repeated one-shot decisions and sequential

decisions. In later chapters, we apply the RL theory to the domain of the Value Learning Task (VLT; Raymond & O'Brien, 2009; Lin et al., 2020) — a simple case of the RL problem that involves only repeated one-shot decisions in three independent conditions (win, loss, and control). We also investigate the individual differences in the VLT.

Furthermore, in the RL theory (Sutton & Barto, 2018), the processes of learning, evaluation, and decision, rely on various external and internal factors relating to the decision problem and the decision maker — external factors include the environment of the decision problem and internal factors include memory (Gershman & Daw, 2017), how fast the decision maker learns, how much the decision maker cares about immediate reward, etc. This indicates that the decision maker often aims to make the (subjectively) best possible choice *without* perfect information and *with* various risks and uncertainty.

Different from typical economic models that focus on decisions based on description, the RL theory emphasizes learning from experience, where the structure of the decision problem must be learned rather than told explicitly (Gershman & Daw, 2017). Research has shown that decisions made with values learned from experience and those learned from description often diverge (Erev, Ert, & Yechiam, 2008; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig & Erev, 2009; Gershman & Daw, 2017). Our work presented in this thesis does not directly address the differences between decisions based on description and decisions based on experience, but our work spans both: value-based decisions in the VLT (Chapters IV, V) are decisions from experience whereas value-based decisions in the ethical domain (Chapters II, III) are decisions from description, where decision makers are presented with multi-attribute options in various ethical scenarios. In the following section, we will provide a brief introduction to multi-attribute decisions.

1.2 Multi-attribute Choice

Multi-attribute decisions require the decision maker to combine the different attributes (or dimensions) of each option given some subjective utility function. Thus, the values of options can be compared on the same scale. While comparing the options in multi-attribute decisions, one often needs to make some tradeoff among attributes. In the car example at the beginning, the decision maker must make a tradeoff between how much the car can be fixed and the cost of service: all three options in this case result in a certain level of goodness of the car accompanied by some cost — on one end, the goodness is high and the cost is also high; on the other end, both attributes have low values; in the middle is the option that results in a potentially satisfactory level of goodness and a medium cost. Given this specific context, decision makers often choose the compromise option in the middle (Huber & Puto, 1983b; Simonson, 1989).

Researchers have extensively studied how choices in multi-attribute decision problems are affected by various choice contexts like this (Wollschlaeger & Diederich, 2020). These effects are known as *context effects*. In a typical multi-attribute choice set that produces context effects, there are often three options with two attributes — two options that have the same (or roughly equal) expected utility but require a tradeoff between the two attributes. The decision maker's preference between the two equal options may depend on how the decision maker weights each attribute. However, the third option (also known as the "decoy") is placed strategically by the experimenter in the choice space, providing a *context* that affects the expressed choice preferences of decision makers. Three typical context effects are the attraction effect (also known as contextual preference reversals, or contextual choice reversals; Huber, Payne, & Puto, 1982; Wedell, 1991; Trueblood, 2012; Trueblood, Brown, Heathcote, & Busemeyer, 2013; Simonson, 1989), the similarity effect (Tversky, 1972; Trueblood, 2012; Trueblood et al., 2013; Berkowitsch, Scheibehenne, & Rieskamp, 2014; Liew, Howe, &

Little, 2016), and the compromise effect (Huber & Puto, 1983a; Simonson, 1989). We summarize the three typical context effects in Table 1.1 below. These context effects are observed both between subjects and within subjects.

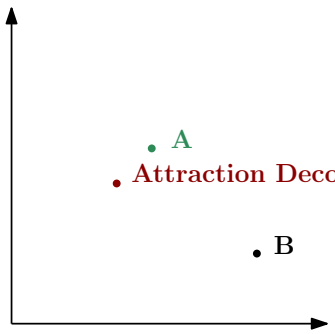
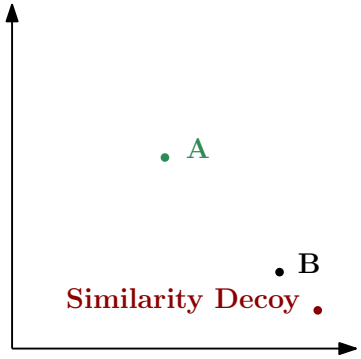
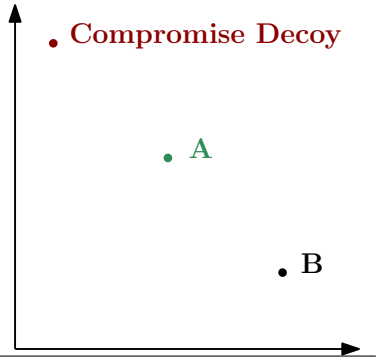
Contexts	Effects	Illustrations
Attraction Decoy: dominated by one of the options, the target; close to the target in the decision space; expected utility is close to, but lower than either option	causes the dominating option, A, to be selected more often	
Similarity Decoy: similar to one of the options (both in terms of decision space and expected utility);	causes the similar option, B, to be selected less often	
Compromise Decoy: has similar expected utility to both options and turns one of the options into a compromise option between decoy and the other option	causes the intermediate item, A, to be selected more often	

Table 1.1: Summary of Three Typical Context Effects
(Wollschlaeger & Diederich, 2020)

According to rational choice theory, the values of options should be independent of other options in the set (Luce, 2012). In other words, the choice preference between two options shouldn't be affected by the decoy. This indicates that context effects violate

the assumptions in rational choice theory and rationality based on utility maximization. However, the framework of bounded rationality (Simon, 1955) motivates models of utility-maximizing strategies adapted to cognitive bounds. These models (Howes, Warren, Farmer, El-Deredy, & Lewis, 2016) provide explanations for context effects while retaining rationality given external and internal influences. We will provide an introduction to this framework in a later section.

The work presented in this dissertation mainly focuses on the attraction effect. The attraction effect has been found in various domains such as choosing gambles (Wedell, 1991) and buying consumer goods (Huber et al., 1982). Our empirical experiments extend the effect to the domain of ethical decisions, where the tradeoffs are more complicated, posing ethical dilemmas to the decision makers. We also explore individual differences in the context effects in multi-attribute ethical decisions.

1.3 Individual Differences

From value learning in value-based decisions to making multi-attribute choices, questions naturally arise about the nature of individual differences. In this section, we discuss individual differences in value learning and in multi-attribute decisions related to our work.

1.3.1 Individual Differences in Learning Values from Experience

Past work on how values can be learned from the decision maker's individual experiences and how value-based decisions are connected to individual memories (Gershman & Daw, 2017) naturally indicates that the differences in experience and memory could lead to different learning results and decisions.

However, it is possible that experience is not the only thing that drives the differences in decisions. The RL computational theory provides us with a method to model

value-based decisions on an individual, trial-by-trial level (Daw, 2011) and gives us insights on each individual's learning rate and how each individual balances exploration vs. exploitation.

In our work on value-based decisions in the VLT, we explore whether the individual differences in value learning are experience driven or are related to further individual characteristics captured by the RL model. To preview our main result, we find that the individual differences in value learning are not purely experience driven, but also affected by individual characteristics such as learning rate.

1.3.2 Individual Differences in Multi-attribute Choice

Although context effects arise in multi-attribute choice in a variety of domains, previous studies have focused on establishing the effects but have not made comparisons across choice domains. Some studies have also shown individual differences across three context effects (participants who show the similarity effect rarely show attraction and compromise effects; Berkowitsch et al., 2014; Liew et al., 2016).

Furthermore, when the attributes in a decision problem are qualitative, there is not a straightforward way to combine and compare them on the same scale (unlike expected values in gambles). Individuals may weigh or evaluate the attributes differently and have different subjective utility functions to combine the attributes, affecting how they compare the options, and in turn affecting the patterns of context effects.

In our work that explores contextual choice reversal in the ethical domain, we investigate several aspects of individual differences. We compare the performances and choice reversal rates between ethical and non-ethical economic decisions. We also use a simple computational model to provide explanations for the differences in the rates of choice reversals among various ethical scenarios by taking into account individual differences in ranking the different levels of ethical attributes.

In both learning values from experience and making decisions among multi-attribute options, individuals face various environmental constraints (e.g., the structure of the tasks, choice contexts) and internal cognitive constraints (e.g., how fast one learns value) when they make decisions.

1.4 Bounded Rationality

Neoclassical economics assumes such decision makers to be *homo economicus*, i.e., the perfectly rational agents who behaves in a way that maximizes utility. Contrary to classical economic theory, bounded rationality (Simon, 1955) offers a perspective where decisions and actions are adapted to the task environment and the mind — as in real life, humans often do not have perfect information and must act while facing uncertain or risky situations.

In the face of uncertainty and risks, decision makers often use a variety of heuristics and are prone to certain biases in making judgments and decisions (Kahneman, 2003). Part of Kahneman and Tversky's main research program provides a comprehensive picture of various heuristics and systematic biases in human decisions such as representativeness, availability, and anchoring heuristics (Kahneman & Tversky, 1972; Tversky & Kahneman, 1973, 1974). Kahneman (2003) summarizes how these heuristics are often useful, but can also lead to sub-optimal consequences and judgment errors compared to optimal choices given by utility maximization.

An alternative to heuristics and biases approach is ecological rationality. Instead of focusing on how heuristics could lead to limitations and biases that violate rationality based on utility maximization, Todd et al. (1999) state that heuristics ought to be viewed together with the environment such as time pressure, the availability of information, limitation of resources, etc. With certain environmental and cognitive constraints, fast-and-frugal heuristics could even lead to better choices compared to more deliber-

ative or comprehensive decision strategies. Gigerenzer & Selten (2001) conceptualize heuristics that individuals use in various environments as an adaptive toolbox, which is an example of Simon's *procedural rationality* (Simon, 1978). Procedural rationality is closely associated with *bounded rationality* — under this framework, the evaluation of the effectiveness of actions takes into account environmental uncertainties and the limited information processing capacities individuals have.

Bounded rationality (Simon, 1955) studies how an individual's rationality is compatible with the characteristics of the environment and the cognitive limitation of the mind. In other words, a decision making agent is not absolutely rational by definition based on traditional economic theory, but is "approximately rational" or at least "intends to be rational" under the influence of the environment and the choosing agent's limited knowledge and ability.

One approach to work in artificial intelligence has formalized bounded rationality as the concept of *bounded optimality*, which states that a bounded agent should do whatever the best program running on its information-processing architecture would do (Russell & Subramanian, 1995). Lewis, Howes, & Singh (2014) applied bounded optimality to psychology with the framework of *computational rationality*. This framework provides a method to take into account the environment bounds as well as the bounds of the agent's cognitive system in rational analyses (Lewis et al., 2014; Gershman, Horvitz, & Tenenbaum, 2015). A similar framework, *resource rationality*, provides a method to analyze and understand how human decision strategies and planning could be optimal given limited cognitive resources (Lieder & Griffiths, 2018; Callaway et al., 2018).

The current dissertation investigates value-based decisions in two different domains using empirical experiments and computational models. We discuss how phenomena in both domains can be connected to the framework of bounded rationality, specifically computational rationality, later in this dissertation. In the following section, we provide

an overview of this dissertation and the specific topics relevant to each line of our research.

1.5 Dissertation Overview

In Chapter II, we present three empirical experiments that demonstrate, for the first time, that contextual choice reversals also arise in decisions involved in ethical dilemmas. Some ethical dilemmas involve attributes such as probability and numbers of lives to save, which allow the decision maker to calculate expected values of options directly. However, some dilemmas contain qualitative attributes, indicating that the decision maker would combine the attributes in certain ways to evaluate options. Our empirical results in Chapter II suggest that contextual choice reversals vary across different scenarios/dilemmas. Thus, in Chapter III, we present a simple computational model that account for some of this variation by predicting choices based on individual decision maker's different ranking of levels in ethical features.

Chapters IV and V focus on behaviors in the Value Learning Task (VLT; Raymond & O'Brien, 2009), a task paradigm developed by psychologists to understand how acquired value impacts how people perceive and process stimuli. The task consists of a series of trials in which participants attempt to maximize accumulated gains as they make choices from a pair of presented neutral images associated with probabilistic win, loss, or no-change outcomes. Despite the task having a symmetric outcome structure for win and loss pairs, people learn win associations better than loss associations (Lin et al., 2020). This asymmetry could lead to differences when the stimuli are probed in subsequent tasks, compromising inferences about how acquired value affects downstream processing. In Chapter IV, we present our investigation of the nature of this asymmetry using a standard error-driven reinforcement learning model with a softmax choice rule. Despite having no special role for valence, the model yields the asymme-

try observed in human behavior, whether the model parameters are set to maximize empirical fit, or task payoff. This asymmetry arises from an interaction between a neutral initial value estimate and a choice policy that exploits while exploring, leading to more poorly discriminated value estimates for loss stimuli. In Chapter V, we further illustrate how differences in estimated individual learning rates help to explain individual differences in the observed win-loss asymmetries, and how the final value estimates produced by the model provide a simple account of a post-learning explicit value categorization task.

Finally, Chapter VI provides a brief summary of the main findings of this dissertation (Table 6.1). We also discuss the connections between our work and the bounded rationality framework, and open questions for research in relevant domains of value-based decisions.

CHAPTER II

Choice Reversals in Ethical Decisions

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 (Wedell, 1991; Huber et al., 1982; Trueblood et al., 2013; O'Curry & Pitts, 1995). Our focus in this work is on *ethical* or *moral* decisions—decisions concerned with the welfare of others (Yu, Siegel, & Crockett, 2019). The primary contribution of our work is the first clear empirical evidence that systematic *contextual choice reversals* and attraction effects arise in ethical decisions. Such 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.

Ethical or moral decision making has been studied in a variety of ways. These include incentivized choices that involve harm or reward (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Rand, Greene, & Nowak, 2012; van Baar Jeroen, Chang, & Sanfey, 2019) (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), and judgments about the moral appropriateness of actions involving the welfare of others (e.g., using hypothetical dilemmas such as the Trolley Problem; Foot, 1967; Thomson, 1976; Awad, Dsouza, Shariff, Rahwan, & Bonnefon, 2020; Kim et al., 2018; Barak-

Corren, Tsay, Cushman, & Bazerman, 2018; Merlhiot, Mermillod, Jean-Luc, Dutheil, & Mondillon, 2018).

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 making judgments or choices based in part 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 short-cuts" (Sunstein, 2005) to make decisions. 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 some of the most fundamental principles of axiomatic rational choice theory: *consistency* and *independence*. More 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 reversals have been shown to arise in multiple choice domains, e.g. economic gambles, consumer goods, political candidates, perceptual size judgements (Wedell, 1991; Huber et al., 1982; O'Curry & Pitts, 1995; Trueblood et al., 2013) and in multiple decision making organisms, including humans, monkeys, slime molds (Huber et al., 1982; Parrish, Evans, & Beran, 2015; Latty & Beekman, 2011). 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).

In the remainder of the chapter 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. We then provide the details of our three 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 yield 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 without clear objective rankings. These scenarios also yield choice reversals, but to varying degrees.

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.

2.1 Contextual Choice Reversals and the Attraction Effect

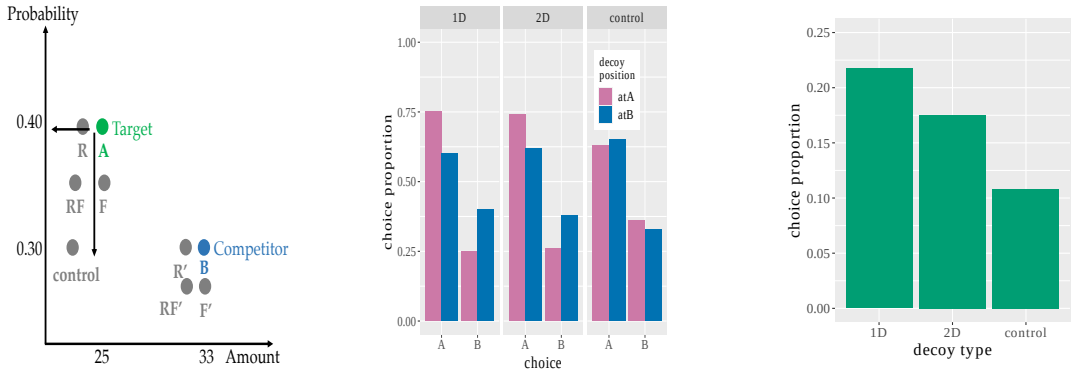
Consider the following decision problem. You are choosing among three video games to purchase, and are weighing price against quality, here assessed as user experience of the game. The choices are a console-version video game that is very expensive and provides you with a great experience, a PC-version of the same video game that is cheaper but a lower quality experience, and a smartphone version that is the same price as the PC-version, but an even lower quality 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 weight 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 do 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 dimensions (2-dimensional dominance), or equal one one and inferior on the other (1-dimensional dominance). 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 expressed preferences is known as the *attraction* effect, because the decoy positioning "attracts" additional choices.

Figure 2.1 (a, left) illustrates possible placements of decoy options using choice among economic gambles (Wedell, 1991) as an example, where each option or gamble is a probability and value pair $\langle p, v \rangle$. Selecting option $\langle p, v \rangle$ means playing ae gamble which pays out v with probability p and 0 with probability $1 - p$. The expected value of each option is thus pv. Figure 2.1 (b, middle) shows the choice proportions reported by Wedell (1991) for both 1-dimensional (1D) and 2-dimensional (2D) variants of the decoy placement. Both the control and 1-D and 2-D 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 1D and 2D decoy placements. Figure 2.1 (c, right) shows the within-subject choice reversal rates for pairs of decision problems reported by Wedell (1991) for 1D and 2D 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 (see Figure A.1, Appendix A.3), 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 Figure 2.1. 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 admit of 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.



(a) Decoy placements

(b) Between-subject choice proportions in Wedell (1991)

(c) Within-subject choice reversals in Wedell (1991)

Figure 2.1: Decoy placement and attraction effect in Wedell (1991).
(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. Options A and B vary in both attributes but have the same expected value. As shown above, decoys R, F, and RF are dominated by A; decoys R', F', and RF' are dominated by C. 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 1-dimensional (1D) and 2-dimensional (2D) 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 1D and 2D dominance problems reported in Wedell (1991).

2.2 Materials & Data Availability

Experiment 1. 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`.

2.3 Experiment 1

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. The dilemma arises when choose 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.

2.3.1 Method

2.3.1.1 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.

2.3.1.2 Materials

In our experiment, decoy position (dominated by A/target versus dominated by B/competitor) and decoy type (1D, 2D, and 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 (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 Qualtrics software (Qualtrics, Provo, UT).

Our ethical dilemmas are isomorphic to the Wedell (1991) tasks while using a scenario of choosing among disaster rescue plans (Figure 2.2): all numerical stimuli (probability and numbers) are identical to Wedell (1991). We created the control (R' decoys) by altering the values of a type of 1D decoy (R decoys) so that the R' decoys were dominated by both targets. Here is an example of a a pair of our task problems:

Decoy at A: 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?

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.

B. A program that leads to a 30% chance of saving 33 people.

Decoy at B: A hurricane hits a small town causing most houses to be destroyed. Three emergency rescue plans have been proposed. Assume that the exact scientific

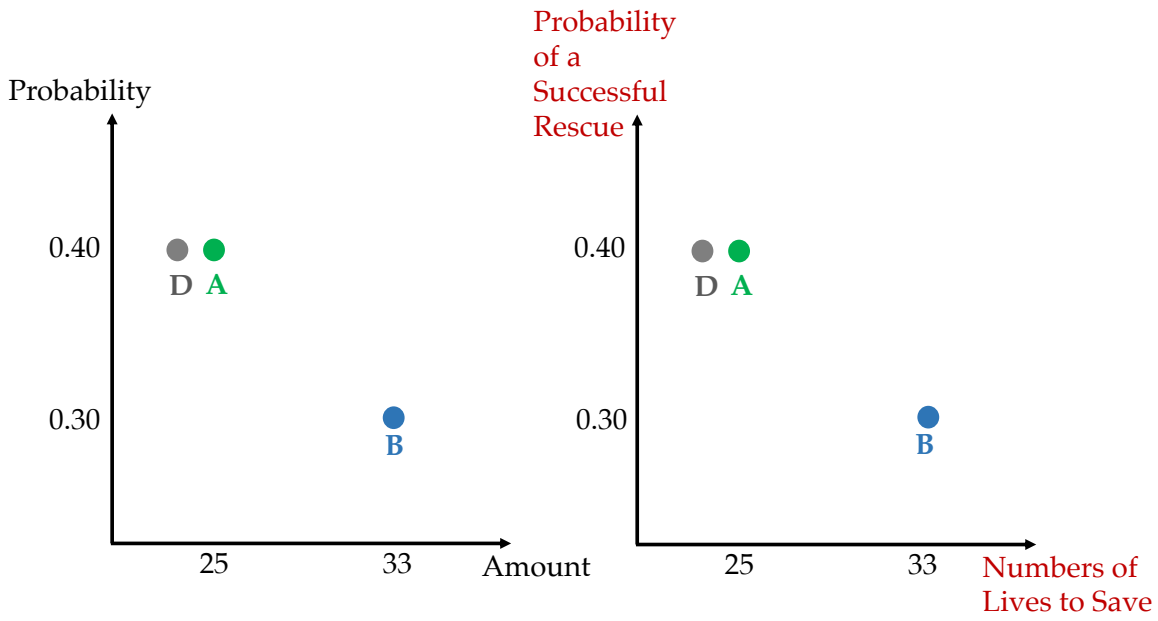


Figure 2.2: The illustration of the task structures of Wedell (1991) decision problem (left) and our ethical dilemmas (right).

estimates of the consequences of the programs are as follows, which program would you choose?

A. A program that leads to a 40% chance of saving 25 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.

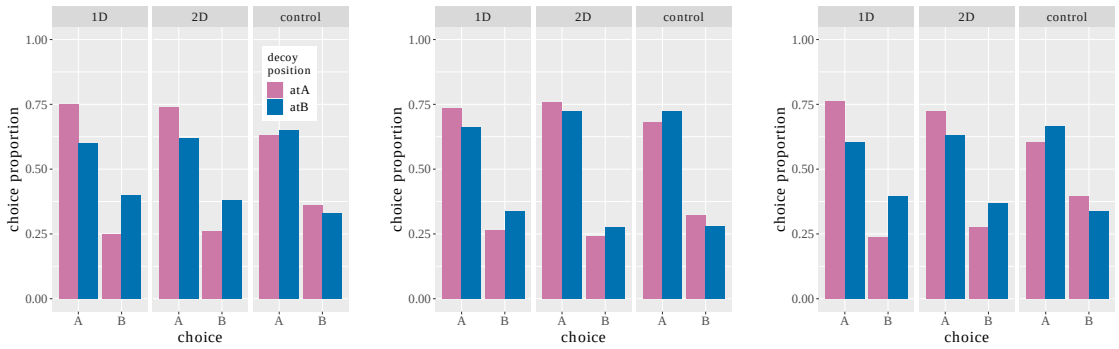
Attention Check. We gave participants one attention check question. This question was always presented at the end of the survey so that it did not interfere with other scenarios. The question stated that the participant can only have time to save people from one out of the three rooms in a burning house 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.

2.3.2 Results

2.3.2.1 Descriptive Analysis

In this section we show the descriptive results for our Experiment 1. As we focus on the contrast between 1D and 2D decoy, we take the mean response rates for R and F decoys as the response rates for 1D decoy in the following analyses. Similar to what we see in Wedell (1991) results (Figure 2.3a), we can observe fairly clear attraction effects across subjects in this experiment (Figure 2.3c) as well as the Wedell (1991) replication study (Figure 2.3b).



(a) Economic gambles Wedell (1991)

(b) Economic gambles (N=150)

(c) Ethical decisions (N=51)

Figure 2.3: Attraction effect across subjects
(a). Attraction effect in Wedell(1991) with economic gambles; (b). Attraction effect in our replication study of Wedell (1991) with economic gambles; (c). Attraction effect in Wedell-isomorphic ethical dilemmas in Experiment 1.

To gain insights on how participants choose the alternatives and to visualize the results, we also show the proportion of within-subject choice reversal for 1D and 2D decoy. In our descriptive analysis, we coded the response data in terms of each participant's choice patterns for each pair of questions (Table A.5) for a clearer comparison

between our results and the original ones in Wedell (1991). For this analysis, in each question, we focused on the pairs presented to subjects and coded the four types of response patterns: 1) choosing the same option for both questions in each pair (both As or both Bs); 2) choosing targets for both questions in each pair (exhibiting a clear choice reversal); 3) choosing opposite non-target options for both questions (i.e., B when decoy is at A and A when decoy is at B); and 4) choosing the decoy at least once.

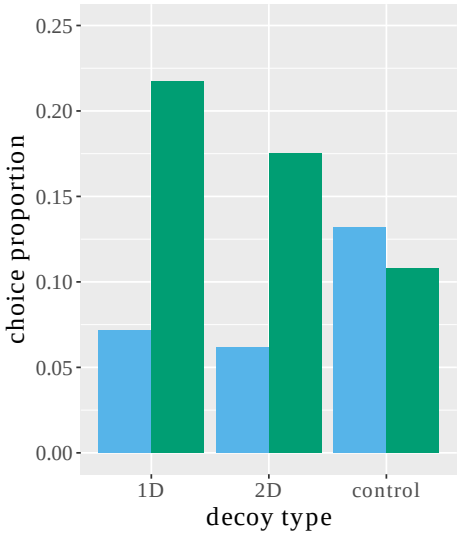
Decoy selection rates were low ($< 10\%$) and the majority participants had consistent choice with in each pair regardless of decoy type. Thus, we focus on the target choice reversal rates and competitor reversal selection rates here (Figure 2.4, full results see Appendix A.3). Generally, within-subject choice reversals occur around 10% to 15% (see Table A.6, Appendix A.3). 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 (Figure 2.4a). A clear and strong choice reversal effect can be observed for the 1D decoy type, whereas the preference reversal effect is less strong for the 2D decoy. In the control (R') condition, we observe the lowest choice reversal rates.

2.3.2.2 Bayesian Statistical Analysis

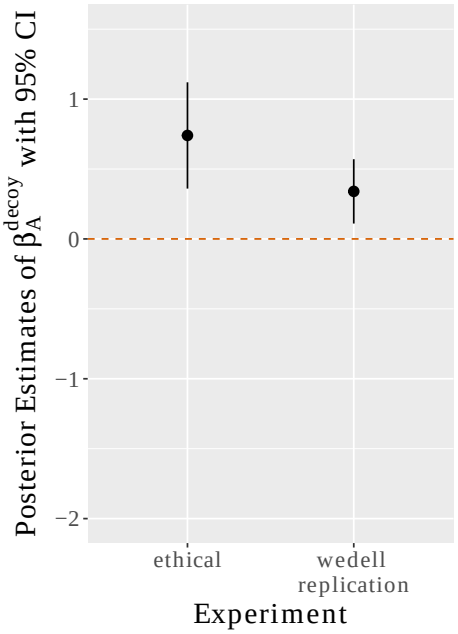
Statistical analysis was conducted in R (R Core Team, 2013) using a Bayesian logistic regression model with RStan (Stan Development Team, 2017).

Statistical Models We used a Bayesian logistic regression model to estimate the main effects and interactions of two predictors: decoy type (1D, 2D, control/R') and decoy position (atA, atB).

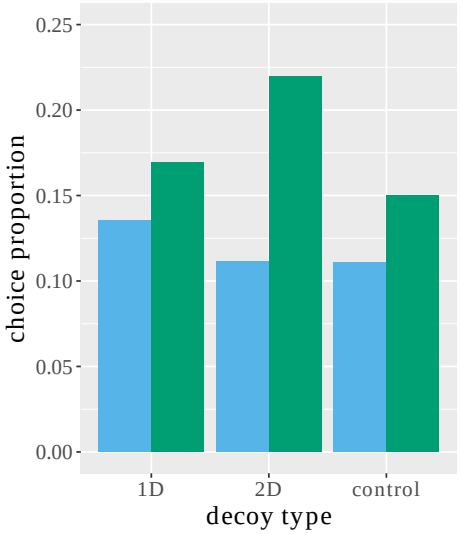
We set choosing B when the one-dimensional (1D) decoy is dominated by B (atB) as the reference category. Thus, we let $P(Y_{ijk} = A)$ denote the probability that the i-th participant's choice is A (where $i \in \{1, 2, ..., N\}$ is the index of subjects, $j \in$



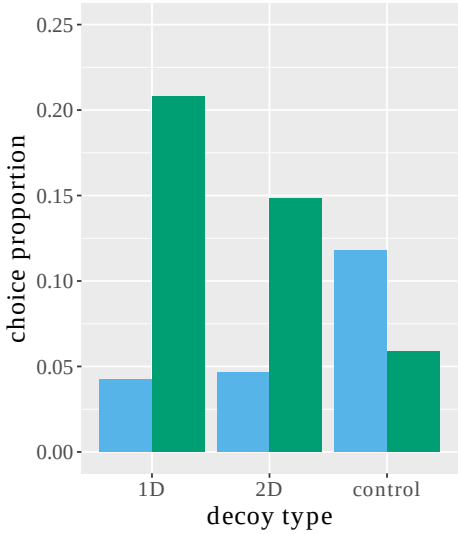
(a) Wedell (1991)



(b) $beta_A^{decoy}$ Posteriors in Wedell (1991) replication and ethical decisions



(c) Wedell (1991) replication (N=150)



(d) Ethical decisions (N=51)

Figure 2.4: Within-subject response patterns.
Within-subject response patterns (competitor reversals & target choice reversals) are shown in (a) for Wedell(1991) original study with economic gambles, (c) for Wedell (1991) replication study and (d) for ethical decisions. The posterior estimates for our main parameter of interest from the Bayesian logistic regression model, β_A^{decoy}, is shown in (b). When this parameter is above zero, it indicates that the logodds of choosing A to choosing B increases when decoy is changed to be dominated by A, suggesting a choice reversal effect.

$\{0,1\}$ is the index of decoy positions, $k \in \{0,1,2\}$ is the index of decoy types, and $\sum_{m\in\{A,B\}} P(Y_{ijk} = m) = 1$).

The full model is given as:

$$\text{logit } P(Y_{ijk} = A) = \beta^{gm} + \beta_A^{type} X_{ik} + \beta_A^{decoy} X_{ij} + \beta_A^{type*decoy} X_{ij} X_{ik}, \tag{2.1}$$

where logit $P(Y_{ijk} = A)$ is the log probability of i-th participant choosing A, and $\boldsymbol{\beta_{gm}}$ estimates the log odds for the baseline category. β_A^{decoy}, our main parameter of interest, estimates the within-subject decoy position effect. Given the baseline, this parameter indicates the change in log odds of choosing A over B when decoy is moved from at B to at A. For example, a β_A^{decoy} of $log(1.3)$ indicates a 30% increase in the log odds of choosing A over B. In other words, a positive β_A^{decoy} indicates within-subject choice reversal.

Parameter β_A^{type} estimates the effects of decoy types on the rate of choice reversals when it is considered together with the interaction term, $\beta_A^{type*decoy}$. Given our model baseline, this parameter, together with the interaction term, informs us whether there is a change in log odds of choosing A over B when decoy is 2D compared to 1D, as the decoy is changed from at B to at A. For example, a coefficient of $log(1.3)$ for the combined two terms indicates a 30% increase in the log odds of choosing A over B.

Besides the full model, we also ran two other models (details in Appendix A.3): the first model (Model 1) was a simple model excluding the control/R' decoy or the interaction between decoy position and decoy type. The model was consistent with the data structure in Wedell (1991)'s original Experiment 1. In the second model (Model 2), we added decoy type control/R' and estimated the interaction between decoy type and decoy position.

Priors Given that we did not have much prior information regarding our model parameters, we chose to select prior distributions that were neither fully informative nor flat. For the βs, we used weakly informative priors: normal$(0, 5)$.

Chain Convergence Evaluation For our full model, we ran four independent chains and each of the four chains contains 2000 samples of each parameter. First 800 samples were part of the *warmup* (or *burn-in* period). This period allowed the sampling process to converge to the posterior distribution, and we analyzed the samples after this period.

Posterior Statistics The posterior estimates for means and 95% credible intervals (CI) for the parameters in all three models are shown in Table A.7, Appendix A.3. For our full model, the central tendencies and 95% credible intervals of the posterior distributions are shown in Figure A.2, Appendix A.3.

All estimates and 95% CIs for our parameter of interest, $beta_A^{decoy}$, are positive, suggesting a clear contextual choice reversal. Particularly, in the full model, the mean estimate for $beta_A^{decoy}$ is 0.74, showing a choice reversal effect (Figure 2.4b). There is 109% increase in log odds of choosing A to B when decoy is moved from B to A. By adding the interaction, we can also see that when decoy is moved from B to A, compared to baseline category (1D), 2D decoy has a reversed effect $(0.12 + (-0.30) = -0.18$, i.e., $\log(.83) = \log(1 - 0.17)$, indicating 17% decrease in log odds of choosing A to B). However, control decoy (R') has a reversed effect $(0.26 + (-0.98) = -0.72$, i.e., $\log 0.49 = \log(1 - 0.51)$, indicating 51% decrease in log odds of choosing A to B).

The results from this model also show that 1D decoy has a stronger effect than 2D decoy (RF), and control/R' decoy has a reversed effect. We will elaborate more on this in the discussion.

2.3.3 A Comparison between Ethical Decisions and Economic Gambles

We conducted an exploratory analysis to compare participants' performances and target reversal rates in economic gambles (i.e., Wedell, 1991 replication, see Appendix A.2) and in ethical decisions (i.e., the Wedell, 1991 isomorphic problems in Experiment 1). All trials with R'/control decoy were excluded from the analyses as these decoys were not expected to produce choice reversals in the first place.

The ethical decisions in Experiment 1 are essentially ethically-significant gambles. This allows us to measure task performance based on expected utility. Each participant's performance is measured by the mean expected value (EU) of their choices throughout the experiment. We found that participants in Experiment 1 (ethical decisions;$N = 51$, $M = 9.96$,$SD = 0.06$) performed significantly better compared to participants in economic gambles ($N = 150$, $M = 9.83$,$SD = 0.22$), $t(195.85) = 6.335$, $p = .00$ (Figure 2.5). Participants in Experiment 1 ($N = 51$, $M = 0.20$,$SD = 0.20$) did not differ significantly from participants in economic gambles ($N = 150$, $M = 0.15$,$SD = 0.16$) in terms of overall target reversal rates, $t(72.03) = 1.43$, $p = .157$, despite of having larger mean target reversal rates (Figure 2.6b).

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, Gabry, Ali, & Brilleman, 2020):

$$\begin{aligned} \text{performance} =& \beta^0 + \beta^1 \text{target reversal} + \beta^2 (\text{experiment=1}) \\ &+ \beta^3 \text{target reversal} * (\text{experiment=1}), \end{aligned} \tag{2.2}$$

where "target reversal" is the centered values of an individual's target reversal rate and "experiment" is a grouping variable indicating whether the values correspond to Wedell (1991) replication study (experiment=0, i.e., the reference group) or ethical

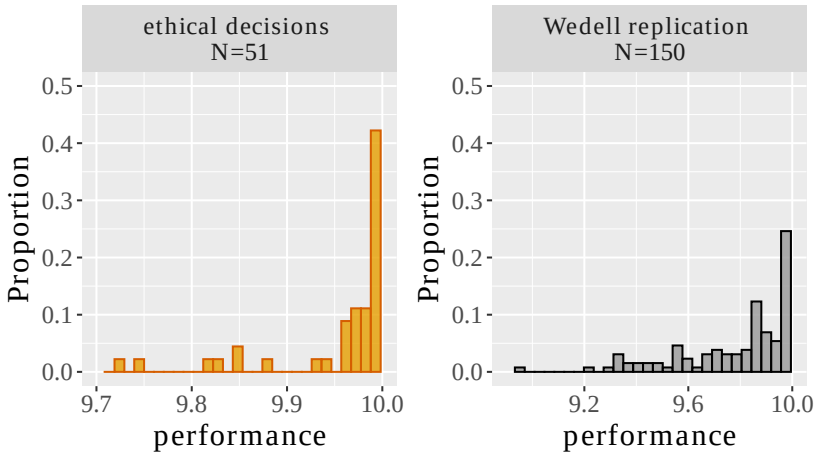


Figure 2.5: Distributions of performances (mean EV) in ethical decisions from Experiment 1 and in economic gambles from Wedell, 1991 replication.

decisions (experiment=1).

We used the default non-informative priors and ran four independent chains (4000 samples each, with the first 200 samples as the *warmup*). The posterior estimates for model parameters and their 95% CIs are fully reported in Table 2.1 below. The predicted mean performance for the reference group, economic gambles, is "$9.78 + 0.60$ target reversal", whereas the predicted mean performance for the ethical decisions group is "$9.78 + 0.60$ target reversal $+ 0.20 - 0.63$ target reversal".

We show individual data points, fitted lines constructed from 500 samples from the total 15200 posterior samples, and fitted lines constructed from medians (of each experiment) of the posterior distributions of model parameters in Figure 2.6a. This result indicates an interaction between type of decision tasks and target reversal rates. In other words, the relationships between performance and target reversal rates are different for ethical decisions and economic gambles. Specifically, in economic gambles,

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β^0 (Intercept)	1.00	17139	9.78	0.02	9.73	9.78	9.82
β^1	1.00	11098	0.60	0.17	0.26	0.60	0.93
β^2	1.00	17954	0.20	0.05	0.10	0.19	0.29
β^3	1.00	12372	-0.63	0.32	-1.26	-0.63	0.01
mean_PPD	1.00	17054	9.83	0.03	9.77	9.82	9.88
log-posterior	1.00	6523	-42.62	1.61	-46.59	-42.29	-40.49
σ	1.00	16642	0.29	0.01	0.27	0.29	0.32

Table 2.1: Experiment 1 — posterior statistics for parameters in the regression model in equation 2.2. Intercept (β^0) represents the effect of the reference group, economic gambles (i.e., Wedell, 1991, replication), on performance.

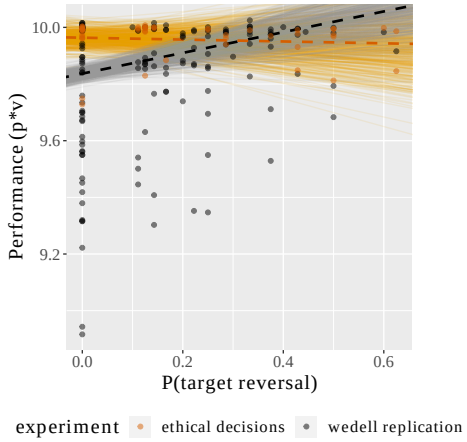
higher reversal rates predict better performances.

2.3.4 Discussion

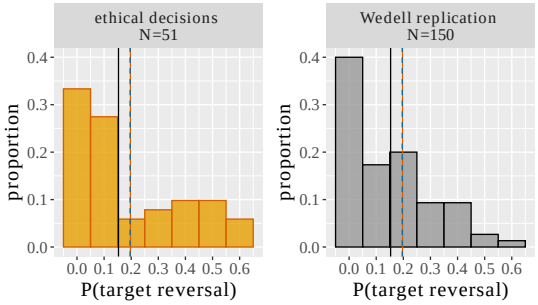
Our study is the first study that investigates and observes contextual choice reversals in the domain of ethical decision making by using rigorously designed tasks with the same structure as classic contextual choice reversal studies. By providing participants with ethical dilemmas that require them to make a trade off between two attributes — probability of saving lives and numbers of lives to save — while choosing among three options: a target, a competitor (both of which have the same expected value), and a decoy option, we found evidence for choice reversals when the decoy was either one-dimensional (R, F) or two-dimensional (RF). One-dimensional decoys had a slightly stronger effect than two-dimensional decoys.

We also compared performance in economic gambles and ethical decisions. The performance in ethical decisions are better, and higher target reversal rates predict better performance in economic gambles but not in ethical decisions. In ethical decisions, the relationship between performance and target reversal rate is difficult to see due to the overall high performance in the ethical decision tasks, which create a ceiling effect.

Although we found evidence for contextual choice reversals in the ethical domain, we acknowledge several drawbacks of this Experiment. First, this task only involves



(a) Relationship between performance and target reversal rate



(b) Distribution of target reversal rates

Figure 2.6: Relationship between performance and target reversal rate and distribution of reversal rates in Experiment 1.
(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 15200 posterior samples and the dashed lines are *medians* from the posterior distributions of model parameters; (b) Distributions of target reversal rates in ethical decisions (Experiment 1) and economic gambles (Wedell, 1991, replication). The lines show overall target reversal rates in two experiments, and the blue dashed line shows the overall target reversal rates in Wedell (1991).

two attributes (the probability of saving lives and numbers of lives that can be saved). Both of these attributes can be measured on continuous scales, making it simple to compare the levels within each attribute (e.g., saving more lives is preferred and higher probability is preferred). Second, this task only involves one scenario — choosing a rescue plan following a natural disaster. In reality, people face various types of ethical dilemmas in all kinds of scenarios — some could be high-level, big 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.

Lastly, the sample size in this study is fairly small (N = 51). This inevitably leads to low precision for estimated posterior distributions of our statistical model parameters. We believe it is necessary to increase sample size in the following studies to achieve higher precision for parameter estimation.

In our following studies, we explore contextual choice reversals in ethical decisions further by creating tasks spanning various ethical domains. These domains include both personal choices such as whether to buy a product from a company that charges more money but provides better health benefits to its employees and choices that are connected to policies that influence groups of people such as choosing a rescue plan.

2.4 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 attributes (probability of saving lives and numbers of lives saved, both of which could be expressed numerically) 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 expand the decision tasks to various domains that involve ethical decisions.

One of the most challenging aspects of designing this experiment is to map the structure of contextual choice reversal tasks to ethical decisions in a greater variety of scenarios (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 (Figure 2.7). If one of the two attributes have only two levels, then the target only dominates the decoy on one attribute. The target and competitor options also need to be equally attractive. The ranking of the levels in each attribute needs to be clear (e.g., the fourth level is always preferred to the first level) among decision makers for the assumed dominance relationship in this structure to hold.

While *number of lives saved* have a clear objective rank as an attribute (i.e., larger numbers of lives saved are better), some specific ethical scenarios require us to use attributes whose levels do not have clear objective ranks. 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

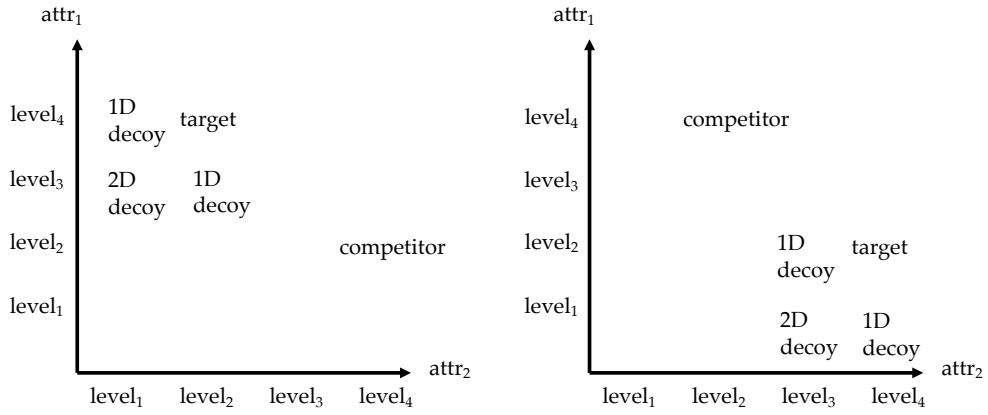


Figure 2.7: The structure for a pair of dilemmas that have the structure of a classic contextual choice reversal task.
If both attributes have four levels, then we can create two-dimensional decoys. If one of the attributes have two levels, we can create a pair of dilemmas with one-dimensional decoys. 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 one-dimensional.

save a loved 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. Thus, we need to gather data to know that, given the sample we have, how the levels in an attributes should be ranked.

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 2.2.

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

Table 2.2: Experiment materials for constructing ethical dilemmas: attributes and their four levels.

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 A.4.2.

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 par-

ticipants 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.

2.4.1 Method

2.4.1.1 Participants

After completing a power analysis (Appendix A.4.1), 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. We included the 475 participants (242 female; age M(SD) = 33(12.19) years) who completed all three parts in the data analyses.

2.4.1.2 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 (1D vs. 2D) and the within-subject variable is decoy position (atA, atB). We have 9 items/scenarios in total. The scenarios and their two attributes/dimensions are shown in Table 2.3 below. In the following section, we provide a brief ethical content analysis of the scenarios. The complete descriptions of the scenarios are in Table A.8, Appendix A.4.3. 7 scenarios (*emergency delivery, jail overcrowding, rescue plan, rescue a survivor, firing an employee, worker welfare, worker welfare 2*) have both 1D and 2D decoys

whereas 2 (*jail overcrowding 2, inevitable injury*) have only 1D decoys. Thus, as each participant sees all 9 items, decoy type is only between-subject for the 7 scenarios that have both 1D and 2D decoys. Each item is a pair of questions: in one question, decoy is at A, and in another, at B.

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 recommitting 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 2.3: The nine scenarios we created for Experiment 2 and their two attributes/dimensions.

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 la-

bor); 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.

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.

2.4.1.3 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.

Part 1 of the experiment contains 16 sets of decision tasks. Each set contains four questions and each question is a decision task with three multi-attribute options. One attribute is the same among the four options and another attribute varies on four different levels. Each participant was randomly presented with one question from each set. In total, each participants completed 16 decision tasks in the first part. We show an example of a question in Table 2.4. In this example, the attribute pollution is the same among the options whereas the attribute speed varies. An example of the full set of questions corresponding to the speed vs. pollution dilemma is given in Appendix A.4.4.

Materials of Part 2 and 3 of the experiment contain the 9 items in pairs. Each part has 9 two-attribute multiple choice decision tasks. To manipulate decoy position and decoy type, we created four different versions of the tasks in Part 2 and 3. The task versions of Part 2 and Part 3 are presented in Table 2.5 below.

Decision problem:	
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?	
Options:	
Pollution	Emergency Delivery Speed
Produces a low amount of pollutants	Delivers overnight
Produces a low amount of pollutants	Delivers in 3 days
Produces a low amount of pollutants	Delivers in 5 days

Table 2.4: An example of a question with the attributes pollution and emergency delivery speed.

Part	Version	Decoy Position	Decoy Type
2	1	atA	1D
	2	atA	2D
	3	atB	1D
	4	atB	2D
3	1	atB	1D
	2	atB	2D
	3	atA	1D
	4	atA	2D

Table 2.5: The four task versions in Part 2 and 3 of Experiment 2.

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. In other words, for the 7 items that have both 1D and 2D decoys, each participant saw either 1D or 2D version of the items but not both. For each item, each participant saw both when the decoy was "atA" and when the decoy was "atB".

All decision tasks were implemented in Qualtrics software (Qualtrics, Provo, UT). The assignment was done as participants signed up for one out of four separate 3-part

studies on Prolific (www.prolific.co). All participants received the same link to the Part 1 questionnaire, and each participant received corresponding links to Part 2 and Part 3 questionnaires depending on which study they signed up for.

2.4.2 Results

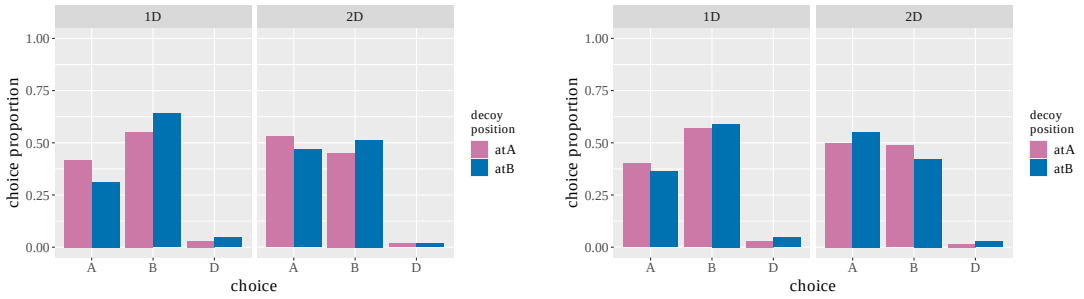
2.4.2.1 Descriptive Analysis

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.

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 (1D decoy: .45; 2D 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 Appendix A.4.5. We also focus on the aggregated data across items. For the complete descriptive results by each item, see Appendix A.4.5.

As in Experiment 1, we present the attraction effect across subjects first. 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 (Figure 2.8). 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.

We then analyze the within-subject choice reversals. Given how choices are made within each pair of dilemmas presented to participants, we code four types of response



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

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

Figure 2.8: Aggregated choice proportions during the first and second session in Experiment 2 (N=475).
We observe an attraction effect across subjects in the first session, but the effect is not present in the second session.

patterns: 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 clear choice reversal); 3) choosing competitors for both questions in each pair (competitor reversal); 4) choosing decoy at least once. Again, 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 Figure A.4, Appendix A.4.5.

As the aggregated data show, choice reversal rates (1D decoy: 0.22; 2D decoy: 0.23) in Experiment 2 are higher than those in Wedell (1991). However, the rates for decoy selection (1D decoy: 0.07; 2D decoy: 0.04) and competitor reversals (1D decoy: 0.15; 2D decoy: 0.23) are slightly higher than those in Wedell (1991) as well. We do not observe a different between 1D and 2D decoys.

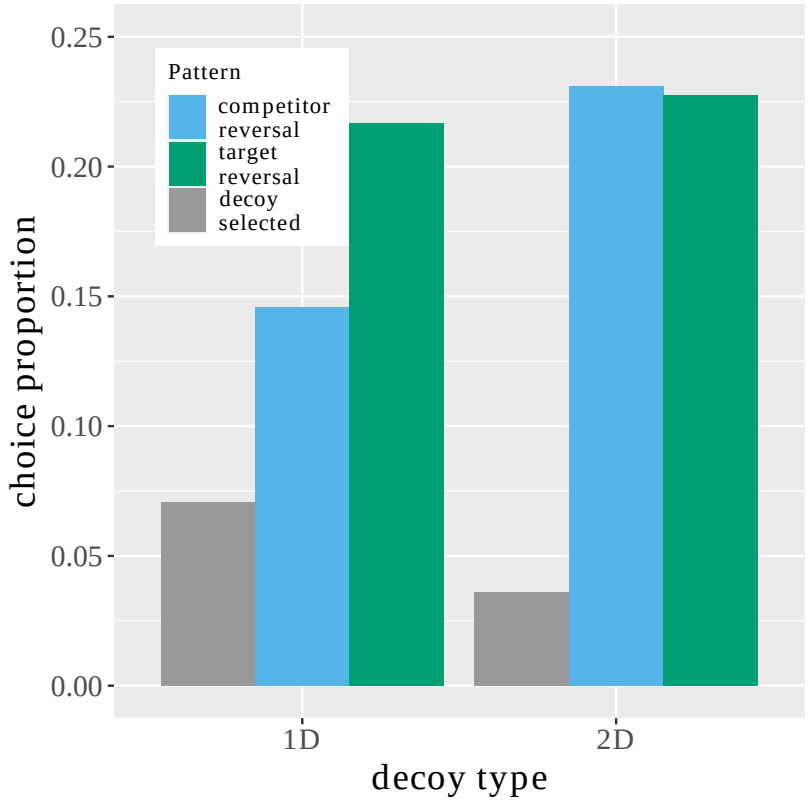


Figure 2.9: Response patterns (competitor reversals & target/choice reversals) aggregated over all items in Experiment 2 (N=475).

2.4.2.2 Bayesian Statistical Analysis

Data analysis was conducted in R (R Core Team, 2013) using a Bayesian multinomial logistic regression model with RStan (Stan Development Team, 2017).

Statistical Models The model setup is very similar to that in Experiment 1. We used a Bayesian multinomial logistic regression model to estimate the decoy type effect (1D vs. 2D), the decoy position effect (atA, atB), and their interactions.

We set choosing B, decoy dominated by B, and decoy being 1D as the reference category. Let $m \in \{1, 2, 3\}$ denote the responses D, A, and B. Let $P(Y_{ijk} = m)$ denote the probability that the i-th participant's choice falls in the m-th category ($i \in \{1, 2, ..., N\}$ is the index of subjects, $j \in \{0, 1\}$ is the index of decoy positions. $k \in \{0, 1\}$ is the index of decoy types). $\sum_m P(Y_{ijk} = m) = 1$. We set choosing B as a baseline and use the logistic regression to compute the log probability of i-th participant choosing m.

We used two models to analyze the data from this experiment. The full model estimates the decoy position and type effects, the interaction between decoy type and position, as well as an item variance:

$$\text{categorical } P(Y_{ijk} = m) = \beta_m^{gm} + w_{item[i]} + \beta_m^{type} X_{ik} + \beta_m^{decoy} X_{ij} + \beta_m^{type*decoy} X_{ij} X_{ik}, \tag{2.3}$$

$$\beta \sim \mathcal{N}(0, 5), \tag{2.4}$$

$$w \sim \mathcal{N}(0, \sigma_w), \tag{2.5}$$

where categorical $P(Y_{ijk} = m)$ is the log probability of i-th participant choosing m, and β_m^{gm} estimates the log odds of choosing m to the baseline. When $m = A$, our main parameter of interest, β_A^{decoy} estimates the within-subject decoy position effect — the change in log odds of choosing A over B when decoy is moved from at B to at

A. Parameter β_A^{type} estimates the effects of decoy types on the rate of choice reversals when it is considered together with the interaction term, $\beta_A^{type*decoy}$.

Besides the full model, we also ran a simple model without item variance on each of the item separately:

$$\text{categorical } P(Y_{ijk} = m) = \beta_m^{gm} + \beta_m^{type} X_{ik} + \beta_m^{decoy} X_{ij} + \beta_m^{type*decoy} X_{ij} X_{ik} \tag{2.6}$$

$$\beta \sim \mathcal{N}(0, 5) \tag{2.7}$$

Priors We used weakly informative priors for all βs: normal(0,5); the prior for item variance was normal(0,1).

Chain Convergence Evaluation For our full model, we ran four independent chains and each of the four chains contains 4000 samples of each parameter. First 800 samples were part of the warmup (or burn-in period). This period allowed the sampling process to converge to the posterior distribution, and we analyzed the samples after this period. To interpret posterior distributions more cautiously and accurately, we first checked chain convergence through traceplots and Rhat (Sorensen, Hohenstein, & Vasishth, 2016). The traceplots for parameters (not included) suggest that the chains have converged for all parameters. The Rhat values from Table A.9 in Appendix A.4.5, have shown convergence as well.

Posterior Statistics The posterior estimates for means and 95% credible intervals (CI) for the parameters in the full model are shown in Table A.9 and Figure A.9 in Appendix A.4.5. Full results for the simple model applied to separate items are also included in Appendix A.4.5.

The posteriors for our main parameter of interest, β_A^{decoy}, for the aggregated data with the full model as well as for each scenario with the simple model are shown in

Figure 2.10.

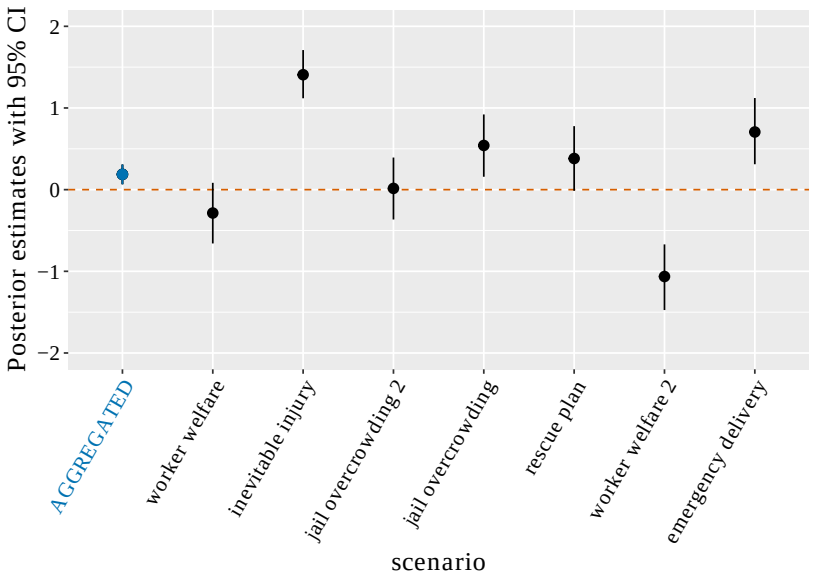


Figure 2.10: Posterior estimates for means and 95% CIs of the main parameter of interest, β_A^{decoy}, for the full model (aggregated data) and for each scenario analyzed with the simpler model.

For the aggregated data, the estimated mean for the parameter β_A^{decoy} is 0.29 and its 95% CI is entirely above 0 (95% CI = [0.16, 0.41]), suggesting that we have a clear overall choice reversal effect in aggregate. By adding an interaction between decoy position and decoy type, we also see that contrary to what we expected, when decoy is moved from B to A, compared to the baseline category (1D), 2D decoys have a slightly stronger effect ($0.69 - 0.28 = 0.41$).

Results from the simple model applied to individual items provide us with more insights on the variations among items. For items without 2D decoy, we simply applied the model to those individual items without β_m^{type} and interaction terms (i.e., categorical $P(Y_{ijk} = m) = \beta_m^{gm} + \beta_m^{decoy} X_{ij}$). We observe very clear choice reversals for four items: *inevitable injury, jail overcrowding, rescue plan*, and *emergency delivery.*

2.4.3 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 preference reversals in the *rescue plan* item, an item taken directly from Experiment 1 and isomorphic to Wedell (1991) tasks. This result replicates Experiment 1. Other items that produced clear contextual choice reversals include *inevitable injury, jail overcrowding, emergency delivery.* In the following discussion, we put more focus on the items that did not produce contextual choice reversals and consider why they didn't work as we expected. The between-subject choice proportions for each item (Figure A.8 in Appendix A.4.5) can provide us with insights.

Using probability as a guide. For the items that involve probability as an attribute, people often use probability as a guide. And we often observe a risk preference effect where people prefer higher probabilities. Item *rescue a survivor* has extremely high decoy selection rate in both 1D-decoy version and 2D-decoy version, especially for

the decoy dominated by option B (Figure A.8a). Such high decoy selection rate (as high as target options) indicates that there is overwhelming preference for the decoy option when the decoy is dominated by B. In this item, people make a trade-off between age of people being rescued and probability of rescuing. When the decoy is dominated by option B, the decoy and target have the same age attribute (65-years-old) and differ in probability (55% and 60%) whereas the competitor shows a 40% probability of rescuing a 35-year-old person. The overwhelming decoy selection rate could be due to a strong preference for high probability in this context. The same general preference for higher probability of rescue can even be seen in the dilemma that produced preference reversals as well — *rescue plan* (Figure A.8f). Similarly, in the *inevitable injury* item (Figure A.8b), we could see a general preference for low probability of injury.

Unbalanced target and competitor. 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; Figure A.8d). 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 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 could be observed in *worker welfare* and *worker welfare 2* (Figure A.8 c & g). 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 aggregate and in some items in Study 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.

2.5 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 1D or 2D version of each item to each participant.

3. **Manipulation of task instructions.** 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.

2.5.1 Method

2.5.1.1 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.

2.5.1.2 Materials

This experiment follows a 2 × 2 × 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 (1D vs. 2D) and decoy position (atA, atB). We have 8 critical items/scenarios (Table 2.6) and 16 filler items/scenarios. The specific descriptions of the critical items can be found in Table A.8, Appendix A.4.3. Among the 8 critical items, 6 items (*emergency delivery, jail overcrowding, rescue plan, firing an employee, worker welfare, worker welfare 2*) have both 1D and 2D decoys whereas 2 items (*jail overcrowding 2, inevitable injury*) have only 1D 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 will see all 24 items (48 questions).

Attention Check. 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

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 recommitting 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 2.6: Critical scenarios in Experiment 3.

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 began to learn English, language used mostly at home, and highest grade completed.

2.5.1.3 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. This part contains 30 sets of decision tasks. Each set contains four questions and each question

is a decision task with three multi-attribute options. One attribute is the same among the four options and another attribute varies on four different levels. Each participant was randomly presented with one question from each set. The attributes in this part includes both attributes that appear in critical items and attributes that appear in fillers. In total, each participant completed 30 multiple-choice decision tasks in this part.

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 created two different versions of the tasks in Part 2, presented in Table 2.7 below.

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

Table 2.7: The two task versions in Part 2 of Experiment 3.

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 and 2 differ in decoy position and filler versions. When each item is presented to a participant *for the first time*, its 1D or 2D decoy version is presented randomly. If an item's 1D decoy version is presented in block 1, then its 1D 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).

2.5.2 Results

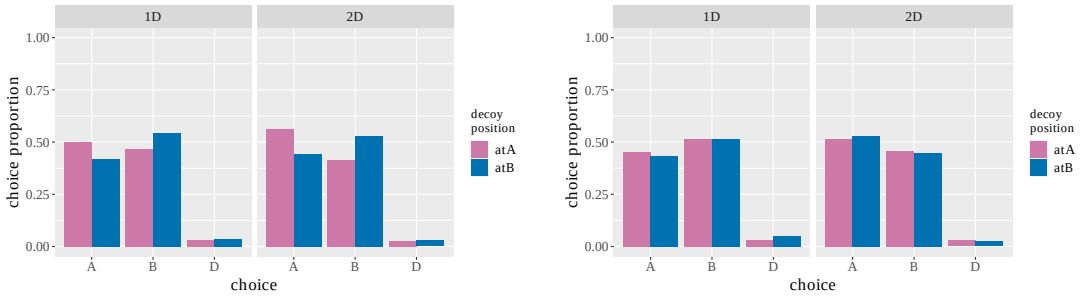
2.5.2.1 Descriptive Analysis

Part 1 results (not included) 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.

In the following analyses, we excluded the *firing an employee* item due to its distinctively high decoy selection rate (1D decoy: .26; 2D 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 Appendix A.4.6. We also focus on the aggregated data across items. The complete descriptive results by each item are in Appendix A.4.6.

As in previous experiments, we first report the attraction effect across subjects (Figure 2.11). 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 2, we observe 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.



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

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

Figure 2.11: 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.

We also show the proportions of choice patterns in Experiment 3 in Figure 2.12 below, excluding rates for consistent choices (see Figure A.11, Appendix A.4.6 for full results). The consistent-selection rates are extremely high (1D decoy: .78; 2D decoy: .82), indicating that participants mostly chose consistently between the pair of questions corresponding to the same item. The rates for contextual choice reversals (1D decoy: .10; 2D decoy: .09) and competitor reversal selections (1D decoy: .05 ; 2D decoy: .04) are fairly low from Experiment 3.

2.5.2.2 Bayesian Statistical Analysis

Data analysis was conducted in R (R Core Team, 2013) using a Bayesian logistic regression model with RStan (Stan Development Team, 2017).

Statistical Models, & Priors: see Experiment 2. We did not find any effect of whether subjects received instruction or not with the full analysis model (Appendix A.4.6.2). Thus, we collapsed the with-instruction and no-instruction group together,

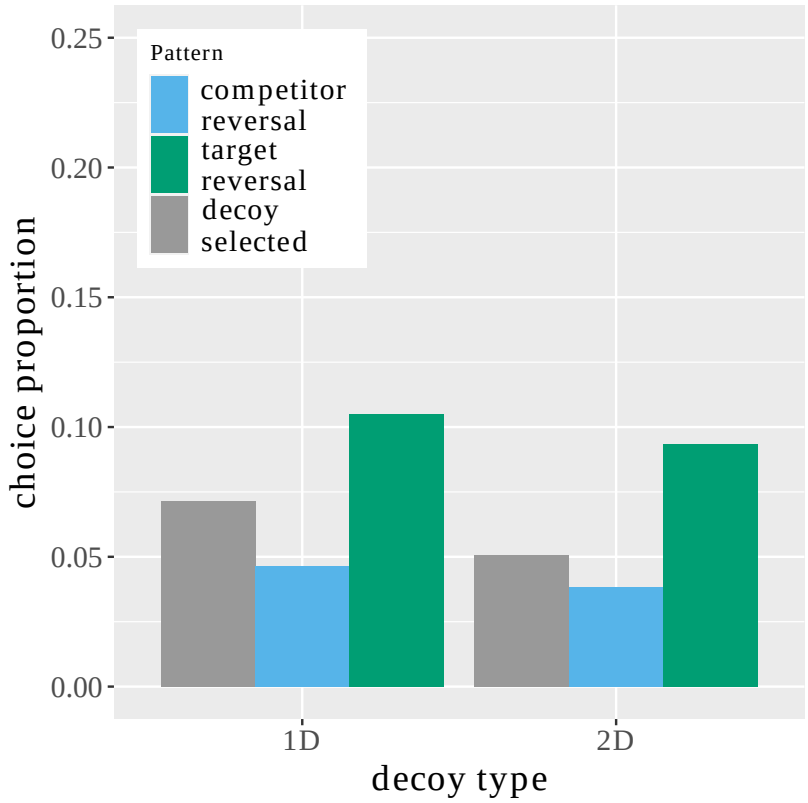


Figure 2.12: Aggregated response patterns for all items (excluding *firing an employee*) in Experiment 3 (N = 456).

and used the full model from Experiment 2. We have also applied the simple model from Experiment 2 to each item separately.

Chain Convergence Evaluation For our full model, we ran four independent chains and each of the four chains contains 4000 samples of each parameter. First 800 samples were part of the warmup (or burn-in period). This period allowed the sampling process to converge to the posterior distribution, and we analyzed the samples after this period. The traceplots for parameters (not included) suggest that the chains have converged for all parameters and the Rhat values from Table A.12 in Appendix A.4.6 have shown convergence as well.

Posterior Statistics The posterior estimates for mean and 95% CIs for the parameters in the full model and the simple model are shown in Table A.12 and Figure A.16 in Appendix A.4.6. Full results for the simple model applied to separate items are included in Appendix A.4.6.

The posteriors for our main parameter of interest, β_A^{decoy}, for the aggregated data analyzed with the full model and for each item analyzed with the simple model are shown in Figure 2.13 below.

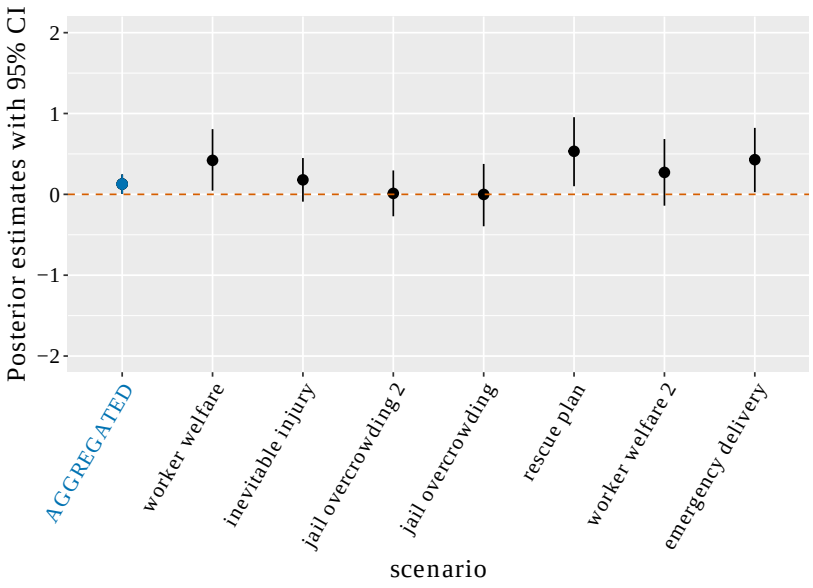


Figure 2.13: Posterior estimates for means and 95% CIs of the main parameter of interest, β_A^{decoy}, for the full model (aggregated data) and for each scenario analyzed with the simple model.

For the aggregated data, the estimated mean for β_A^{decoy} is 0.19 and its 95% CI is entirely above 0 (95% CI = [0.07, 0.31]), suggesting that we have a clear overall within-subject choice reversal effect. By adding an interaction between decoy position and decoy type, we can also see that results suggest that when decoy is moved from B to A, compared to the baseline category (1D), 2D decoys have a slightly stronger effect (0.22 + 0.02 = 0.24). However, the effect of decoy type remains inconclusive as the 95% CI for the estimated interaction is ambiguous ($\beta_A^{type*decoy}$ 95% CI = [-0.19, 0.23]).

Similar to Experiment 2, results for individual items (Figure 2.13) show that choice reversals are present for some items (*worker welfare, worker welfare 2, inevitable injury, rescue plan, emergency delivery*), but not present in *jail overcrowding* item and *jail overcrowding 2* item.

2.5.3 Discussion

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

This study 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.

Similar to Experiment 2, we found evidence for contextual choice reversals in aggregate. But we also again found item variations — we did not observe contextual choice reversals in all items. In addition, we also found very high consistent selection rates. This could 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 1D and 2D 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.

We hope to explore further how item variations could be affected by individual participants' ranking of levels in the attributes. As discussed in Experiment 2, we

constructed the items based on the study where we found attributes with levels that majority of participants had a preferred ranking. This suggests that there still remains individual variations in terms of preferred ranking for the levels in attributes.

2.6 Combined Analyses

To appropriately combine the items and compare results from Experiment 2 and Experiment 3, we present the following combined analyses.

First, we present the results of between-subject attraction effect. Table 2.14 below contains the between-subject 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 Appendix A.4.7). 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.

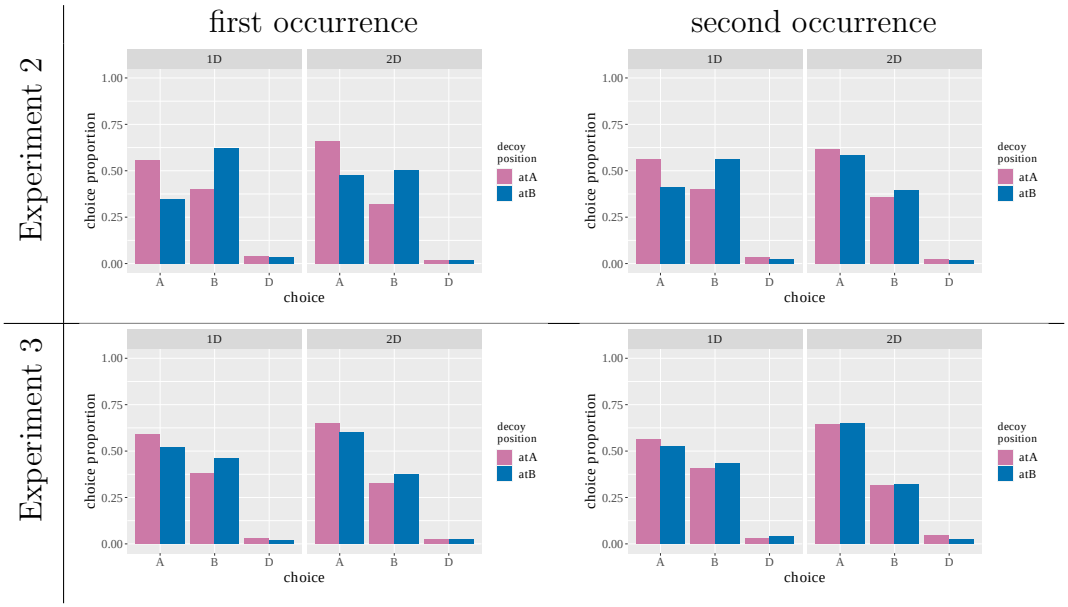


Figure 2.14: Attraction effect across subjects shown as aggregated choice proportions for first and second occurrences of shared items (*emergency delivery, jail overcrowding, inevitable injury, rescue plan*) in Experiment 2 (N=475) and 3 (N=456).

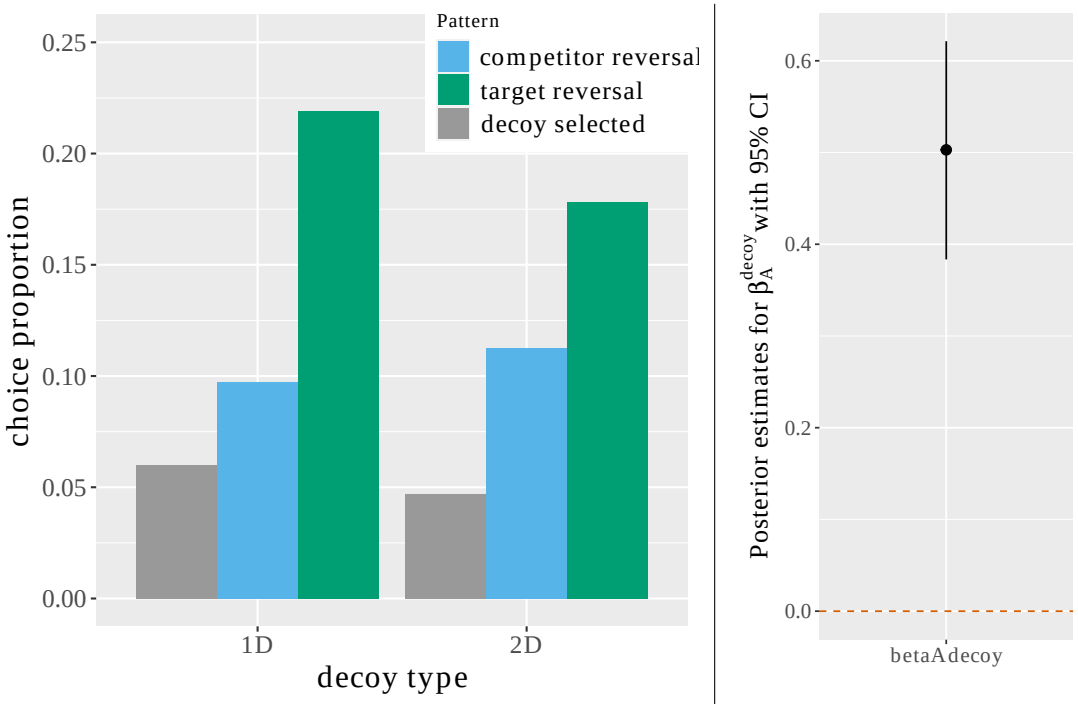
Besides the attraction effect, we also show the choice patterns (excluding consistent-selection rate) of the four shared items in aggregate (N=931) below in Figure 2.15a, where the rates of choice reversals in 1D and 2D decoy conditions are quite high. We ran the full model (Equation 2.3) on the combined data from the shared items in Experiment 2 and 3. The results suggest a clear aggregated choice reversal (Figure 2.15b, β_A^{decoy} mean: 0.50, 95% CI = [0.38,0.62]). The estimate of the interaction between decoy type and decoy position is extremely small (β_A^{int} mean: -0.21, 95% CI = [-0.41, -0.01]). Combined with the estimate of decoy type, results show a slightly stronger effect for 2D decoy compared to baseline category (0.51-0.21=0.30). Full descriptive and statistical results for the aggregated data can be found in Figure A.19, Appendix A.4.7.

We also present the results of attraction effects across subjects for the revised scenarios (*worker welfare 2, worker welfare, jail overcrowding 2*, Figure 2.16). We observe an attraction effect in Experiment 3 after the revision of items, especially in the first occurrences of items.

We have also investigated how much time on average participants spent on a question during each session in Experiment 2 and Experiment 3 (Table 2.8). 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.

2.7 Summary

To explore the contextual choice reversals in the ethical domain, we created various ethical dilemmas that have the strict structure of multiple-choice problems with two distinctive attributes. These tasks differ from previous studies of ethical decisions not only in their underlying structure, but also in their content: we moved away from



(a) Combined analysis: response patterns for shared items in Experiment 2 & 3.

(b) Combined analysis: posterior estimate for β_A^{decoy}.

Figure 2.15: Descriptive and statistical results for shared items in Experiment 2 and 3. Data from these items from Experiment 2 and 3 are combined (N=931).
(a). Response patterns aggregated over shared items in Experiment 2 and 3 from both experiments. (b). Posteriors from the full model analysis of these items from Experiment 2 and 3 for β_A^{decoy}. This result indicates that within-subject choice reversals are present.

trolley problems as an attempt to ground the decisions in more realistic scenarios and to address more specific concerns.

In Experiment 1, we focused on choices in ethical dilemmas under a single scenario and isomorphic to Wedell (1991) tasks. Through these tasks, we found very clear choice reversals. As the ethical dilemmas in Experiment 1 are essentially ethically-significant gambles, we were able to calculate task performance based on expected utility and compare the performance in ethical decisions and economic gambles. We found that general performances in ethical decisions were better than those in economic

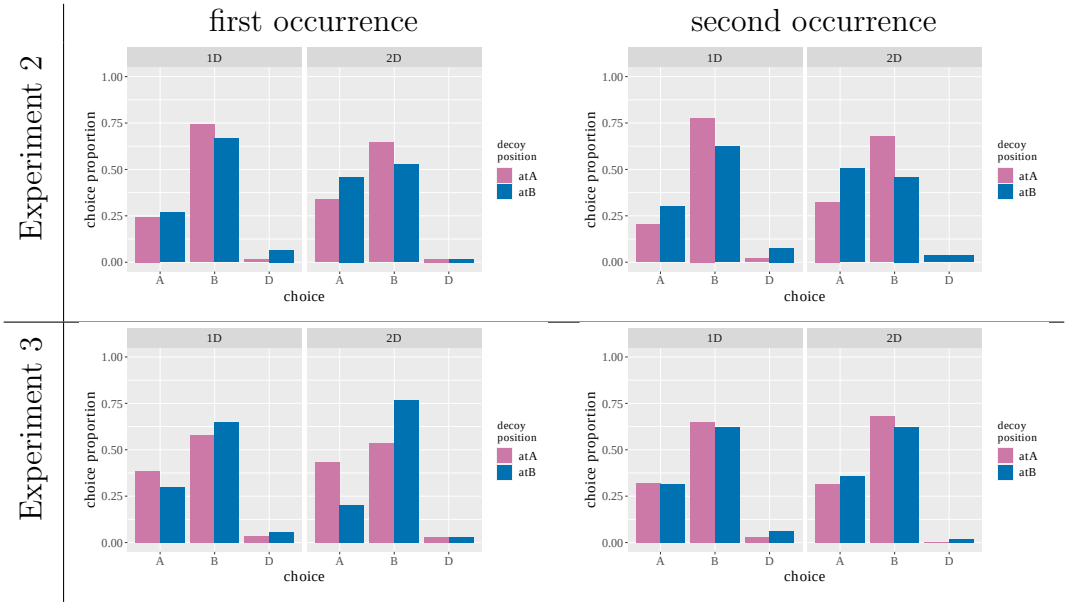


Figure 2.16: Attraction effect across subjects shown as aggregated choice proportions for first and second occurrences of revised items (*worker welfare 2, worker welfare, jail overcrowding 2*) in Experiment 2 (N=475) and 3 (N=456).

Experiment 2	session 1	23.47s
	session 2	20.97s
Experiment 3	critical items	20.59s
	first-seen critical	24.55s
	second-seen critical	16.58s

Table 2.8: Average time spent per item

gambles. Higher target reversal rates predict better performances in economic gambles but not in ethical decisions. In Experiment 2, we included ethical scenarios that involve qualitative attributes, where we found choice reversals but also variations of reversal rates among different ethical scenarios. In Experiment 3, we made an attempt to revise the scenarios, improved the randomization procedure on Experiment 2, and added fillers. Similar to the results in Experiment 2, we found choice reversals in aggregate but item variations still remained. Through all three experiments, the effect of 1D versus 2D decoys remained inconclusive.

Given that Experiment 2 and 3 had 4 shared items and 4 different items, we con-

ducted further analyses to compare the results from Experiment 2 and 3. Generally, we observed smaller effects of choice reversals in Experiment 3 in aggregate, especially among the 4 shared items. This could be partially due to that participants completed the pairs of items in separate sessions during Experiment 2 but in the same session in Experiment 3. In other words, participants remembered the items well — even with 16 fillers added in Experiment 3. This is also consistent with the result that the rates of choosing the same options within a pair were extremely high in Experiment 3. The time participants spent on these decisions also provide some insight into the lowered rates of choice reversals in Experiment 3, as some studies have suggested that choice reversals tend to diminish as time pressure increases (Pettibone, 2012).

In our experiments, the time participants spent on each item on average in Experiment 2 was longer than the time participants spent on critical items on average in Experiment 3, consistent with previous studies' findings. However, we acknowledge that the time recorded by us via Qualtrics software (Qualtrics, Provo, UT) was the time participants clicked the last time on a choice and thus included at least their reading time and cannot be viewed as equal to decision time. Among the four unique items in Experiment 2 and 3, we could see a clear change: contextual choice reversals were found in the changed items after we revised them according to our analyses in Experiment 2.

In Experiment 3, we found that variations in different items still remained. This suggests that it is necessary to further investigate how individual differences may affect the variations of choice reversals among items.

2.8 Discussion

This work, for the first time, demonstrates contextual choice reversals in the ethical domain with the common paradigm used in studies of contextual choice reversals —

first with tasks isomorphic to economic gamble tasks, then with tasks spanning multiple ethical scenarios with qualitative attributes.

In general, these reversals are seen as violations of human rationality in expected utility maximization. Based on rational choice theory, a rational choice should maximize expected utility. In a task such as Wedell (1991), one's preference shouldn't change when the decoy is switched from being dominated by A to dominated by B. This is because that the selection between target and competitor is a matter of risk preference, such as the observed preference for higher probability options among some participants. Similarly, in a choice reversal task that involves ethical features, as long as the basic ranking assumption (i.e., decision makers have the same rankings for levels in attributes) holds, one's expressed preference shouldn't change when the decoy position changes. Here we discuss a perspective allowing us to bring contextual choice reversals in ethical decisions into contact with the framework of bounded rationality (Simon, 1955).

In the bounded rationality framework (Simon, 1955), a decision making agent is not absolutely rational by definition based on traditional economic theory, but is "approximately rational" or at least "intends to be rational" under the influence of the environment and the choosing agent's limited knowledge and ability.

An account on the bounded rationality of contextual choice reversals given by a recent model (Howes et al., 2016) shows that these reversals are inevitable when people maximize expected values given the assumed perceptual and cognitive bounds. Specifically, the model combines two noisy observations of each option: the first one is a noisy ordinal observation of its attributes (i.e., partial orderings for each attribute) and the second one is a noisy calculation of its subjective expected utility. Given these two noisy observations, the model makes a choice that maximizes expected value and predicts choice reversals. This suggests that the types of systematic choice reversals observed in human choice are a signature of boundedly rational expected utility maximization.

Given that we observed contextual choice reversals in ethical decision tasks that have the same structure as a classic choice reversal task (Wedell, 1991), we can extend the same theoretical model to the choice domain of ethical decisions. In other words, contextual choice reversals naturally arise in multi-attribute ethical decisions as the decision maker makes the utility maximizing choice given noisy observations of the attributes in each option and noisy calculations of each option's expected utility based on the decision maker's subjective expected utility. This opens the possibility for rigorous accounts of the bounded rationality of ethical decision making that are consistent with — instead of conflicting with — rational choice theory. It also provides us with the possibility of finding the similarities between the principles under which ethical decision making operates and the principles under which other types of decisions operate.

The boundedly rational account of contextual choice reversals is also consistent with our exploratory finding in Experiment 1 where higher target reversal rates predict better task performances in economic gambles. Although we did not find this relationship between target reversals and performance in Wedell (1991)-isomorphic ethical decisions due to the ceiling effect, we observed that the overall reversal rates in ethical decisions are just as high as those in choices in economic gambles (if not higher). This suggests that we need future investigations to systematically compare the differences between context effects in economic gambles and ethical decisions and provide explanations for these differences.

Finally, this work also has some potential broader implications. We believe that understanding context effects in ethical decisions may help us understand decision making in medical situations and the domain of public policy better, as they often include complex multi-attribute and multi-option decisions.

CHAPTER III

Explaining Variation in Contextual Choice Reversals across Ethical Dilemmas: An Individual Differences Account

Contextual preference reversals, or choice reversals, refer to the change of choice preference between two options in the presence of a third unchosen decoy. They have been studied in various domains (Huber et al., 1982; Wedell, 1991; Trueblood, 2012; Trueblood et al., 2013; O'Curry & Pitts, 1995) in the past. Through our previous empirical experiments (presented in the previous chapter), we have demonstrated that these reversals also occur in ethical decisions. In total, we presented nine scenarios across two experiments: *emergency delivery, jail overcrowding, jail overcrowding 2, inevitable injury, rescue plan, rescue a survivor, firing an employee, worker welfare, worker welfare 2* (Table A.8, Appendix A.4.3). Among these scenarios, five were identical in the two experiments. Despite observing within-subject choice reversals (i.e., choosing target options in a pair of items that have the same scenario, with decoy position switched) in the aggregate in both experiments, we also observed variation among these scenarios. More specifically, the rates of within-subject choice reversals differ significantly across scenarios.

These scenarios include both ones that are isomorphic to economic decisions —

where the decision makers make a tradeoff between probability and some numeric value — and ones that involve qualitative attributes — where the decision makers have to combine the attributes in certain ways to evaluate the options.

As different individuals have different subjective ways to evaluate the attributes, the individuals would have potentially different scales to evaluate the options. 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. Even when the individuals differ in risk preference, or in how probability and value are combined, or in the shape of the subjective value curve, a higher value is still a higher subjective value for everyone (plausibly) as is for probability.

However, for the ethical dilemmas, this is not the case. Individual participants' different rankings for levels in attributes in a scenario could affect our task structure, and further affect the preference reversal rates in a scenario. Consider the example below (illustrated in Figure 3.1).

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.

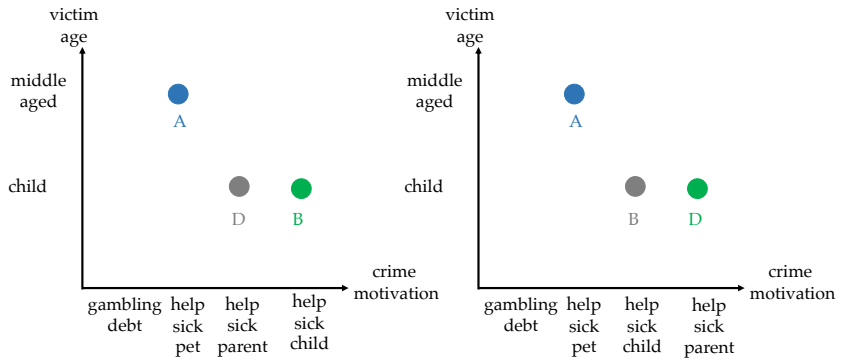


Figure 3.1: Task structures for the *jail overcrowding 2* dilemma according to our assumed ranking for levels in the crime motivation attribute (left) and according to a different yet possible ranking for levels in the crime motivation attribute (right).

In this chapter, to show how the individual differences in ranking of the levels in

the attributes systematically affect the variation in choice reversal rates, we present a generative model that takes into account the individual preferences of feature rankings in different attributes while generating choices for each ethical dilemmas. Our model predictions show that the variation among scenarios can be partly explained by the individual differences in the feature rankings.

3.1 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 2.6). 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. Consequently, what is intended to be a target may become a competitor or even a decoy given some individual's rankings, changing the structure of the decision problem completely. 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 3.1. 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, $\mathbf{p_{consistent}} \in [0, 1]$. 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 a Wedell (1991)-like structure, the model selects the same options in the questions in that pair with $\mathbf{p_{consistent}}$.

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. In the next section, 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.

3.2 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 at-

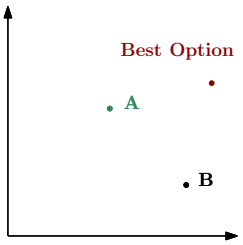
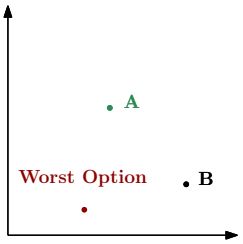
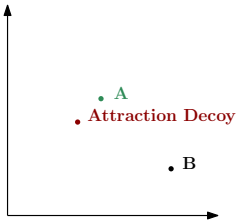
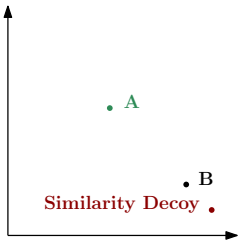
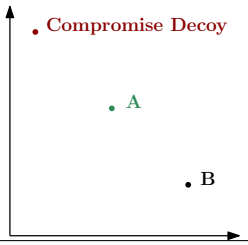
Structure/Configuration	Illustrated example
Best Option Configuration: a best option dominates the other two options in both attributes	
Worst Option Configuration: a worst option is dominated by the other two options in both attributes	
Attraction Configuration: a target, competitor, and decoy can be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

Table 3.1: Possible structures/configurations given all possible feature rankings.

tribute.

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 3.1):

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).

3.3 Generating Choices Given Decision Problems' Structures

Recall that in Experiment 3, each subject sees all 8 scenarios, among which 6 scenarios have both 1D decoy and 2D decoy, and 2 scenarios only have 1D decoy. Thus, each subject randomly sees a total of 16 questions – the 1D-decoy or 2D-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.

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 3.2) 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 3.3).

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 (ϵ)

b. this question has the *Worst Option Configuration*: select worst option with error ϵ

c. this question has the *Similarity Configuration*: sample a choice given the pattern in Trueblood (2012) data (Table 3.4).

d. this question has the *Compromise Configuration*: sample a choice given the pattern in Trueblood (2012) data (Table 3.5). Since we don't 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.

decoy type	pattern	prob
1D	consistent choice	.64
	target reversal	.19
	competitor reversal	.06
	decoy selected	.08
2D	consistent choice	.69
	target reversal	.16
	competitor reversal	.06
	decoy selected	.08

Table 3.2: Distribution of Choice Patterns in Wedell (1991) after re-normalizing.

3.4 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

decoy position	decoy type	option	prob
atA	1D	A	.65
		B	.25
		D	.08
atB	1D	A	.52
		B	.38
		D	.08
atA	2D	A	.66
		B	.24
		D	.08
atB	2D	A	.56
		B	.34
		D	.08

Table 3.3: Marginal Distributions of Choices Wedell (1991) after re-normalizing.

option	prob
decoy	.2
focal option	.5
non-focal option	.3

Table 3.4: Distribution of choices in Trueblood (2012) — similarity effect. Focal option refers to the option that is enhanced by the decoy.

option	prob
compromise	.48
extreme options	.26

Table 3.5: Distribution of choices in Trueblood (2012) — compromise effect.

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 (ϵ) was set to a low and reasonable value (0.05) and the probability of choosing consistently ($\mathbf{p_{consistent}}$) was set to 0.7, as people choose consistently in about 70% trials with an *Attraction Configuration* empirically.

The simulation generally produced the pattern of choice patterns from the empirical data (Table B.1, Appendix B). However, our simulation generally predicted "consistent choice" pattern rates lower than empirical data and "competitor reversal" choice pat-

tern rates higher than empirical data (Figure B.1, Appendix B). The complete results including the *firing an employee* item can be found in Appendix B. We focus on the results of simulated and empirical decoy selection rates and choice reversal rates in Figure 3.2 below.

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 1D decoys than in dilemmas with 2D decoys. Besides reversal rates, the model also predicts decoy selection rates well for both dilemmas with 1D decoys and dilemmas with 2D decoys. This suggests that some item variations can indeed be accounted for by the variations of people's individually varying rankings.

3.5 Discussion

In this chapter, we show how a generative model that takes individual differences in feature rankings can partly account for the variation of choice reversal rates among ethical scenarios.

Our model seems complex, yet it is driven by a simple theoretical idea that individual differences in feature rankings may pose challenges to the structure of multi-attribute choice problems. This model provides some insights into the assumption of the classical choice reversal paradigm and how the structure of multi-attribute choice problems can change given subjective rankings. Classical choice reversal paradigm assumes that decision makers have the same feature rankings on the attributes involved in the tasks. This assumption may not be problematic when the attributes have clear numeric values such as probability, the amount of money to win, or the quality rating of a restaurant. However, this assumption may not hold when the attributes are qualitative and concern different individual ethical values. In our ethical

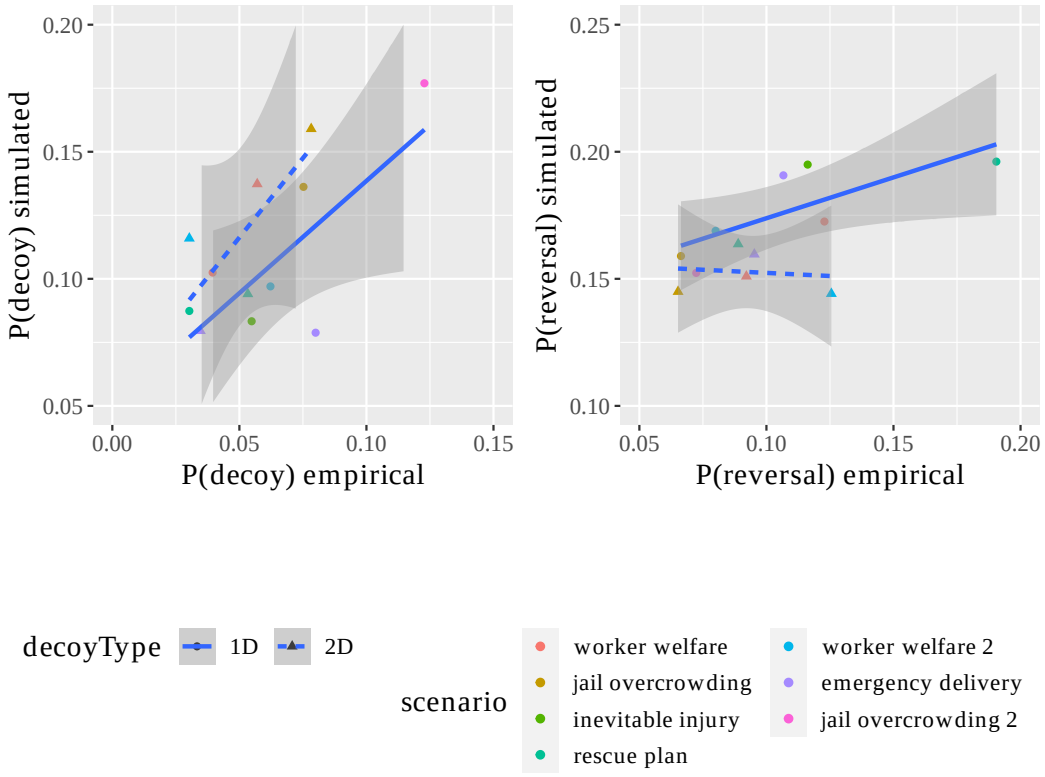


Figure 3.2: Simulated and Experiment 3's empirical decoy selection rates (left) and choice reversal rates (right) for each scenario and decoy type.

decision tasks, individuals' rankings directly affect their subjective representation of the multi-attribute ethical dilemmas — making the structure of the choice reversal tasks different from our intended structure in many cases. As a consequence, a target option may become the decoy option, and a structure with an attraction decoy may become one with a similarity decoy instead.

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.

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.

CHAPTER IV

Explaining Valence Asymmetries in Value Learning: A Reinforcement Learning Account

Learning that certain objects or actions are associated with value can impact many aspects of human behavior, including what we pay attention to, what we desire, and what we learn. To understand how acquired value influences behavior, laboratory tasks have been developed to establish associations between otherwise neutral items and win or loss outcomes. The impact of value on subsequent processing has been examined in a variety of cognitive processing domains such as attention (Della & Chelazzi, 2009; Raymond & O'Brien, 2009), motor control (Painter, Kritikos, & Raymond, 2014), and memory (Aberg, Müller, & Schwartz, 2017). We focus here on a task which we refer to as the *Value Learning Task* (VLT; Raymond & O'Brien, 2009). The VLT has been used to examine how learned value impacts the cognitive processing (e.g. visual attention, perceptual and motor processing) of stimuli that were previously associated with wins or losses that occurred with low or high probability.

Research adopting the VLT has largely focused on examining the cognitive processing of stimuli previously associated with wins or losses, and not the learning itself or possible valence asymmetries in the learning. But a recent meta-analysis of several VLT experiments (Lin et al., 2020) provides evidence that people learn win associa-

tions better than loss associations. Furthermore, in two new empirical studies, Lin et al. (2020) demonstrated that this learning asymmetry was evident with both monetary earnings and non-monetary points, and was evident regardless of whether participants received explicit instructions about the outcome contingencies. However, the underlying cognitive basis for this learning asymmetry remains unclear. The aim of the present computational study is to provide a clear explanation of the observed asymmetries. We next describe the VLT and key empirical findings in more detail before introducing the computational learning model.

4.1 The Value Learning Task

The Value Learning Task (VLT) involves a choice game where a pair of images is presented on each trial, and participants select one image from each pair, receiving a probabilistic positive, negative, or zero reward as feedback. The participants' goal is to maximize earnings (points or money) by learning and exploiting the expected value of each stimulus.

Condition	Stimulus	Outcomes and Probabilities	Expected Value
Win pair	A	+5 (p = 0.8), 0 (p = 0.2)	4
Win pair	B	0 (p = 0.8), +5 (p = 0.2)	1
Loss pair	C	-5 (p = 0.8), 0 (p = 0.2)	-4
Loss pair	D	0 (p = 0.8), -5 (p = 0.2)	-1

Table 4.1: The standard symmetric payoff structure used in the VLT
The high probability win (A) and high probability loss (C) stimuli have the same absolute rewards and expected values, as do the low probability win (B) and low probability loss (D) stimuli.

An example of the probabilistic structure of a typical VLT paradigm is given in Table 4.1. There are pairs of images in *win*, *loss*, or *no-change* conditions. In the *win* condition, a selection between a pair of images results in a win of 5 points 80% of time,

and 0 points 20% of time; in the *loss* condition, a selection between a pair of images results in a loss of -5 points 80% of time and 0 points 20% of time; in the *no-change* condition, a selection between a pair of images always results in 0 points (Table 4.1).

The structure of the task is *symmetric* in that corresponding stimuli from each valence condition have the same absolute expected values, as a consequence of the symmetry of the probabilities and rewards. To maximize earnings, participants must learn to select the image associated with the highest expected value within each pair. In other words, the optimal choice for the win pair is the high probability win image (80% win), whereas the optimal choice for the loss pair is the low probability loss stimulus (20% loss).

The VLT has been adopted by many researchers to examine the impact of learned value on perceptual and attentional processing by presenting the VLT stimuli in a variety of secondary tasks where the reward schedule is discontinued and no longer task relevant. Despite numerous studies using the same VLT, the conclusions drawn from the secondary tasks have varied. For example, Raymond and O'Brien (2009) reported two effects of acquired value on old versus new recognition of faces when attentional capacity was limited. First, stimuli previously associated with high probability outcomes (either win or loss) showed a processing advantage (i.e. greater recognition accuracy) regardless of available attention (reduced versus full) compared to stimuli previously associated with low probability outcomes. Second, win-associated stimuli showed processing advantages (versus loss-associated stimuli) when available attention was reduced. In another example, a reach-to-grasp task showed faster reaches toward stimuli previously associated with high probability outcomes (versus low probability) but more efficient reaches toward stimuli previously associated with wins (versus loss or no-change; Painter et al., 2014).

Asymmetries in learning wins and losses. But inferences about asymmetric valence effects on subsequent processing depend, at least implicitly, on the assumption that the values of win and loss stimuli have been learned equally well—otherwise the subsequent processing differences may be due to learning differences rather than valence per se.

Lin, et al. (2020) conducted a meta-analysis of studies adopting the VLT to compare learning for win and loss outcomes. In each study, the probabilistic structure was symmetric as in the example in Table 4.1 above. Nevertheless, the results of the meta-analysis showed that the probability of optimal choice was significantly higher for win-associated stimuli compared to loss-associated stimuli, suggesting a valence-based asymmetry. Furthermore, when Lin et al. (2020) conducted new experiments using the VLT, they found that the learning asymmetry was observed regardless of whether the outcome led to monetary or point earnings, and was also observed when participants were provided with a description of the task structure (with information about the specific probabilities and payoffs but not the association between stimuli and outcomes).

Asymmetries in explicit memory for wins and losses. One approach that Lin et al. (2020) have pursued to further understand the nature of the learning asymmetry is to probe participants' explicit knowledge of the outcomes associated with each stimulus by using a post-learning memory task. In the studies conducted by Lin et al. (2020) participants completed a forced choice recognition memory task in which participants indicated the outcome most likely associated with each image from the VLT (e.g. "very likely to win" for the 80% win scene). Performance on the post-learning memory task was consistent with the learning asymmetry in the VLT: memory accuracy was superior for optimal win scenes versus optimal loss scenes.

4.2 The Computational Reinforcement Learning Model

We apply computational reinforcement learning (RL) theory (Sutton & Barto, 2018) to build models of the VLT in order to provide new insights and possible explanations for the observed win-loss asymmetry. Our model is simple, but it yields interesting explanations of qualitative phenomena from the results of trial-level simulations, and it also provides some insights into performance on the subsequent outcome memory task described above.

RL theory (Sutton & Barto, 2018) provides a formal definition of the problem of learning from experience and insights on how to act so as to maximize cumulative rewards. In the standard RL problem formulation, a decision-maker, or an *agent*, determines at each time step t what action, a_t, to take at a given state, s_t (or observation), and at the next time step receives some reward r_{t+1} and transition to a new state or observation. The agent's goal is to maximize the expected cumulative future rewards. For example, in the VLT, the actions that result in maximum total reward are those actions that select the high probability win scene in the win condition and the low probability loss scene in the loss condition.

The Value Learning Task is a special case of the general RL problem in that it does not involve *sequential decision making*; i.e., each choice affects only immediate reward and not future rewards. The win and loss pairs in the VLT are thus each equivalent to a *two-armed bandit task*. Despite their simplicity, bandit tasks are nevertheless interesting in RL theory and algorithm development because they are the minimal setting which imposes the challenge of learning value from probabilistic outcomes along with the need to balance exploration and exploitation.

Sutton and Barto (2018) provide a number of algorithms for solving bandit tasks, including sophisticated methods that approach optimal exploration strategies. We adopt here a simple incremental algorithm that learns expected values via an error-

driven learning rule. The form of the rule is shared by many RL algorithms and theoretical approaches to human and animal learning.

We denote the estimated value of action a at trial t as $Q_t(a)$. In the VLT, the value of an action is its expected reward. For example, the value for the high probability win scene (80% win with a reward of 5) is $5 * 0.8 + 0 * 0.2 = 4$. The error-driven update is:

$$Q_{t+1}(a) = Q_t(a) + \alpha(r_t - Q_t(a)) \tag{4.1}$$

where α is the agent's learning rate; $\alpha \in [0, 1]$. When α is 0, there is no learning, and when α is 1, the agent only takes into account the feedback from the previous trial, giving rise to a *win-stay-lose-shift strategy.*

At each trial, the agent makes a selection according to a *choice rule* that converts current action value estimates into choices while balancing exploration and exploitation. There are several common choice rules, including *greedy* (always choose the action with the highest estimated value) and *epsilon-greedy* (choose a random action with probability ϵ otherwise choose greedily). We adopt here another standard choice rule for balancing exploitation and exploration: the *softmax* rule. According to the softmax rule, at trial t, the probability of choosing an action A given the value estimates for action A and B is:

$$P(A|Q_t(A), Q_t(B)) = \frac{\exp(\beta * Q_t(A))}{\exp(\beta * Q_t(A)) + \exp(\beta * Q_t(B))} \tag{4.2}$$

where β is the inverse temperature parameter, and larger β corresponds to greedier choices (e.g., Daw, 2011). The computed probabilities thus define a multinomial distribution from which an action is sampled; actions with higher value estimates are sampled more frequently, but lower-valued actions always have a nonzero probability. Table 4.2 provides an example of how the model updates action values, converts the

t	Condition	$P(C)$	$P(D)$	Model choice	Reward	$Q(C)$	$Q(D)$
0						0	0
1	Loss	0.5	0.5	C	-5	-1.15	0
2	Loss	0.09	0.91	D	0	-1.15	0
3	Loss	0.09	0.91	D	-5	-1.15	-1.15
4	Loss	0.5	0.5	D	0	-1.15	-0.886
5	Loss	0.37	0.63	D	0	-1.15	-0.68
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4.2: Model simulation example of a sequence of trials at the start of the Loss pair condition with $\alpha = 0.23$, $\beta = 2$.
The model chooses between stimulus C and stimulus D in the pair. Value estimates for both choices, denoted $Q(C)$ and$Q(D)$ start at 0. On each trial, the model converts value estimates for choices into choice probabilities $P(C)$ and $P(D)$, makes a selection by sampling a choice according to these probabilities, receives a reward, and updates its value estimates using the error-driven update rule.

values into choice probabilities, and samples an action choice for several trials in the loss condition given a specific pair of parameters ($\alpha = 0.23$, $\beta = 2$).

Adopting the softmax rule has the analytic advantage of directly giving a nonzero probability for each choice on each trial conditioned on the learners value estimates, which allows us to use maximum likelihood estimation to find the best fitting parameters to our data. The ability of RL models to make contact with human data at the individual trial level is a significant theoretical benefit of their use (Daw, 2011). In the following section we provide details on how we select model parameters and modeling learning and choice in the VLT at the trial level.

Error-driven learning rules with fixed learning rates such as the rule we adopt in Eq. 4.1 may be contrasted with the simple method of keeping a running average of experienced rewards as value estimates. Rules such as Eq. 4.1 are effectively computing a *weighted average* of experienced rewards, where more recent rewards are weighted more than rewards in the distant past. Such rules have the advantage that they allow the agent to adapt to non-stationary environments where the probabilistic payoffs may

be changing over time. They also require an *initial value estimate*, which can be a locus of prior knowledge about the environment. In the absence of prior knowledge, common initial value estimates are zero, very small random values with mean zero, or random values with a small positive mean; positive initial values estimates build in an optimism that is one method for encouraging exploration (Sutton & Barto, 2018). For our model we fix the initial value estimate to be zero and explore its implications.

The model thus has two free quantitative parameters that correspond to learning rate (α) and the balance between exploration and exploitation (β). These parameters influence how $Q_t(a)$ is updated and how the agent makes the selection at each trial. In our simulations below we explore two methods for setting the parameters: maximizing empirical fit to human data, and maximizing reward in the task.

4.3 Data and material availability

Data and source code are available at `https://osf.io/4vc3p/?view_only=78c4e692a08649a9abb68640f154166a`.

4.4 Simulating the Value Learning Task

Experiment structure. We simulate first the VLT in Lin et al. (2020). This task has three pairs of stimuli: one pair in the win condition, one pair in the loss condition, and one pair in the no-change condition, with payoffs and probabilities as in Table 4.1. The task has 300 trials across 5 blocks: 100 win pair trials, 100 loss pair trials, and 100 no-change trials.

Over these 300 trials the model thus estimates six values: the win-correct option, the win-incorrect option, the loss-correct option, the loss-incorrect option, and the two no-change options. (We focus here only on the values for the win and loss pairs as no learning happens for the no-change pair.) All initial values were set to zero in our

analyses. On each trial, the model makes a choice given the condition and the softmax choice rule (Eq. 4.2), receives a reward probabilistically according the parameters in task (Table 4.1), and updates the value of the corresponding choice according to the incremental update rule (Equation 4.1).

Because the point schemes and monetary currencies vary arbitrarily across VLT experiments, and any such points or currencies must be transformed by humans into an internal reward signal (Singh, Lewis, Barto, & Sorg, 2010), we use a standardized reward (1 and -1) in all of our subsequent analyses.

Setting model parameters. We simulated the VLT with parameters set in two ways: in the *data-driven approach* we estimate α, β for each individual participant to maximize fit to their choice data (details below). In the *theory-driven* approach we find optimal settings of α and β—settings that maximize expected reward in the task. This represents a simple bounded optimality analysis to find *computationally rational* (Lewis et al., 2014) parameter settings that determine the upper bound on performance given the constraints of the learning algorithm.

We use maximum likelihood estimation to find the pair of parameters that yields choices that best fit each human participant's choices. The likelihood is given directly by the softmax rule (Daw, 2011), and the likelihood or probability of the entire observed sequence of choices from one participant is the product of the probabilities of their choices on all trials:

$$\prod_t P(c_t = A | Q_t(A), Q_t(B)). \tag{4.3}$$

The product in Eq.4.3 is often an extremely small number and so it is usually better

to compute the summed log-likelihood instead, which is

$$\begin{aligned}&\sum_t \log(P(c_t = A|Q_t(A), Q_t(B)))\\&= \sum_t \beta * Q_t(A) - \sum_t \log(\exp(\beta * Q_t(A)) + \exp(\beta * Q_t(B))).\end{aligned} \tag{4.4}$$

To find α,β that maximize the quantity in Eq. 4.4 we use a simple randomized grid search, sampling 100 α's from a uniform distribution, $\mathcal{U}(0, 1)$, and 100 β's from a uniform distribution, $\mathcal{U}(0, 10)$, which resulted in 10,000 pairs of parameter settings. For each pair we computed the log-likelihood for each individual participant's data given each pair of parameters and the model. Finally, we chose the pair of parameters that produced the largest log-likelihood as the maximum likelihood estimation of each individual's learning rate and selection strategy (Daw, 2011).

We found an approximation of the optimal pair of parameters for the task with the same randomized grid search: we sampled 100 α's from a uniform distribution $\mathcal{U}(0, 1)$, and 100 β's from a uniform distribution $\mathcal{U}(0, 10)$, yielding 10000 pairs of parameters. For each pair of parameters, we calculated the mean sum of rewards in the task over 500 simulated runs (thus 5M total simulations). Finally, we chose the pair of parameters that produced the largest mean sum of rewards as an approximation to the optimal parameters. Simulating the VLT with these parameters allows us to see whether the qualitative empirical effects—in particular any win-loss asymmetries—persist when using the best possible parameters for the learning algorithm. The optimal parameter values also provide some insight into the nature of the task itself—what the task structure is demanding of the learner. The simple randomized grid search method also allows us to visualize the 2-D payoff surface (Figure 4.1, described below).

Main results. We simulated the VLT for all N=191 participants in Lin et al. (2020) who exceeded a minimal learning threshold, defined as achieving at least 65% correct

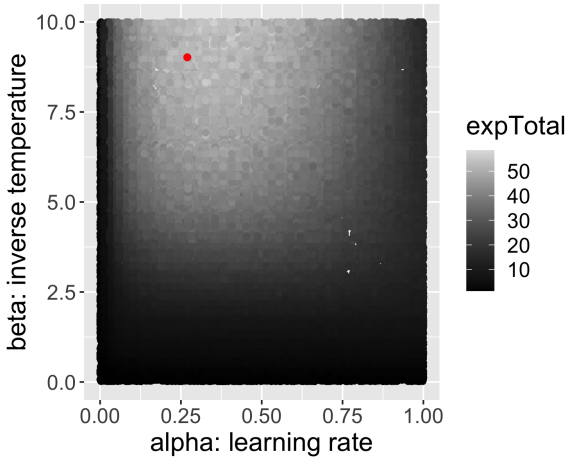
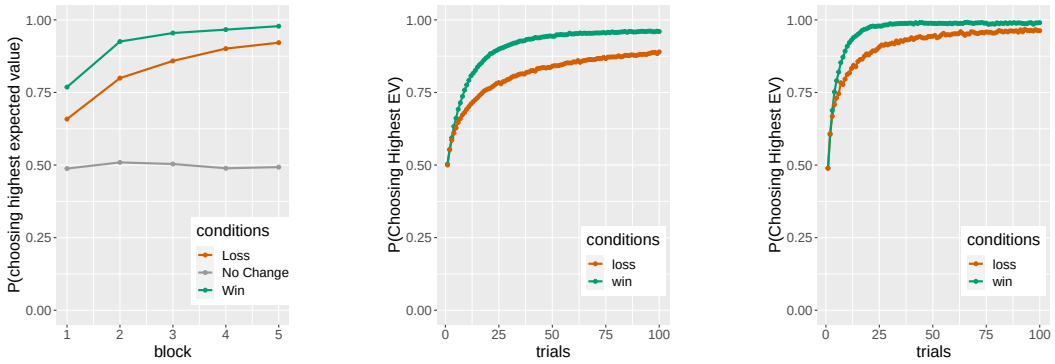


Figure 4.1: Total payoff given different values of α and β (model simulations). The pair of parameters that produce the highest reward is $\alpha^* = 0.27, \beta^* = 9.02$, shown as a red dot. We have applied an exponential transform ($1.07^{\text{total reward}+5}$) to the simulated accumulated rewards to make the visualization clear.

selection in the final block; we discuss the remaining poorly-performing participants below. The model was run 200 times for each participant and so the aggregated results represent a mean of 191×200=38,200 model runs.

Figure 4.2b shows the aggregate results of the model simulating the 191 participants. The results are very similar to the empirical results (Figure 4.2a), and in particular, there is a clear asymmetry in performance on the win and loss stimuli: the win pairs are learned better than the loss pairs. This difference diminishes with learning but persists through the final block.

We also simulated learning with optimal parameters—the settings of α and β that maximize expected reward. Figure 4.2c, shows the performance averaged over 5000 runs of the optimal parameter setting. The payoff surface is shown in Figure 4.1, which plots the total expected reward earned in the task given different values of α and β. The optimal values are $\alpha^* = 0.27, \beta^* = 9.02$, indicating that the best strategy is to update value estimates aggressively from recent past trials and accordingly exploit the learned better option in each condition. This is the result of the structure and



(a) Humans Lin et al. (2020)

(b) Model with individual best-fit parameters

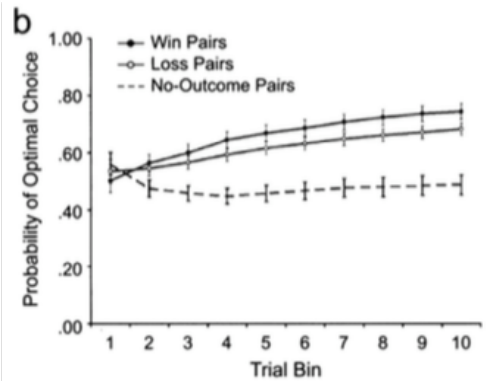
(c) Model with optimal parameters

Figure 4.2: Human (from Lin et al., 2020) and model performances in the VLT. (a) Human participant results ($N = 191$) from Lin et al. (2020): Mean probability of selecting the correct stimulus from win pairs and loss pairs, across the 5 blocks (100 trials total), showing better performance for win pairs than loss pairs. (b) Model simulation of probability of correct selection for the 191 participants using best-fitting parameters for individuals. (c) Model simulation of probability of correct selection using optimal parameters; the asymmetry persists in this model, though it is quantitatively diminished.

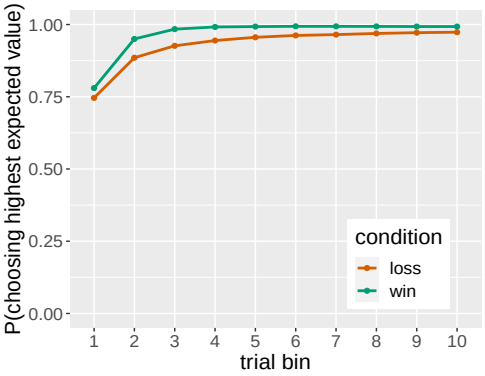
setup of the task, and is a function of the probabilities (0.8 and 0.2) and number of trials. Probabilities closer to 0.5 (say 0.65 and 0.35) would impose a more difficult learning task and result in lower optimal α levels.

Even at optimal learning and exploration rates, the simulation results show that win trials are learned better than loss trials, though the asymmetry is quantitatively diminished. This suggests that the explanation of the win-loss asymmetry cannot be simply that participants have adopted suboptimal learning or exploration rates.

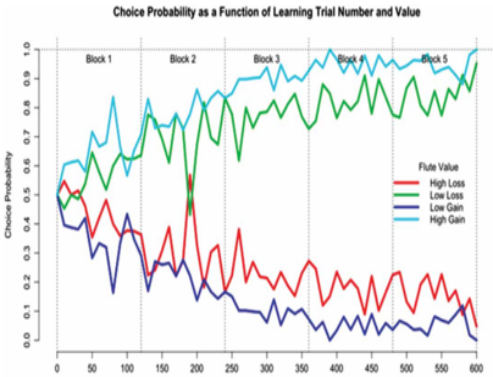
Simulation of three other VLT studies. In this section we show that the learning asymmetry exhibited by the model of the VLT in Lin et al. (2020) also occurs when simulating three other studies that use the same general paradigm, but with different numbers of stimuli pairs and number of trials (Raymond & O'Brien, 2009; Rothkirch, Tonn, Köhler, & Sterzer, 2017; Painter et al., 2014). Because we did not have access



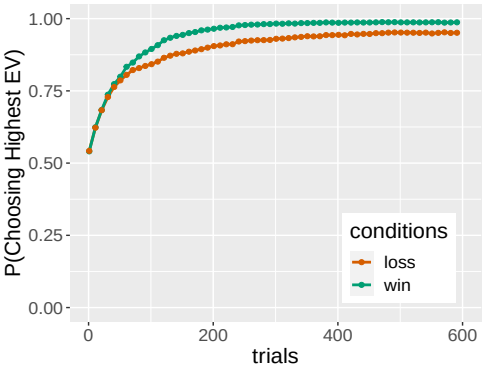
(a) Humans
(Raymond & O'Brien, 2009)



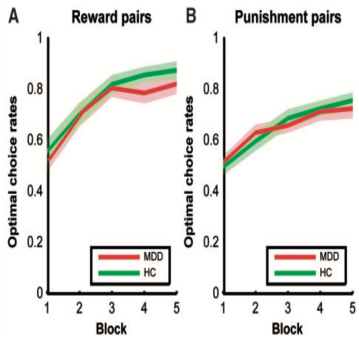
(b) Model simulations
Raymond & O'Brien (2009)
Optimal parameters ($\alpha = 0.32, \beta = 9.62$)



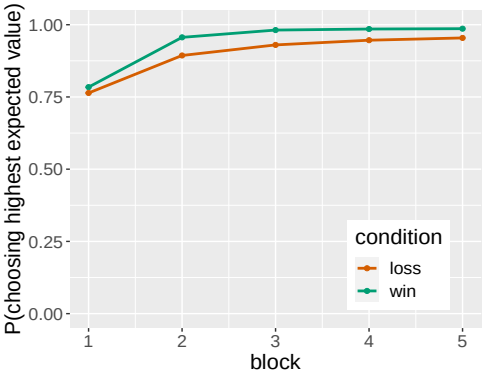
(c) Humans
(Painter et al., 2014)



(d) Model simulations
Painter et al (2014)
Optimal parameters ($\alpha = 0.42, \beta = 9.57$)



(e) Humans
Rothkirch et al (2017)



(f) Model simulations
Rothkirch et al (2017)
Optimal parameters ($\alpha = 0.44, \beta = 9.75$)

Figure 4.3: Simulations of three studies using the VLT. The difference in learning in wins and losses persists in these studies although they have different pairs stimuli or numbers of trials from Lin et al.(2020).

to individual participant data from these studies, we simulated the tasks using approximately optimal settings for α and β, using the method described above for find the optimal parameters.

In Raymond and O'Brien (2009) the VLT consisted of of six pairs of faces: two win pairs, two loss pairs, and two control pairs. Each pair was presented 100 times randomly in each block for a total of 6 blocks, yielding a total of 600 trials. On each trial, a choice led to a monetary outcome for win and loss trials (5 pence) with a probability of either 0.8 or 0.2 (Raymond & O'Brien, 2009). We simulated the probability of correct choice (mean of 10000 runs with optimal parameters: $\alpha = 0.32\,\beta = 9.62$) within ten 10-trial bins to match the data display in Raymond and O'Brien (2009) (Figure4.3 (a) and (b)). Both model and human participants show the win-loss asymmetry, though the model's performance with optimal parameters is much higher than the humans.

Painter et al. (2014) used twelve pairs of flute glasses as stimuli, six of which were win pairs and the other six were loss pairs. Each pair was presented 10 times in each block for a total 5 blocks, yielding a total of 600 trials. The monetary outcome for win and loss trials (20 cents in AUD) occurred with a probability of either 0.8 or 0.2. During the last block, participants no longer received any feedback, indicating that participants only learned during the first 4 task blocks and were tested for their learning during the last block (Painter et al., 2014). The empirical results and optimal-parameter model simulation (mean of 10000 runs) is shown in Figure 4.3 (c) and (d). Again, both the empirical and simulated results show that the win condition was learned better than the loss condition, and in this experiment the human participants are much closer to the performance of the optimal model.

The final study that we simulated was Rothkirch et al. (2017). Their task consisted of four pairs of stimuli, two of which were win pairs and the remaining two were loss pairs. Each pair was presented 10 times in each block for a total of 5 blocks, yielding a total of 200 trials. The monetary outcome for win and loss trials (5 cents) occurred

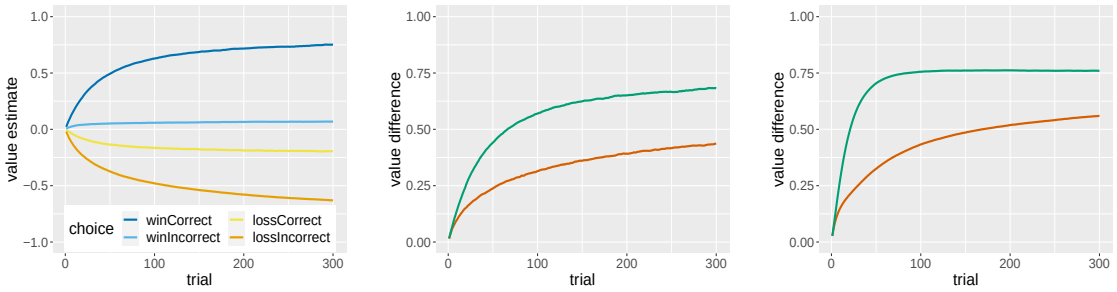
with a probability of either 0.8 or 0.2 (Rothkirch et al., 2017). The empirical results and optimal parameter model simulation (mean of 10000 runs) are shown in Figure 4.3 (e) and (f). There were no differences in participants' learning of the two pairs of stimuli *within* the win condition (called "Reward" in Rothkirch et al. (2017)) and the loss condition (called "Punishment"). But again, both the empirical and simulated results show win pairs were learned better than loss pairs over the blocks.

4.5 Explaining the Win-Loss Asymmetry

The VLT paradigm in Lin et al. (2020) and the three experiments above each have seemingly symmetric payoff structures (Table 4.1). But our model predicts that asymmetric learning of wins and losses will occur across all the experiments. What gives rise to the asymmetry?

An examination of the evolving value estimates in the model reveals that they exhibit a different pattern for win and loss pairs over the course of the simulated experiment. The mean trial-by-trial value estimates for all choices in win and loss conditions for the 191 models with α and β fit to individual participants is shown in Figure 4.4a, and Figure 4.4b shows the corresponding *differences* in values between stimuli in the win and loss pairs. These differences are key because they are monotonically related to differences in probability of choice for each option. It is clear that the stimuli in the win pair are more sharply discriminated than the stimuli in the loss pair.

Why is this the case? In the win condition, the value estimates for the win-correct option approach the true expected value of 0.8 within the first 150 trials; this is not surprising because the choice is sampled frequently. The win-incorrect option has still not approached the true expected value of 0.2 by the end of the experiment because it is sampled much less frequently and the initial estimate of zero still has its influence. Similarly, the loss-incorrect option is more slowly approaching the true expected value



(a) Value estimates by trial (individual models)

(b) Value differences by trial (individual models)

(c) Value differences by trial (optimal α and β)

Figure 4.4: Value estimates and differences by trial (model simulations). (a) Evolving mean value estimates for the four stimuli in win and loss pairs; mean computed from 30 runs of each of the 191 individual participant models. (b) Evolving mean differences in value estimates for win and loss conditions from the 191 individual participant models. (c) Evolving mean differences from the model with optimal α and β (5000 runs).

of -0.8 because it is sampled less frequently than the loss-correct option, which is approaching the true expected value of -0.2. But the result is that the value estimates of the loss pair stimuli are closer together, leading to comparatively greater choices of the incorrect loss option than the incorrect win option; put differently, model choices in the loss pair are noisier. The asymmetry persists when α and β are set to their optimal values (Figure 4.4c). In short, throughout the task, the loss stimuli remain more poorly discriminated than win stimuli.

4.6 Modeling a Learning Outcome Memory Task

Following the VLT, Lin et al. (2020) administered a post-learning memory task that aims to probe participants' explicit knowledge of the outcome associated with each stimulus (scenes) that appeared in the VLT [1]. The task included the 6 VLT scenes and 12 new scenes. VLT scenes were presented 4 times each and 12 new images each

[1] Lin et al. (2020) reported memory task performance for all participants who met learning criteria (N=191).

appeared twice. Participants indicated the outcome associated with each image as follows: 1) *very likely to win*, 2) *occasionally win*, 3) *no change*, 4) *occasionally lose*, 5) *very likely to lose*, 6) *none* (indicating a new image).

Figure 4.5, left panel, shows the human results. There is a clear interaction: the Win-80 stimulus was very accurately categorized but the Win-20 stimulus was categorized poorly. Each of the two Loss stimuli were categorized about equally well, better than Win-20 but not as accurately as Win-80. In short, there is a clear valence difference but also an interesting interaction. And given this interaction, when collapsing across the paired stimuli, accuracy on the loss stimuli is slightly overall higher than win stimuli—a counter-intuitive result given the choice performance asymmetry.

We extended the learning model to also provide an account of the performance on the memory probe task, for those stimuli that were part of the VLT. The simple hypothesis we pursued is the following: participants would make the categorical judgments based on their learned values estimates for each stimuli, using a set of reasonable thresholds over these estimates to yield the five categories.

In our initial exploration, we hand-picked the following intuitively reasonable ranges for the thresholds or breakpoints for mapping value estimates into the five categories (we also found empirical best-fit thresholds, described in Table 4.3). Recall that the observed rewards for the model were $+1$, -1, or 0.

Note that when these thresholds are applied to the *true* values of stimuli, they yield the intuitively correct categorizations of stimuli that were used as the definition of the correct responses for computing the empirical accuracy scores reported in Lin et al. (2020).

We then took the value estimates for win and loss stimuli from the models for each of the 191 participants who reached our learning criterion and sampled 1000 sets of thresholds from their plausible ranges to create simulated responses to the memory task.

Category	Definition	Threshold & Range
very likely win	value estimates $\geq$ highest threshold	highest threshold $\in [+0.50, +0.70]$
occasional win	value estimates $\geq$ high threshold	high threshold $\in [+0.17, +0.23]$
no change	value estimates $\geq$ low threshold	low threshold $\in [-0.23, -0.17]$
occasional loss	value estimates $\geq$ lowest threshold	lowest threshold $\in [-0.70, -0.50]$
very likely loss	value estimates $\leq$ lowest threshold	lowest threshold $\in [-0.70, -0.50]$

Table 4.3: Thresholds for mapping value estimates into the five categories.

Figure 4.5, middle panel, shows the probability of categorizing stimuli correctly based on the cutoffs above for the simulated value estimates. Figure 4.5, right panel, shows the results with cutoff thresholds chosen to maximize empirical fit (minimize mean-squared error between predicted and observed accuracies). It is clear that the modeling results recover the key qualitative patterns in the human data.

4.7 Discussion

The value learning task (VLT) developed by Raymond and O'Brien (2009) is a simple and popular paradigm for studying value learning and the effects that learned value have on subsequent processing of valued stimuli. But the standard paradigm, despite the apparent symmetry in payoff structure, yields a contrast between wins and losses: choice performance on win stimuli is better than loss stimuli (Lin et al., 2020; Rothkirch et al., 2017), and this pattern holds whether participants receive points or monetary rewards, and even when they are explicitly instructed about the structure of the task.

We developed a simple model of the VLT based on a standard error-driven learning

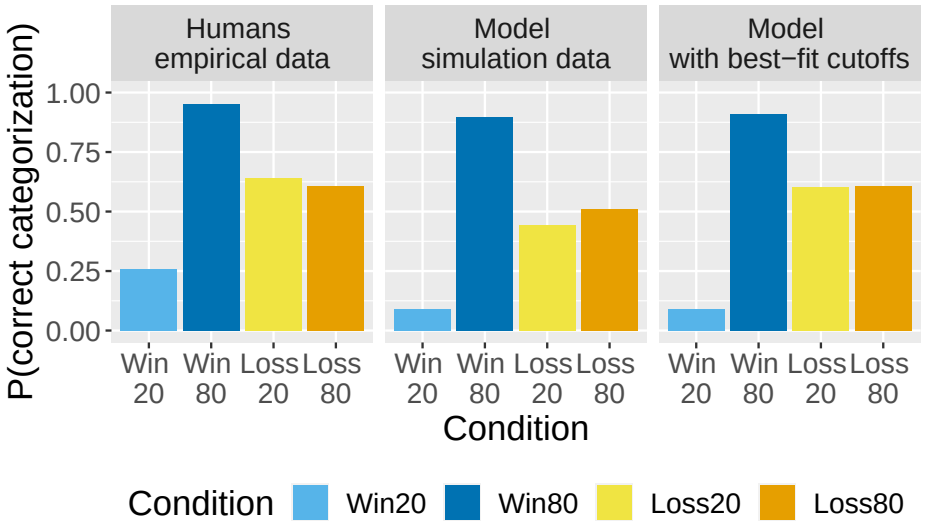


Figure 4.5: Memory task results (left) from human participants (N=191) and categorization of stimuli given simulated value estimates for the participants.
Simulation data (middle) show the mean probability of correct categorization based on 1000 sets of plausible cutoffs. Simulation with best-fit cutoffs (right) shows the probability of correct categorization based on the set of cutoffs that fit empirical data the best. The best-fit set (1: *very likely win* (value estimates $> +0.51$), 2: *occasional win* (value estimates $> +0.17$), 3: *no change* (value estimates > -0.17), 4: *occasional loss* (value estimates > -0.51), and 5: *very likely loss* (value estimates ≤ -0.51)) are decided by the minimum mean squared error between P(correct categorization) from simulation and empirical data.

rule, soft-max choice, and neutral (zero) initial value estimates. This model produces the asymmetry in learning gains and losses that is evident in human performance. This is the case despite (a) the task itself having a symmetric design; (b) the learning and choice rules having no special role for valence; and (c) allowing the learning and choice rule parameters to vary widely and include optimal settings for the task. The model furthermore yields an explanation: the asymmetric learning pattern arises from an interaction of incremental learning, exploitation while exploring, and neutral initial value estimates. As a consequence the learned values of the loss stimuli are discriminated less well than the win stimuli. We have shown this asymmetric learning pattern arises

in three other experimental tasks that have a very similar structure to the VLT in (Lin et al., 2020).

A simple extension of the model that uses the learned value estimates to simulate a post-learning outcome memory task provides further evidence for the asymmetric value estimates that the model naturally produces, and thus indirectly for our assumption of an initial neutral value estimate.

It is worth noting that, for this model, valence is relevant insofar as exploitation wants to pursue greater reward. But valence in the sense of positive/negative does not play a special role. Therefore, such a computational model is very useful for researchers to have as a baseline model for any value learning task, to draw out the implications of the simplest set of assumptions that don't assume a special role for positive/negative valence. In this sense, it is also a way to put into sharper focus any real valence-related differences that do emerge.

The model-based analysis provides some insights into what we could do to reduce the asymmetry in learning in the VLT, without compromising the task's symmetric design. Again, the asymmetric pattern is a result of the interaction of incremental learning, the balance between exploration and exploitation, and zero initial values. One clear way reduce the asymmetry in learning is to adjust the initial values for the actions by allowing an extra block at the beginning of the experiment as a purely exploration phase, where participants are instructed to learn as much about each option as possible, without concern for exploitation. Adjusted initial values could lead to a smaller difference in learned value estimates between win and loss conditions, and may subsequently produce smaller differences in performance for categorizing win and loss stimuli. This solution needs to be tested with further empirical work.

Finally, the asymmetric learning pattern for win and loss stimuli in the VLT does not arise uniformly across participants: a subset of the participants learn wins and losses nearly equally well. This suggests that a natural next step is to investigate

the individual differences in performances in the VLT. In the following chapter, we present and discuss the individual differences in the VLT and how our model may help explain the variation in terms of individual variation in the learning and exploration parameters.

CHAPTER V

Individual Differences in Value Learning

We have shown that a simple RL error-driven learning model provides an explanation of win-loss learning asymmetries in the superficially symmetric VLT paradigm, and have shown that this asymmetry persists whether the learning and exploration parameters are set to maximize empirical fit to individual participants, or are set to the computationally rational optimal setting to maximize task reward. We have also demonstrated how final value estimates of the stimuli from this simple RL model can provide explanations for the results in the post-learning explicit value categorization task in Lin et al. (2020). However, empirical data from Lin et al. (2020) suggest that the striking asymmetric learning pattern does not characterize all individuals: a subset of participants learned both conditions nearly equally well.

In this chapter, we examine the extent to which the asymmetry persists for all participants, and whether variation in the model's learning parameters can account for individual differences. We also examine whether the same individual differences occur in the post-learning categorization task and discuss its indications.

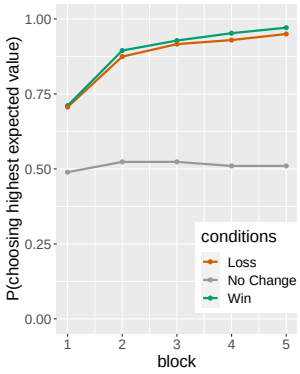
5.1 Individual Differences in Performances in the VLT

Despite of the consistent win-loss learning asymmetry observed in aggregate (Lin et al., 2020), empirical data also suggest that this pattern does not arise in all participants. To help visualize this participant variation, we characterized the learning asymmetry for each participant (in the N=191 who achieved at least 65% correct selection in the last block) by computing the difference between mean probabilities of correct selection of win and loss stimuli across the 5 blocks. We then separated participants into two groups using a median split on this difference measure. We refer to the group with lower win-loss differences as the *Nearly Equal Learner Group*, and the group with greater win-loss differences as the *Unequal Learner Group*.

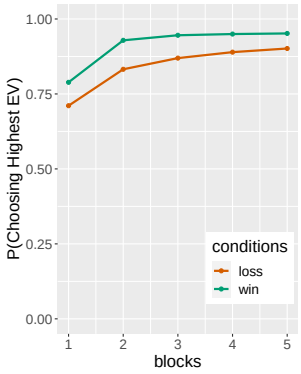
Figure 5.1 shows the empirical (left panel) and best-parameter-fit model-simulation (right panel) learning curves for the Nearly Equal Learners (top row) and Unequal Learners (bottom row). Note that these model simulations are identical to the ones presented in the previous chapter for the N=191 participants in Lin et al. (2020) [1]; we are simply splitting those results into the two different groups. The key result here is that the asymmetry is diminished in the simulation of the Nearly Equal Learners, though not to the extent observed in the empirical means. This suggest that variation in α and β provides a partial account of the individual variation in the win-loss asymmetry.

We also explored the effect of the individual parameter variation on predicted estimated values for participants in the two groups (Figure 5.2). Consistent with the analysis presented above, the mean differences in value estimates for win and loss conditions from models of participants in the Nearly Equal Learner group are smaller than those from the models of participants in the Unequal Learners' group.

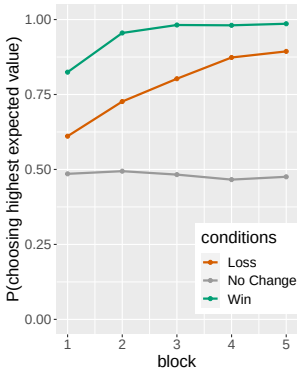
[1] This individual difference analysis was not included in Lin et al. (2020)



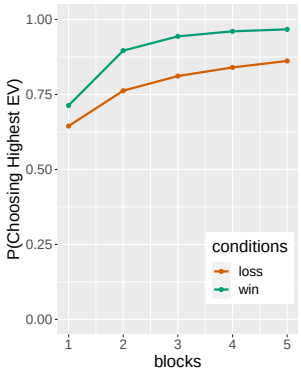
(a) Human Nearly Equal Learners (N=95)



(b) Model Nearly Equal Learners (N=95; best fit individual parameters)

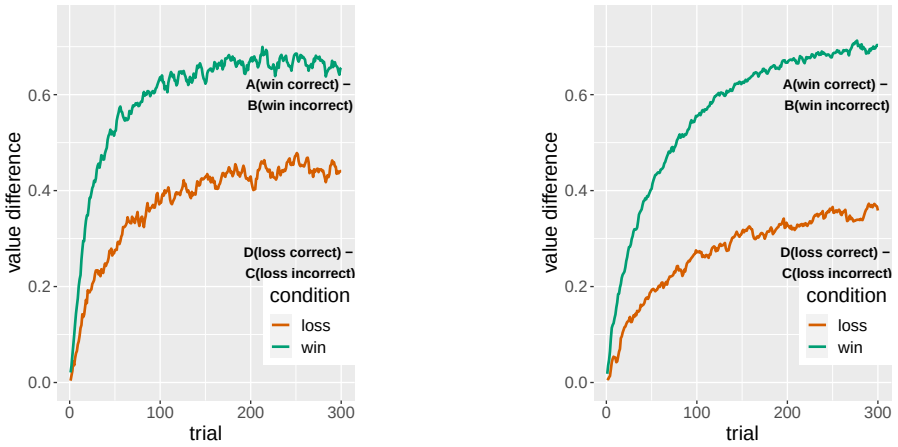


(c) Human Unequal Learners (N=96)



(d) Model Unequal Learners (N=96; best fit individual parameters)

Figure 5.1: Human data and model simulations for two groups of participants created by a median split on the learning asymmetry; see text for details.



(a) Model-produced value differences by trial
Nearly Equal Learners (N=95)

(b) Model-produced value differences by trial
Unequal Learners (N=96)

Figure 5.2: Value differences by trial for Nearly Equal and Unequal Learners. (a) Evolving mean differences in value estimates for win and loss conditions from the 95 models of the Nearly Equal Learner participants. (b) Evolving mean differences in value estimates for win and loss conditions from the 96 models of Unequal Learners.

5.2 Effect of Experience on Individual Performances

Each individual participant also differed in the specific experiences on each trial. Although it seems unlikely that these experience differences could account for the individual differences we observed, we also ran simulations of the model using the actual experience of each individual participant—that is, forcing the model to experience the exact same trial conditions in the same order as the participants. We then computed optimal learning parameters for these individual experiences to assess whether experience alone might lead to upper bounds on performance that vary enough to account for some of the observed performance differences. We did not observe any differences in the optimal model simulations, suggesting that random experience differences cannot account for the observed variation in individual performance (Simulation results in Appendix C).

5.3 Model Simulations of Poor Performers

Finally, the model simulates the performance of most participants who did not achieve the 65% correct selection threshold (some of whom were operating nearly at chance). Setting either α or β to very low levels yields poor performance. It is possible to further divide the poor performing participants into subgroups who learned neither win or loss associations (N=23), or who learned wins slightly better than losses (N=17), or losses slightly better than wins (N=8). Only the latter small group of participants (8 of 287) cannot be accounted for by the model. Simulation results of these four subgroups are in Appendix C).

5.4 Individual Model Parameters

Figure 5.3 shows the best-fitting α and β parameters for each of the 287 participants, color coded for each of the three groups: *Unequal Learners*, *Nearly Equal Learners*, and *Poor Performers*. What is clear from this plot is that the Nearly Equal Learners have parameter values closer to the optimal parameters. The model thus predicts that these participants will have the highest overall performance, a prediction that is confirmed empirically (See Appendix C).

5.5 Individual Differences in the Outcome Memory Task

Recall that Lin et al. (2020) administered a post-learning memory task that aims to probe participants' explicit knowledge of the outcome associated with each stimulus that appeared in the VLT. Participants indicated the outcome associated with each image as follows: 1) *very likely to win*, 2) *occasionally win*, 3) *no change*, 4) *occasionally lose*, 5) *very likely to lose*, 6) *none* (indicating a new image).

Here we present and discuss the empirical and simulation results from the catego-

rization task by two sub-groups of participants: Nearly Equal Learners and Unequal Learners. Note again that the data and simulation are both the same as those in Lin et al. (2020) and the previous chapter — we are simply dividing the results into the two groups.

Figure 5.4, top panel, shows the human results for the two groups of participants. We observe the same interaction as shown in aggregated data from before in both sub-groups: both Nearly Equal Learners and Unequal Learners categorized the Win-80 stimulus very accurately but categorized the Win-20 stimulus poorly. Both groups categorized each of the two Loss stimuli almost equally well.

We applied the same set of thresholds for mapping value estimates into the five categories (Table 4.3). Figure 5.4, second row, show the probability of categorizing stimuli correctly based on the thresholds for the two groups of simulated value estimates. Figure 5.4, third row shows the results with thresholds chosen to maximize empirical fit for each group separately (minimize mean-squared error between predicted and observed accuracies). The modeling results recover the qualitative patterns in the human data for both groups. We observe that despite differences in learning of the win- and loss-stimuli in the VLT, the qualitative effects on subsequent memory performance are the same.

5.6 Discussion

The asymmetric learning pattern for win and loss stimuli in the VLT does not arise uniformly across participants. In this chapter, we show how the model partially explains this variation in terms of individual variation in the learning and exploration parameters. From simulating each participant's individual task experience with both the optimal parameters and best-fitting parameters, we are also able to rule out random variations in task experience as the source of the individual differences.

We also show the learned value estimates from the VLT predicts the performance in the post-learning outcome memory task for both human Nearly Equal Learners and Unequal Learners. The simulation yields the observed win-loss interaction in the human data in both groups and even accounts for the surprising finding that accuracy in categorizing outcomes of loss-stimuli is slightly better than win-stimuli (Results of overall correct categorization for win- and loss-stimuli in Appendix C). However, despite nearly equal or unequal learning of the win- and loss-stimuli in the behavioral task, the qualitative effects on subsequent memory performance were the same. This suggests that the win-loss asymmetry in learning does not directly drive effects on subsequent tasks. Instead, the learned value estimates were better predictors of the subsequent memory task. Thus, a promising avenue for future work is to quantitatively model value learning as we have done here, and use the learned value estimates as parameters of computational models of downstream tasks.

Finally, the individual differences in the VLT provide us with insights on some limitations of our model. First, although the model simulations reflect the general characteristics of human performances by people in different groups, the model cannot account for the performance of the small percentage ($< 3\%$) of individuals who performed better in the loss condition than the win condition. It is possible that individuals who learned losses better than wins have a different internal reward function that transforms point or monetary observations into an internal reward signal, but this could be very challenging to estimate. Second, our model does not take into account the possibility that humans may also learn the *structure* of the task in ways that allow them to update value estimates for the stimulus in the pair *other* than the one that is chosen. In other words, in the VLT, feedback on one stimulus in a pair does provide information about the value of the other stimulus. It is possible that this more efficient task structure learning accounts for some of the performance differences of participants in the Nearly Equal Learners group.

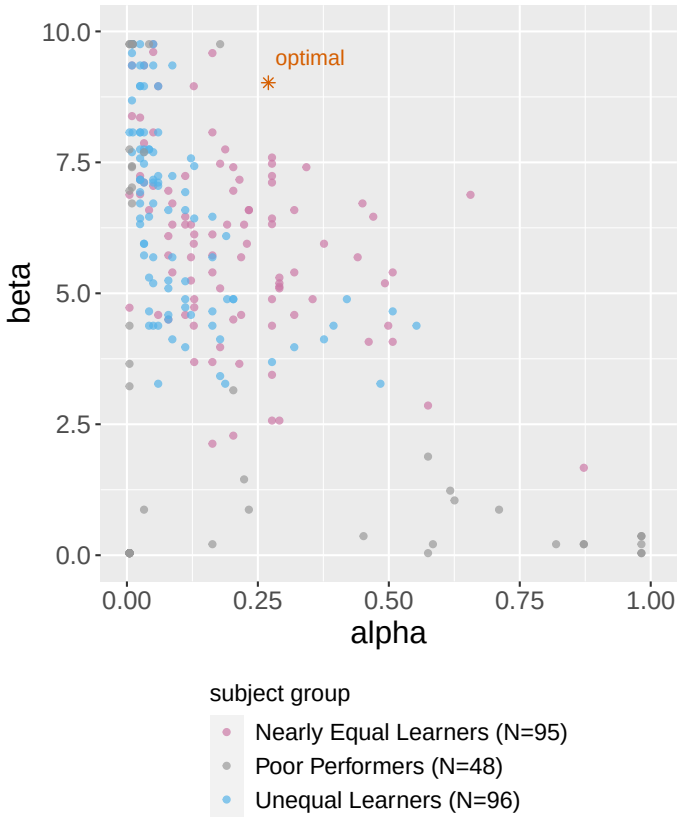


Figure 5.3: Individual best-fit parameters for all participants.
Individual best-fit parameters for all participants, distinguishing the three groups: *Nearly Equal Learners*, *Unequal Learners*, and *Poor Performers*. Parameter settings for the *Nearly Equal Learners* are closer to the optimal setting (see Figure 4.1); the model thus predicts that these participants will have higher overall performance, a prediction that is consistent with the data.

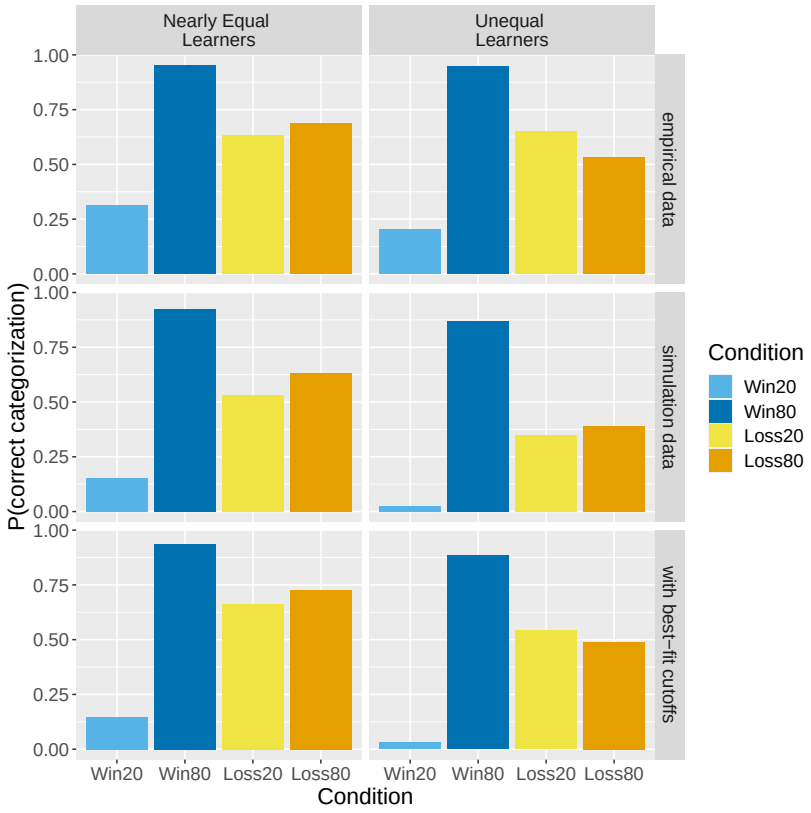


Figure 5.4: Memory task results from human learners and categorization of stimuli given simulated value estimates for the two groups of participants.
Simulation data (second-row panels) show the mean probability of correct categorization for the two groups based on 1000 sets of plausible cutoffs. Simulation with best-fit cutoffs (third-row panels) shows the probability of correct categorization for the two groups based on the set of cutoffs that fit empirical data the best.

CHAPTER VI

General Discussion

In this dissertation, we present work on value-based decisions in two different domains: multi-attribute ethical dilemmas (Chapters II and III) and the Value Learning Task (VLT; Chapter IV and V; Raymond & O'Brien, 2009). Key results in this dissertation are summarized in Table 6.1 below. In the ethical decision tasks, the values are subjective utilities of each available option calculated through combining the attributes given certain subjective utility functions. In the VLT, the values are estimates of expected reward associated with novel stimuli, updated through hundreds of trials of learning. Both domains require the decision maker to represent the decision problem (i.e., identify available options, attributes, etc.), evaluate the alternatives, and make the decision.

In Chapter II, our work established for the first time that contextual choice reversals (or the attraction effect) occur in the domain of ethical decisions. Our empirical studies built on a classical paradigm in contextual preference reversal studies. We started with dilemmas isomorphic to economic gambles, and then extended the choice tasks to multiple specific ethical dilemmas that include qualitative attributes. We found both between-subject and within-subject choice reversals in both ethical dilemmas isomorphic to economic gambles and ethical dilemmas that involve qualitative attributes.

Chapter	Key Results
Chapter II	1. We established for the first time that both between-subject and within-subject contextual preference reversals can be found in ethical decisions with either quantitative or qualitative attributes. 2. Performance in ethical decisions is better and higher target reversal rates predict better performance in economic gambles but not in ethical decisions.
Chapter III	1. We used a computational model that predicts choices in ethical dilemmas while taking into account individual differences to show that the individual differences in rankings of ethical features can affect the original configurations of the ethical dilemmas designed to produce context effects. This partly explains the variation of within-subject choice reversal rates across specific ethical dilemmas.
Chapter IV	1. We used a simple reinforcement learning model with the softmax choice rule to provide explanations for the win-loss learning asymmetry in the VLT. The asymmetry occurs as the result of an interaction between a neutral initial value estimate and a choice policy that exploits while exploring, leading to more poorly discriminated value estimates for loss stimuli. 2. We also recovered the key qualitative patterns in the human data by mapping final value estimates from the models to categories of stimuli.
Chapter V	1. The win-loss learning asymmetry diminished in the simulation of the Nearly Equal Learners, though not to the extent observed in the empirical means. 2. Individual differences in the learning asymmetry does not affect the qualitative patterns of the subsequent memory performance.

Table 6.1: Summary of Key Results in this Dissertation.

However, we also discovered that within-subject contextual choice reversals do not arise to the same extent across specific ethical dilemmas.

In Chapter III, we used a simple computational model to show that the variation of contextual choice reversals across dilemmas are partially explained by individual

differences in the representation of the decision problem. Specifically, decision makers differ in how they rank the levels of attributes (especially qualitative ones). As a result, the original configurations of the ethical dilemmas change according to the decision maker's feature rankings. The configuration from a classical paradigm in contextual choice reversal studies (Table 1.1) has three multi-attribute options in the two-dimensional feature space. There is a target option, a competitor option that is as attractive as the target option, and a decoy option that is dominated by the target. When the decision maker's feature rankings differ from the assumed rankings, an attraction effect configuration could change to a similarity effect configuration, for instance. These changes further affect within-subject choice reversal rates.

While our results demonstrate the commonality between decisions in economic gambles and in ethical domains, our exploratory findings also suggest that there are also differences in contextual choice reversals between choices in economic gambles and ethical decisions. We found that the performance in ethical decisions is better and that higher target reversal rates predict better performance in economic gambles but not in ethical decisions.

A further direction along this line of research is to investigate whether other established context effects: similarity effect and compromise effect; (Wollschlaeger & Diederich, 2020) can be found in ethical decisions as well. Furthermore, it is worth to explore whether the differences in performance and differences in the relationship of reversal rates and performance between choices in economic gambles and ethical dilemmas also arise in other context effects. This direction will allow us to explore individual differences across context effects and choice domains (economic vs. ethical) in more depth. Specifically, we can explore how individuals' subjective utility, or decision strategies affect context effects in multi-attribute choices in economic and ethical domains.

In Chapters IV and V, we investigated decisions in the VLT — a paradigm developed for studying how people learn values associated with neutral stimuli and the effects that learned value have on subsequent processing of the valued stimuli (Raymond & O'Brien, 2009). The VLT consists of a series of trials in which participants are presented with a pair of neutral images associated with win, loss, or no-change outcomes with certain probabilities. Participants attempt to maximize accumulated winnings as they make choices and learn how the images associate with wins or losses. Despite the symmetrical structure of the VLT, results from various studies show a clear contrast between wins and losses, where wins are consistently learned better than losses (Lin et al., 2020). In Chapter IV, we provide an explanation for the asymmetry in learning wins and losses using a simple reinforcement learning model: the asymmetry arises from an interaction between a neutral initial value estimate and a choice policy that exploits while exploring, leading to more poorly discriminated value estimates for loss stimuli. We also show that the final value estimates produced by the model provide a simple account of the subsequent explicit value categorization task. Specifically, we recovered the key qualitative patterns in the empirical data by mapping final value estimates from the models to categories of stimuli with a set of thresholds.

In Chapter V, we show that individual differences in learning rates and exploration rates help explain individual differences in the observed win-loss asymmetries. We first separated human participants into two groups: the Nearly Equal Learner Group (where participants showed lower win-loss differences) and the Unequal Learner Group (where participants showed greater win-loss differences). Then, we divided the model simulation results of all participants in the same way, and we found that the asymmetry diminished in the simulation of the Nearly Equal Learners, though not to the extent observed in the empirical means. This key result suggests that individual differences in learning rates and exploration rates partly explains individual differences in the observed win-loss asymmetries. Furthermore, we found that individual differences in

the learning asymmetry does not affect the qualitative patterns of the subsequent memory performance. Instead, performance of the explicit value categorization task is affected by the final value estimates of stimuli.

Our results suggest that researchers should use this computational model for any value learning task before drawing inferences about how learning asymmetries affect subsequent tasks. This model can help draw out the implications of the simplest set of assumptions that don't assume a special role for positive/negative valence.

We propose two potential future directions in this line of work. First, our model does not take into account how learning the task structure may help the individuals to update the value estimates of the non-chosen option. Thus, it is natural to consider the development of a more sophisticated structured Bayesian RL model to account for such learning. For example, in such a Bayesian RL model, each participant's learning parameter and exploration parameter are drawn from a common population distribution (Daw, 2011). Second, as the asymmetric pattern is a result of the interaction of incremental learning, the balance between exploration and exploitation, and zero initial values, one potential way to reduce the asymmetry in learning is to adjust the initial values for the actions by allowing a purely exploration phase for the participants. This naturally suggests a further empirical direction to test this solution for reducing the learning asymmetry.

Our two lines of work investigate decisions in different domains, but they both touch on the basic processes involved in value-based decisions (Rangel et al., 2008) — our empirical studies on multi-attribute ethical decisions explore how decision makers combine attributes and evaluate choices based on subjective utilities and our computational work on the VLT focuses on how the decision makers learn values from experiences and make decisions based on learned value estimates.

In addition, our studies in both domains of decisions reflect the various issues

present in value-based decisions that pose challenges to neoclassical economic theory which assumes perfect rationality yet can be explained or modelled in the framework of bounded rationality (Simon, 1955). Below, we briefly summarize and discuss how each line of work is informed by the framework of bounded rationality.

Our work on context effects in ethical decisions is conceptually tightly connected to the framework of bounded rationality. Context effects like contextual choice reversals are traditionally considered as violations to human rationality based on utility maximization. However, recent computational models explore the underlying processes of multi-attribute choices predict contextual choice reversals while retaining bounded rationality. Howes et al. (2016) show that these reversals are a natural consequence of utility maximization given observation and calculation noise. Given our results of finding these reversals in ethical decision tasks, we hope to extend the same theoretical model to the domain of ethical decisions and to provide the possibility of rigorous accounts of the bounded rationality of ethical decisions. A challenge in extending the theoretical model to the domain of ethical decisions is identifying individual decision maker's subjective utility functions. To establish a subjective utility function, we need to make assumptions about how the decision maker weighs and combines different pieces of information. This is particularly difficult given the qualitative nature of many ethical dilemmas and the individual differences in rankings of specific ethical features.

In our work on decisions in the VLT, we apply the framework of computational rationality to analyze to what extent are human performances affected by the task structure. One of the ways that we simulate the VLT with our RL model is to use the pair of parameters that maximizes total rewards, i.e., the optimal parameters. Essentially, the optimal parameters represent the computationally rational (Lewis et al., 2014) parameter settings that yield the optimal performance in the VLT given the learning algorithm, i.e., our RL model. As a result, the model still produces the win-loss asymmetry, albeit the asymmetry is smaller than that observed empirically in Lin et al.,

(2020). This analysis allows us to rule out the possibility that the empirically observed win-loss asymmetry is fully due to sub-optimal learning by the task participants.

Furthermore, the optimal parameter settings provide us with a new perspective to explore individual differences and what affects those differences — besides comparing the learning asymmetries between individuals who learned wins and losses nearly equally well and those who learned wins better than losses, we are able to compare the model parameter values and performances of two groups of individuals with the optimal parameters. Individuals who learned wins and losses nearly equally well have parameter values closer to the optimal parameters, and higher overall performance.

Our work illustrate that simple theoretical ideas often lead to complex phenomena. The ethical choice model that reflects individual differences is a simple generative model, yet it provides partial explanations of the variation of context effects across specific ethical dilemmas, and insights on how ethical decisions differ from economic gambles: individual differences in feature rankings challenge the configuration of context effects tasks in ethical dilemmas. The reinforcement learning model of the VLT is also a simple computational model, yet it explains the win-loss learning asymmetry observed in the VLT empirically. On the other hand, theoretical ideas also help unify complex phenomena across specific domains. Models of boundedly rational multi-attribute decision making provide us with a common ground to investigate economic and ethical decisions, and the reinforcement learning theory allows us to explore learning and memory together.

I would like to end this dissertation with a reminder from Allen Newell's last lecture in 1991: all of ultimate scientific questions are so deep about the universe that one can be held by them for an entire life and still be just a little ways into them (Anderson, 2007). Our work in no way comprehensively explores the intricate area of value-based decisions. However, we still hope that our new empirical findings and computational

work in the domains of ethical decisions and value learning can provide novel perspectives on the environmental and cognitive factors that influence value-based decisions, and that our results will motivate further research in the area of value-based decisions.

APPENDICES

APPENDIX A

Supplemental Materials for Choice Reversals in Ethical Decisions

A.1 Background

Behaviors	Heuristics	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 wrong-doing for a fee	Rules-of-thumb

Continued on next page

Continued from previous page

Behaviors	**Heuristics**	**Type**
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 A.1: Summary of inconsistent behaviors and their underlying moral heuristics.

A.2 Wedell (1991) Replication Study

A.2.1 Method

A.2.1.1 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.

A.2.1.2 Wedell (1991) Replication Materials

We constructed a questionnaire that contained 40 pairs of questions (80 in total) with the original stimuli from Wedell (1991), shown in Table A.2 below. 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.

	Decoy bets		
Target 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 A.2: Gambles used in Wedell (1991)'s original Experiment 1.
R = range decoy; F = frequency decoy; RF = range-frequency decoy (Wedell, 1991).

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 A.2, except for control (R' decoys), which were constructed by altering the values of R decoys so that the R' 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), there were two within-subject variables – 10 pairs of questions and decoy position (atA, atB) – and two between-subject variables: decoy type (R, F, RF in Experiment 1 and R and R' in Experiment 2) and presentation order. In our replication study, we had three within-subject variables: each person completed 10 pairs of questions, decoy position (atA, atB), and decoy type (R, F, RF, control/R').

A.2.1.3 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.2.2 Results

A.2.2.1 Descriptive Analysis

The descriptive analysis was the same as that in our ethical decision making study.

Table A.3 shows an overview of within subject choice reversals for each type of decoy. This result suggests that choice reversals occur around 20% of time within subjects and that single dimensional decoys (R and F) have stronger effects than RF or R'.

Decoy Type	Subjects	Total Pairs Done	# of PR	Proportions of PR
R	142	382	82	0.21
F	137	367	83	0.23
RF	142	371	63	0.17
control (R')	145	380	63	0.17

Table A.3: Proportions of Within Subject Choice Reversals (PR) Occurrences in Wedell (1991) Replication Study

Figure 3b in the main paper shows the complete proportions of choice patterns in this Wedell (1991) replication study. The decoy selection rates (R decoy: 20.41%, F decoy: 16.89%, RF decoy: 17.25%, R' decoy: 14.21%) in our replication study were fairly high compared to Wedell (1991). However, the replication study results in terms of proportions of choice patterns are very similar to Wedell (1991) original results. We observed that the majority participants had consistent choice within each pair regardless of decoy type. A clear and strong choice reversal effect can be observed for R and F decoy type, whereas the choice reversal effect is weaker for RF decoy. In the control, or R' decoy, condition, least numbers of subjects exhibited choice reversal. In general, descriptive plots suggest that we replicated Wedell (1991) results.

A.2.2.2 Bayesian Statistical Analysis

The statistical analysis was the same as that in our ethical decision making study Experiment 1.

Chain Convergence Evaluation For all of our models, we ran four independent chains and each of the four chains contains 2000 samples of each parameter. First 800

samples were part of the *warmup* (or *burn-in* period). This period allowed the sampling process to converge to the posterior distribution, and we analyzed the samples after this period.

To interpret posterior distributions more cautiously and accurately, we first checked chain convergence through traceplots and *Rhat* (Sorensen et al., 2016). The traceplots for parameters (not included) suggest that the chains have converged for all parameters the *Rhat* values from Table A.4 have shown convergence as well.

Posterior Statistics The central tendencies and 95% credible intervals of the posterior distributions are reported in Table A.4. We will focus on parameters $beta_A^{decoy}$ and types of decoy in our results.

For our main parameter of interest, $beta_A^{decoy}$, the mean estimate is 0.34, showing a choice reversal effect. There is 40% increase in log odds of choosing A to B when decoy is moved from B to A. After adding the interaction, we can also see that when decoy is moved from B to A, compared to baseline category R and F (1D), decoy type RF has a weak positive effect ($0.28 - 0.15 = 0.13$, indicating 14% increase in log odds of choosing A to B). However, R' has a reversed effect ($0.27 - 0.54 = -0.27$, indicating 24% decrease in log odds of choosing A to B).

A.2.3 Discussion

This experiment successfully replicated Wedell (1991), helping us move forward into the domain of ethical decisions by creating tasks that are isomorphic to those in Wedell (1991).

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
Full Model							
β^{gm} (Intercept)	1.00	2789	0.69	0.08	0.53	0.69	0.84
$\beta_A^{R'}$	1.00	2682	0.27	0.14	-0.02	0.27	0.55
β_A^{2D}	1.00	2664	0.28	0.15	-0.00	0.28	0.58
β_A^{decoy}	1.00	2498	0.34	0.12	0.11	0.34	0.57
$\beta_A^{R'*decoy}$	1.00	2494	-0.54	0.20	-0.93	-0.54	-0.14
$\beta_A^{2D*decoy}$	1.00	2528	-0.15	0.21	-0.57	-0.14	0.27
mean_PPD	1.00	4209	0.71	0.01	0.69	0.71	0.73
log-posterior	1.00	1993	-1630.18	1.71	-1634.53	-1629.83	-1627.79
Model 1							
β^{gm} (Intercept)	1.00	4942	0.55	0.10	0.37	0.55	0.75
β_A^{R}	1.00	4448	0.31	0.12	0.07	0.31	0.54
β_A^{RF}	1.00	4705	0.36	0.12	0.12	0.36	0.60
β_A^{decoy}	1.00	5767	0.30	0.10	0.11	0.30	0.49
mean_PPD	1.00	5207	0.71	0.01	0.69	0.71	0.74
log-posterior	1.00	2210	-1200.90	1.37	-1204.30	-1200.59	-1199.19
Model 2							
β^{gm} (Intercept)	1.00	2117	0.54	0.11	0.32	0.54	0.76
β_A^{R}	1.00	2402	0.30	0.16	-0.01	0.30	0.62
$\beta_A^{R'}$	1.00	2287	0.42	0.16	0.10	0.41	0.73
β_A^{RF}	1.00	2432	0.43	0.16	0.12	0.43	0.75
β_A^{decoy}	1.00	1842	0.34	0.16	0.02	0.34	0.67
$\beta_A^{R*decoy}$	1.00	2133	0.01	0.24	-0.45	0.01	0.48
$\beta_A^{R'*decoy}$	1.00	2076	-0.54	0.23	-0.98	-0.54	-0.09
$\beta_A^{RF*decoy}$	1.00	2347	-0.15	0.24	-0.61	-0.15	0.31
mean_PPD	1.00	4821	0.71	0.01	0.69	0.71	0.74
log-posterior	1.00	2000	-1629.76	1.99	-1634.54	-1629.43	-1626.88

Table A.4: Posterior Statistics for Wedell (1991) Replication Data

A.3 Experiment 1

A.3.1 Data Structure and Descriptive Analysis

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

subject	pair	decoy type	decoy dimension	pair choice pattern
1	1	R	1D	consistent choice
...	...	...	...	...
1	21	RF	2D	target reversal
2	11	F	1D	target reversal
...	...	...	...	...
2	31	R'	1D	decoy selected
...	...	...	...	...
J	11	F	1D	target reversal
...	...	...	...	...
J	21	RF	2D	competitor reversal

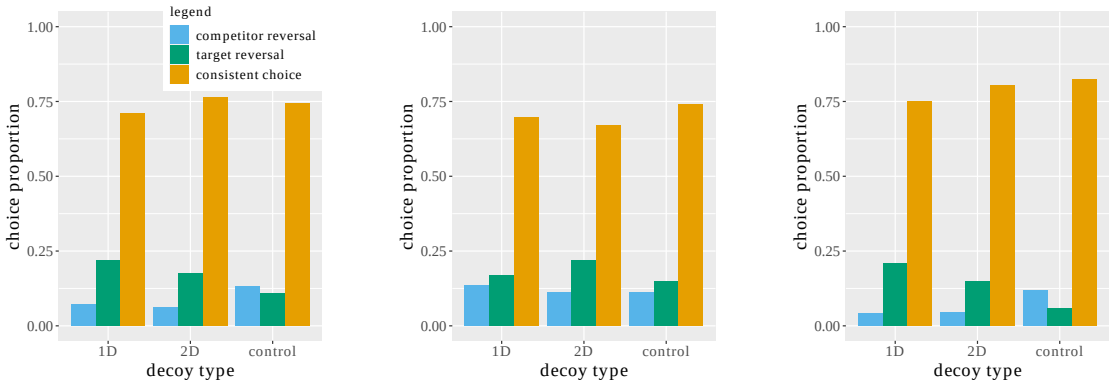
Table A.5: An Example of the Data Coded as Choice Patterns for J Subjects

The table below shows the proportions of choice reversals in Experiment 1, 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.

Decoy Type	Subjects	Total Pairs Done	# of PR	Proportions of PR
R	47	129	14	0.11
F	49	150	20	0.13
RF	47	112	16	0.14
control (R')	49	109	19	0.16

Table A.6: Proportions of Within Subject Choice Reversals (PR) Occurrences in Experiment 1

A.3.2 Full Description Results



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

Figure A.1: Descriptive results — full response patterns in: (a). Wedell(1991); (b). Wedell (1991) replication; (c). Ethical decisions.

A.3.3 Additional Statistical Models

A.3.3.1 Model 1

$$\text{logit } P(Y_{ijk} = A) = \beta^{gm} + \beta_A^{type} X_{ik} + \beta_A^{decoy} X_{ij} \tag{A.1}$$

A.3.3.2 Model 2

In Model 2, we add decoy type R' and estimate the interaction between decoy type and decoy position.

$$\text{logit } P(Y_{ijk} = A) = \beta^{gm} + \beta_A^{type} X_{ik} + \beta_A^{decoy} X_{ij} + \beta_A^{type*decoy} X_{ij} X_{ik} \tag{A.2}$$

A.3.3.3 Priors

Given that we do not have much prior information regarding our model parameters, we choose to select prior distributions that are neither fully informative nor flat. For

the βs, we use weakly informative priors: normal$(0, 5)$.

A.3.4 Full Statistical Results

Parameter	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
Full Model							
β^{gm} (Intercept)	1.00	2773	0.42	0.12	0.18	0.42	0.67
$\beta_A^{R'}$	1.00	2770	0.26	0.23	-0.20	0.26	0.72
β_A^{2D}	1.00	2610	0.12	0.23	-0.34	0.13	0.57
β_A^{decoy}	1.00	2805	0.74	0.19	0.36	0.74	1.12
$\beta_A^{R'*decoy}$	1.00	2462	-0.98	0.33	-1.64	-0.98	-0.34
$\beta_A^{2D*decoy}$	1.00	2496	-0.30	0.34	-0.96	-0.30	0.37
mean_PPD	1.00	4309	0.67	0.02	0.63	0.67	0.71
log-posterior	1.00	1989	-639.18	1.75	-643.46	-638.86	-636.73
Model 1							
β^{gm} (Intercept)	1.00	5231	0.34	0.14	0.07	0.34	0.63
β_A^{R}	1.00	5032	0.25	0.18	-0.11	0.25	0.61
β_A^{RF}	1.00	4355	0.09	0.19	-0.28	0.09	0.47
β_A^{decoy}	1.00	5272	0.66	0.16	0.35	0.66	0.98
mean_PPD	1.00	4987	0.68	0.02	0.63	0.68	0.73
log-posterior	1.00	2155	-479.86	1.43	-483.34	-479.53	-478.12
Model 2							
β^{gm} (Intercept)	1.00	2205	0.29	0.16	-0.03	0.28	0.61
β_A^{R}	1.00	2451	0.31	0.25	-0.18	0.31	0.78
$\beta_A^{R'}$	1.00	2381	0.40	0.25	-0.09	0.39	0.89
β_A^{RF}	1.00	2404	0.25	0.26	-0.25	0.25	0.76
β_A^{decoy}	1.00	1919	0.80	0.25	0.31	0.80	1.30
$\beta_A^{R*decoy}$	1.00	2206	-0.12	0.37	-0.86	-0.12	0.59
$\beta_A^{R'*decoy}$	1.00	2186	-1.05	0.37	-1.76	-1.05	-0.35
$\beta_A^{RF*decoy}$	1.00	2271	-0.36	0.39	-1.11	-0.36	0.43
mean_PPD	1.00	4040	0.67	0.02	0.63	0.67	0.71
log-posterior	1.00	1749	-641.02	2.07	-646.17	-640.62	-638.08

Table A.7: Posterior Statistics for Experiment 1.

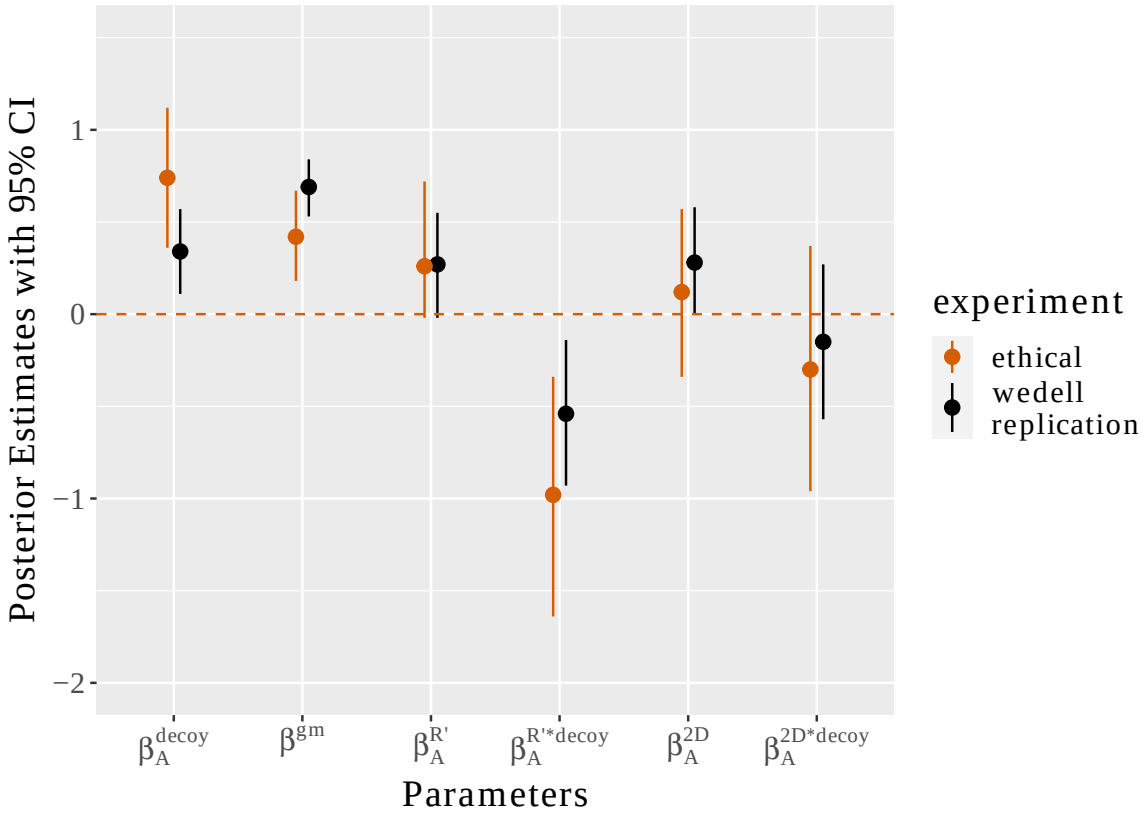


Figure A.2: All posterior estimates for means and 95% CIs of all parameters specified in the logistic regression model in Experiment 1.
Intercept describes the baseline situation — log odds of choosing B when decoy is "atB" and decoy type is 1D. The parameter β_A^{decoy} is our main parameter of interest, which indicate the change in logodds of choosing A to B when decoy is changed from "atB" to "atA". The parameters $\beta_A^{R'}$ and β_A^{2D} describe the effect of decoy type on choice proportions, i.e., the change in logodds of choosing A over B when decoy type is changed from 1D to R'/control or RF/2D. The parameters $\beta_A^{R'*decoy}$ and $\beta_A^{2D*decoy}$ indicates the interaction between decoy position and decoy type effects when considered together with $\beta_A^{R'}$ and β_A^{2D}. For example, by combining β_A^{2D} and $\beta_A^{2D*decoy}$, we learn how logodds of choosing A to B change when decoy is changed from "atB" to "atA" and from 1D to 2D.

A.4 Experiment 2 & 3

A.4.1 Power Analysis

We conducted a power analysis by simulating data and generate posteriors of simulated data with our full model (Equation 2.3) so that we could collect data from a number of subjects that would let us achieve the following level of precision: the width of 95% credible interval (CI) of our main parameter of interest is less than 1. The 95% CI is the central portion of the posterior distribution that contains 95% of the probable effect values.

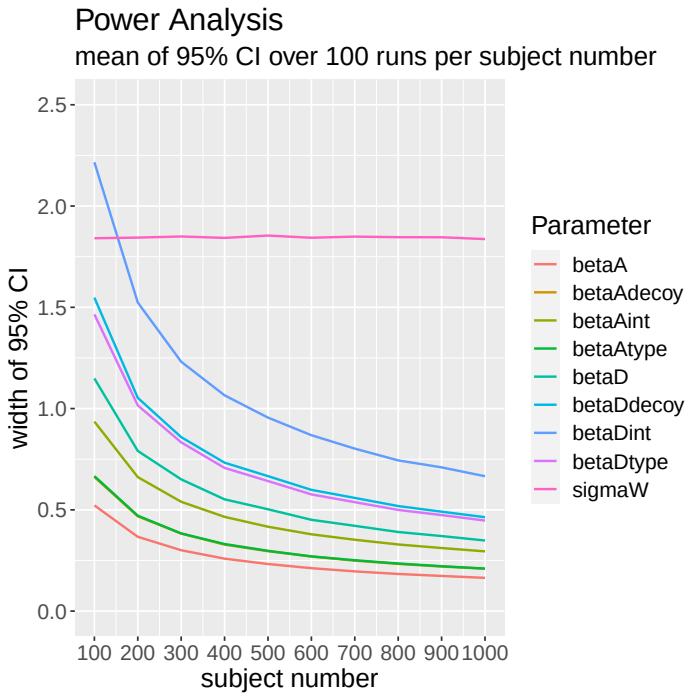


Figure A.3: Power analysis result: change of the width of 95% CI as subject number decreases.

A.4.2 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?

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

A.4.3 Scenarios in Experiment 2 & 3

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.)

Continued on next page

Continued from previous page

Item	Scenario
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 [1]	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.)

Continued on next page

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

Continued from previous page

Item	Scenario
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.)

Table A.8: Items (scenarios) appeared in Experiment 2 and Experiment 3. The *rescue a survivor* item only appeared in Experiment 2.

A.4.4 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 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.

A.4.5 Experiment 2 Results

A.4.5.1 Descriptive Analysis

Analysis with Full Response Patterns Here we include the full response patterns for aggregated data in Experiment 2.

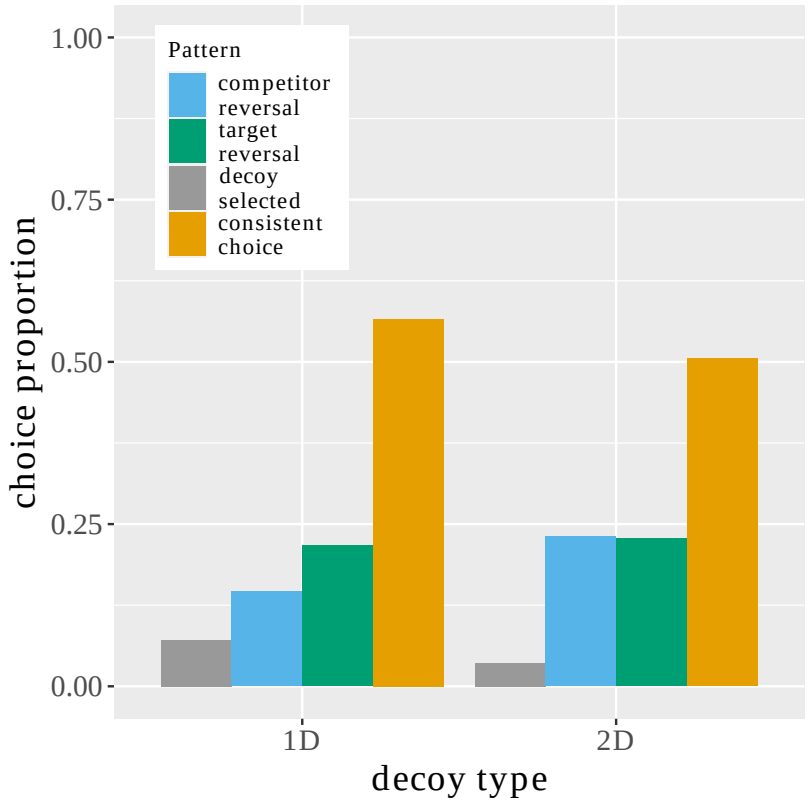


Figure A.4: Full response patterns for data aggregated over all items in Experiment 2 (N=475).

Analyses with all Items Here we include descriptive analyses and data analyses with the full model including the *rescue a survivor* item, which we have excluded in the main chapter.

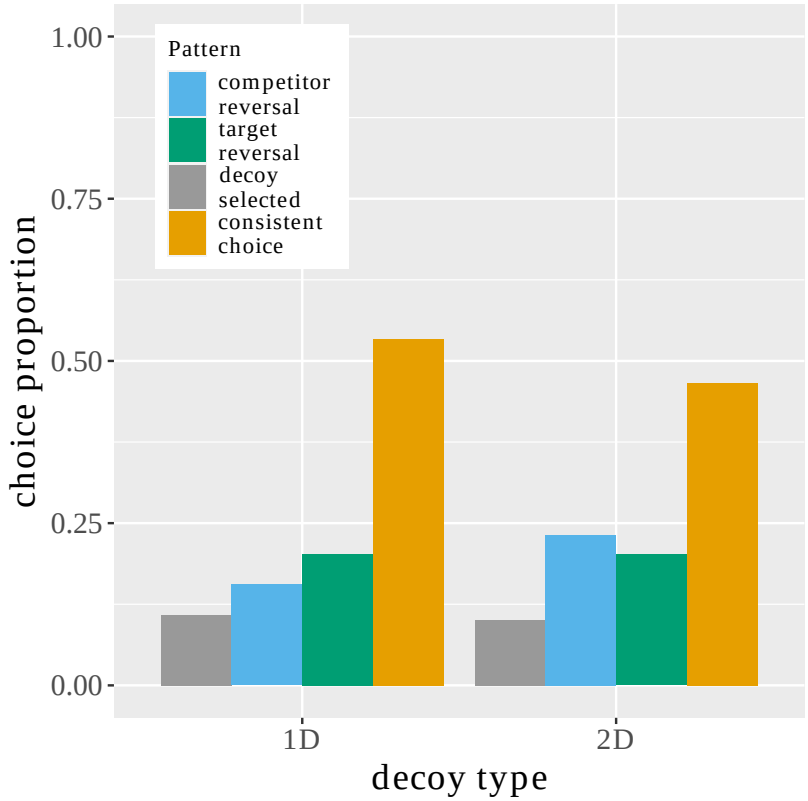


Figure A.5: Response patterns aggregated over all 8 items in Experiment 2 (N=475).

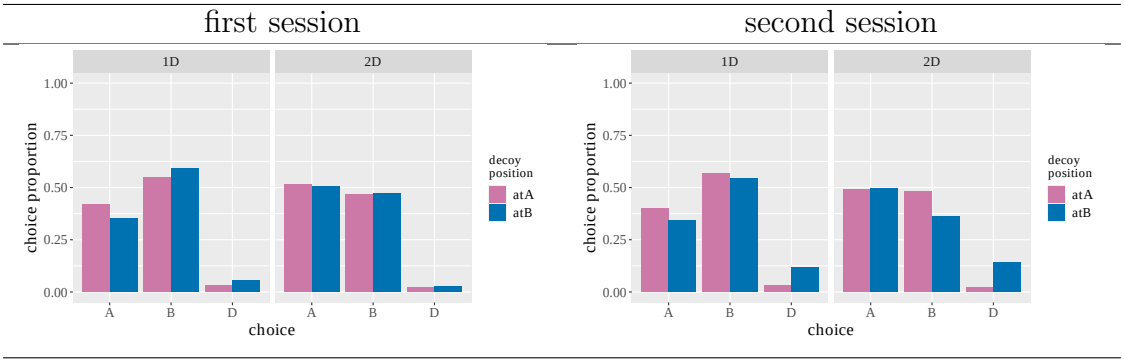
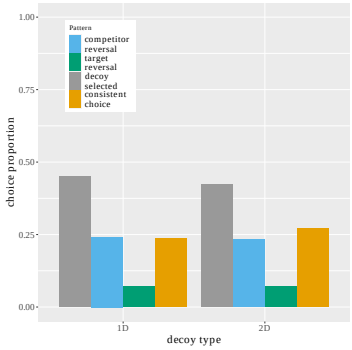
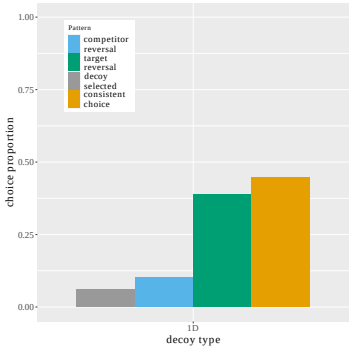


Figure A.6: Aggregated choice proportions for all 8 items during the first and second session in Experiment 2 (N=475).

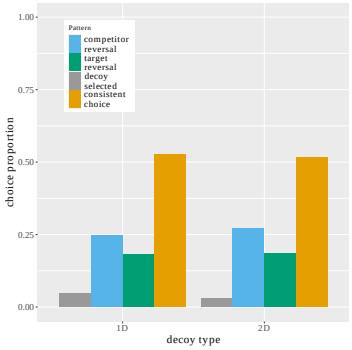
Descriptive Analysis by Item Here we present the descriptive results of choice patterns and choice proportions by item.



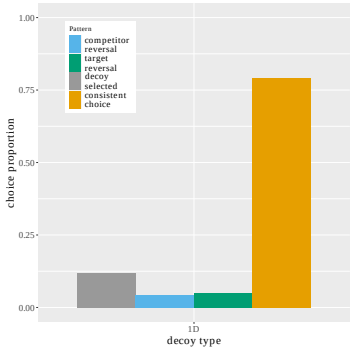
(a) rescue a survivor



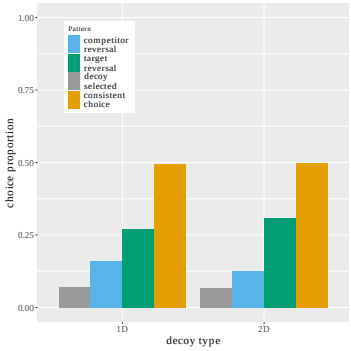
(b) inevitable injury



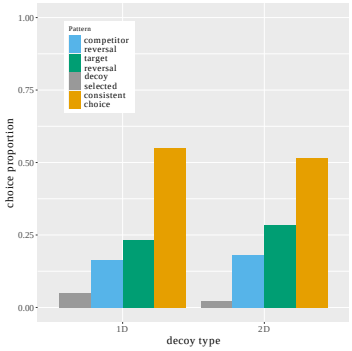
(c) worker welfare



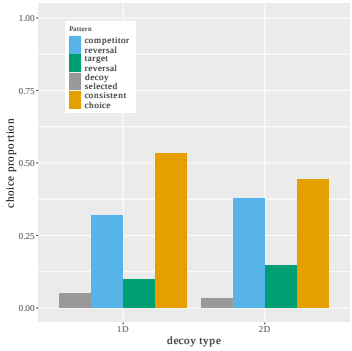
(d) jail overcrowding 2



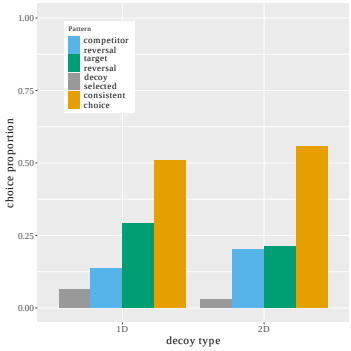
(e) jail overcrowding



(f) rescue plan

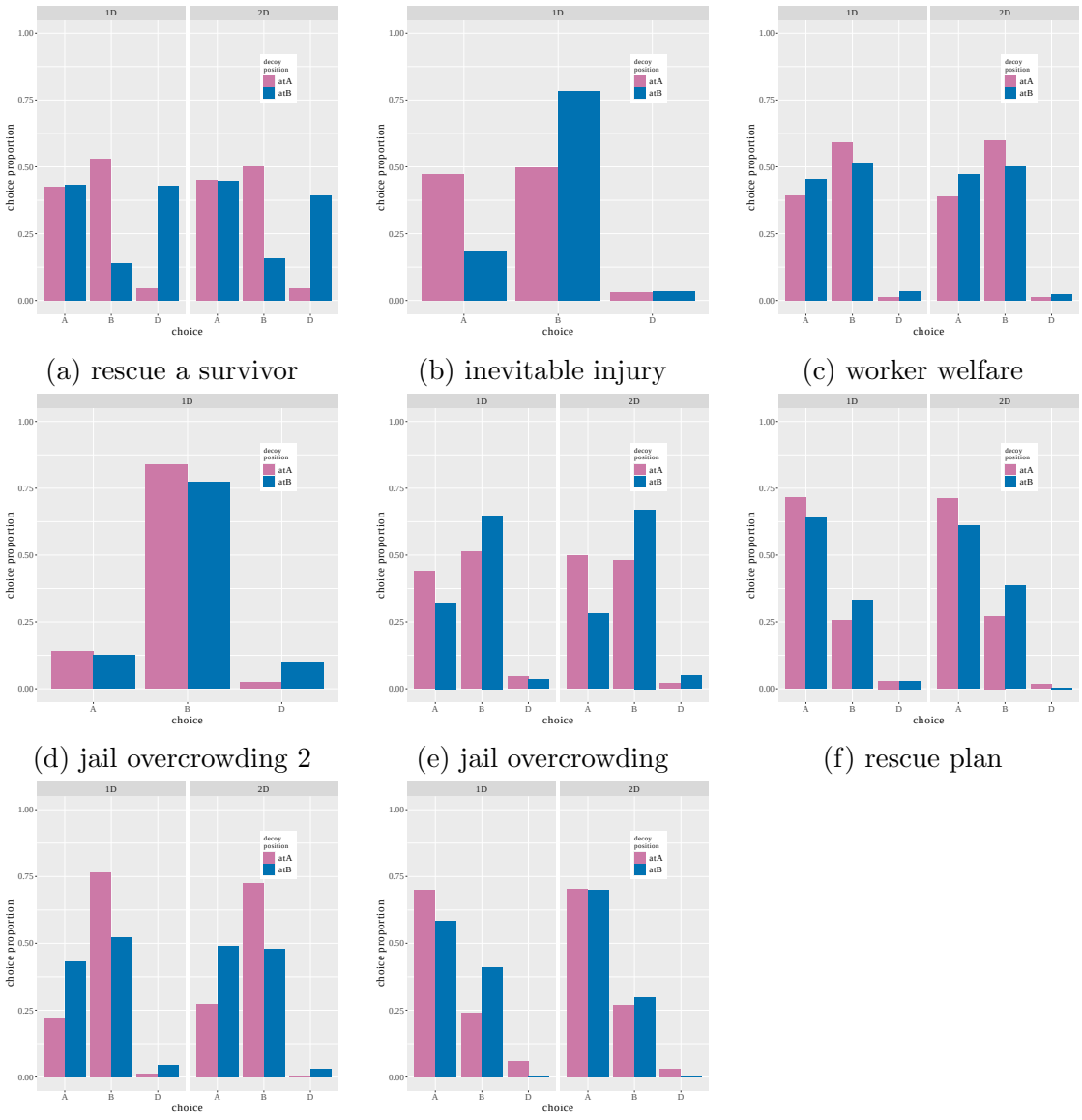


(h) worker welfare 2



(i) emergency delivery

Figure A.7: Choice patterns for all 8 ethical dilemmas.



(a) rescue a survivor

(b) inevitable injury

(c) worker welfare

(d) jail overcrowding 2

(e) jail overcrowding

(f) rescue plan

(g) worker welfare 2

(h) emergency delivery

Figure A.8: Choice proportions for each ethical dilemma.

A.4.5.2 Statistical Results

Full Model First we show the full statistical results for the full model with 7 items (excluding *rescue a survivor* item).

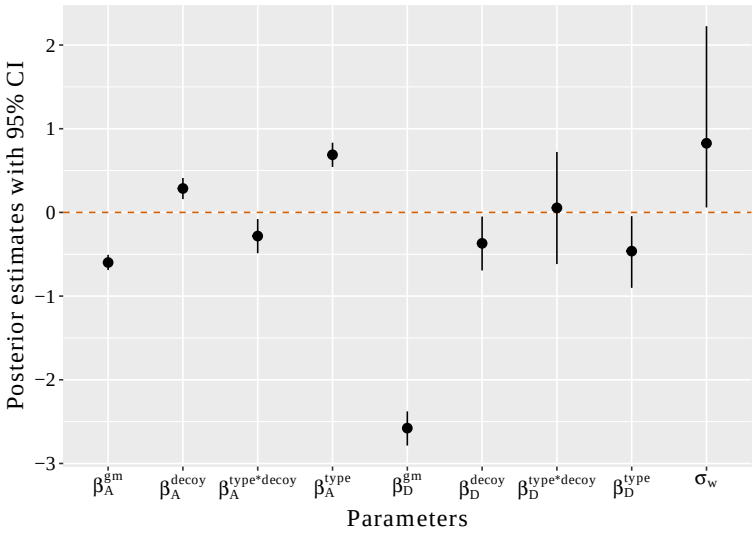
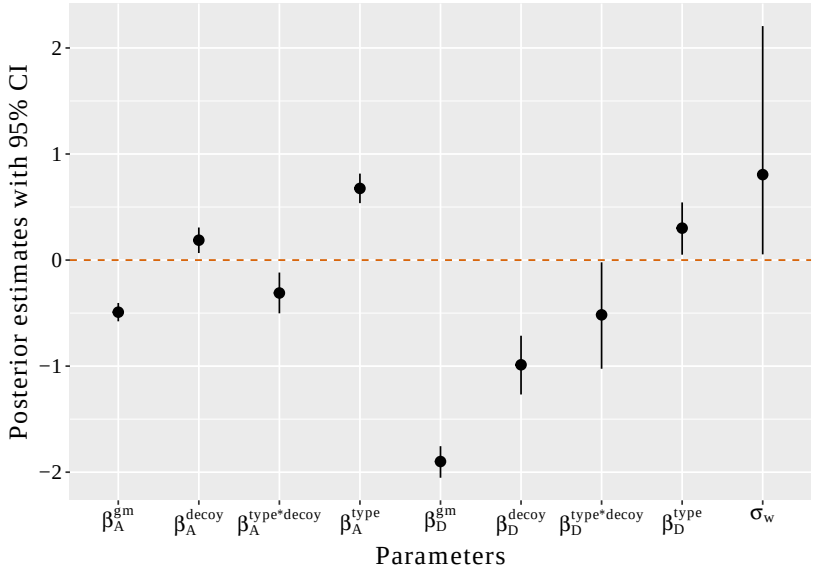


Figure A.9: Experiment 2: posteriors for the full model (7 items).
β_A^{gm} and β_D^{gm} indicate the log odds of choosing A and D when decoy is "atB" and "1D". β_A^{decoy} and β_D^{decoy} indicate the change in log odds of choosing A & D over B when decoy is changed from "atB" to "atA". β_A^{type} and β_D^{type} indicate the change in logodds of choosing A & D over B when decoy is changed to 2D. $\beta_A^{type*decoy}$ and $\beta_D^{type*decoy}$ indicate the interaction of decoy position and type effect on the change in log odds of choosing A & D over B. σ_w estimates the **sd** of the item variations' distribution.

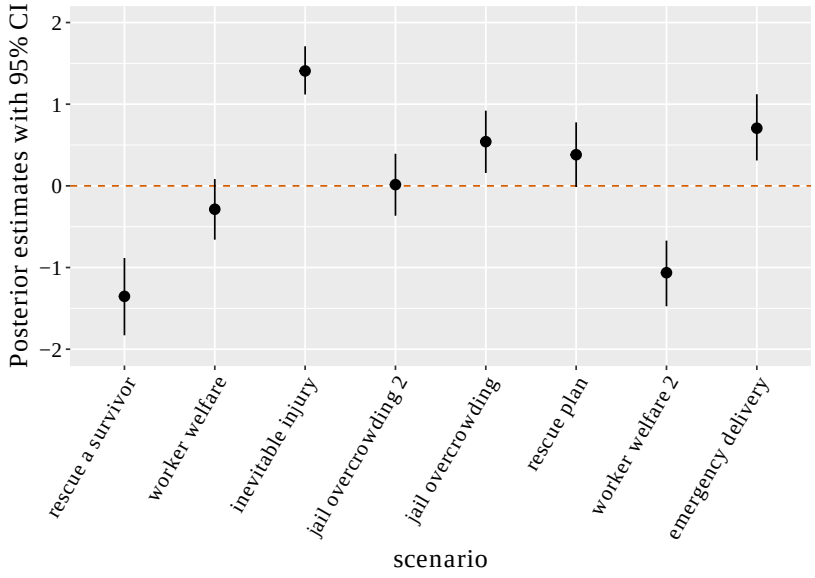
Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	6443.98	-2.58	0.10	-2.79	-2.58	-2.38
β_A^{gm}	1.00	5985.28	-0.60	0.05	-0.69	-0.60	-0.51
β_D^{decoy}	1.00	6344.87	-0.37	0.16	-0.69	-0.37	-0.05
β_A^{decoy}	1.00	5856.80	0.29	0.06	0.16	0.29	0.41
β_D^{type}	1.00	6452.02	-0.46	0.22	-0.90	-0.46	-0.04
β_A^{type}	1.00	6447.36	0.69	0.07	0.54	0.69	0.83
$\beta_D^{type*decoy}$	1.00	5925.03	0.05	0.34	-0.62	0.05	0.72
$\beta_A^{type*decoy}$	1.00	6374.15	-0.28	0.10	-0.49	-0.28	-0.08
σ_w	1.00	870.90	0.83	0.59	0.06	0.71	2.23

Table A.9: Posterior mean and 95% CIs for full model parameters (Experiment 2, excluding *rescue a survivor* item).

Below we show the full statistical results for the full model and simple model including *rescue a survivor* item (Figure A.10).



a. Posteriors for the full model, including all 8 items.



b. Posteriors for β_A^{decoy} for all 8 scenarios.

Figure A.10: Posterior estimates for means and 95% CIs of all parameters in the full model and those for the main parameter of interest, β_A^{decoy}, for each scenario in the simpler model (Experiment 2, including *rescue a survivor* item).

Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	6306.78	-1.90	0.08	-2.05	-1.90	-1.75
β_A^{gm}	1.00	5971.07	-0.49	0.04	-0.58	-0.49	-0.40
β_D^{decoy}	1.00	6601.99	-0.99	0.14	-1.27	-0.99	-0.71
β_A^{decoy}	1.00	5767.18	0.19	0.06	0.07	0.19	0.31
β_D^{type}	1.00	6283.85	0.30	0.13	0.05	0.30	0.54
β_A^{type}	1.00	5722.37	0.68	0.07	0.54	0.67	0.82
$\beta_D^{type*decoy}$	1.00	7459.82	-0.52	0.25	-1.02	-0.52	-0.02
$\beta_A^{type*decoy}$	1.00	5616.11	-0.31	0.10	-0.50	-0.31	-0.12
σ_w	1.01	652.81	0.81	0.59	0.05	0.68	2.21

Table A.10: Posterior mean and 95% CIs for the full model parameters (Experiment 2, including *rescue a survivor* item).

Simple Model Table A.11 below shows the full statistical results of the simple model applied to each item individually.

Scenario	Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
emergency	β_D^{gm}	1.00	3684.58	-4.63	0.94	-6.82	-4.51	-3.11
delivery	β_A^{gm}	1.00	5152.08	0.36	0.13	0.10	0.36	0.62
	β_D^{decoy}	1.00	3808.32	3.19	0.98	1.52	3.08	5.41
	β_A^{decoy}	1.00	5523.29	0.71	0.21	0.31	0.70	1.12
	β_D^{type}	1.00	3923.83	-0.16	1.42	-3.17	-0.10	2.60
	β_A^{type}	1.00	5091.55	0.50	0.19	0.12	0.50	0.88
	$\beta_D^{type*decoy}$	1.00	4019.51	-0.67	1.49	-3.58	-0.72	2.45
	$\beta_A^{type*decoy}$	1.00	5436.92	-0.59	0.29	-1.17	-0.59	-0.04
worker	β_D^{gm}	1.00	6530.94	-2.45	0.32	-3.11	-2.44	-1.87
welfare 2	β_A^{gm}	1.00	6083.69	-0.19	0.13	-0.45	-0.19	0.07
	β_D^{decoy}	1.00	7326.18	-1.80	0.70	-3.30	-1.76	-0.58
	β_A^{decoy}	1.00	5900.86	-1.06	0.21	-1.47	-1.06	-0.67
	β_D^{type}	1.00	6666.71	-0.43	0.51	-1.43	-0.42	0.55
	β_A^{type}	1.00	5667.67	0.20	0.19	-0.16	0.20	0.56
	$\beta_D^{type*decoy}$	1.00	6510.93	-0.97	1.40	-4.01	-0.86	1.56
	$\beta_A^{type*decoy}$	1.00	5375.78	0.06	0.28	-0.48	0.06	0.62
worker	β_D^{gm}	1.00	6276.44	-2.77	0.37	-3.53	-2.75	-2.08
welfare	β_A^{gm}	1.00	5172.19	-0.12	0.13	-0.39	-0.12	0.13
	β_D^{decoy}	1.00	6110.84	-1.22	0.72	-2.76	-1.18	0.09
	β_A^{decoy}	1.00	5378.79	-0.29	0.19	-0.66	-0.29	0.08
	β_D^{type}	1.00	6175.17	-0.30	0.56	-1.42	-0.31	0.79
	β_A^{type}	1.00	5280.66	0.07	0.19	-0.30	0.07	0.43

Continued on next page

Continued from previous page

Scenario	Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
	$\beta_D^{type*decoy}$	1.00	5803.19	0.25	1.04	-1.80	0.24	2.28
	$\beta_A^{type*decoy}$	1.00	5283.28	-0.09	0.27	-0.61	-0.09	0.44
jail over-	β_D^{gm}	1.00	4647.71	-2.97	0.36	-3.72	-2.96	-2.31
crowding	β_A^{gm}	1.00	4566.31	-0.69	0.14	-0.98	-0.69	-0.42
	β_D^{decoy}	1.00	4769.79	0.53	0.48	-0.41	0.53	1.48
	β_A^{decoy}	1.00	4285.51	0.54	0.20	0.16	0.54	0.92
	β_D^{type}	1.00	4667.20	0.34	0.47	-0.56	0.34	1.28
	β_A^{type}	1.00	4393.61	-0.18	0.20	-0.58	-0.18	0.22
	$\beta_D^{type*decoy}$	1.00	5153.47	-1.11	0.73	-2.58	-1.10	0.32
	$\beta_A^{type*decoy}$	1.00	4196.47	0.37	0.28	-0.18	0.37	0.91
rescue a	β_D^{gm}	1.00	3645.03	1.12	0.20	0.74	1.12	1.52
survivor	β_A^{gm}	1.00	2922.31	1.13	0.20	0.75	1.13	1.53
	β_D^{decoy}	1.00	4749.89	-3.58	0.37	-4.32	-3.57	-2.88
	β_A^{decoy}	1.00	3406.11	-1.35	0.24	-1.83	-1.35	-0.88
	β_D^{type}	1.00	3680.75	-0.21	0.27	-0.74	-0.21	0.34
	β_A^{type}	1.00	2922.15	-0.09	0.27	-0.62	-0.09	0.45
	$\beta_D^{type*decoy}$	1.00	4531.21	0.23	0.52	-0.79	0.23	1.25
	$\beta_A^{type*decoy}$	1.00	3436.46	0.21	0.33	-0.44	0.21	0.84
rescue	β_D^{gm}	1.00	6549.29	-2.49	0.40	-3.34	-2.46	-1.76
plan	β_A^{gm}	1.00	5654.01	0.66	0.14	0.39	0.66	0.93
	β_D^{decoy}	1.00	6542.05	0.28	0.58	-0.86	0.28	1.40
	β_A^{decoy}	1.00	5377.63	0.38	0.20	-0.01	0.38	0.78
	β_D^{type}	1.00	4503.86	-2.32	1.13	-4.87	-2.19	-0.45
	β_A^{type}	1.00	5665.24	-0.20	0.19	-0.57	-0.20	0.17
	$\beta_D^{type*decoy}$	1.00	4631.77	1.60	1.31	-0.74	1.53	4.37
	$\beta_A^{type*decoy}$	1.00	5563.97	0.12	0.28	-0.43	0.12	0.68
jail over-	β_D^{gm}	1.00	7408.31	-2.04	0.15	-2.35	-2.04	-1.75
crowding 2	β_A^{gm}	1.00	6505.96	-1.82	0.14	-2.10	-1.82	-1.55
	β_D^{decoy}	1.00	6908.14	-1.58	0.35	-2.30	-1.57	-0.94
	β_A^{decoy}	1.00	5772.39	0.02	0.19	-0.37	0.01	0.39
inevitable	β_D^{gm}	1.00	5808.69	-3.11	0.25	-3.63	-3.09	-2.65
injury	β_A^{gm}	1.00	4989.14	-1.47	0.12	-1.71	-1.46	-1.24
	β_D^{decoy}	1.00	5817.88	0.24	0.38	-0.51	0.24	0.98
	β_A^{decoy}	1.00	4944.34	1.41	0.15	1.12	1.41	1.71

Table A.11: Complete results for simpler model applied to each scenario in Experiment 2.

A.4.6 Experiment 3 Results

A.4.6.1 Descriptive Analysis

Analysis with Full Response Patterns Here we include the full response patterns for aggregated data in Experiment 3.

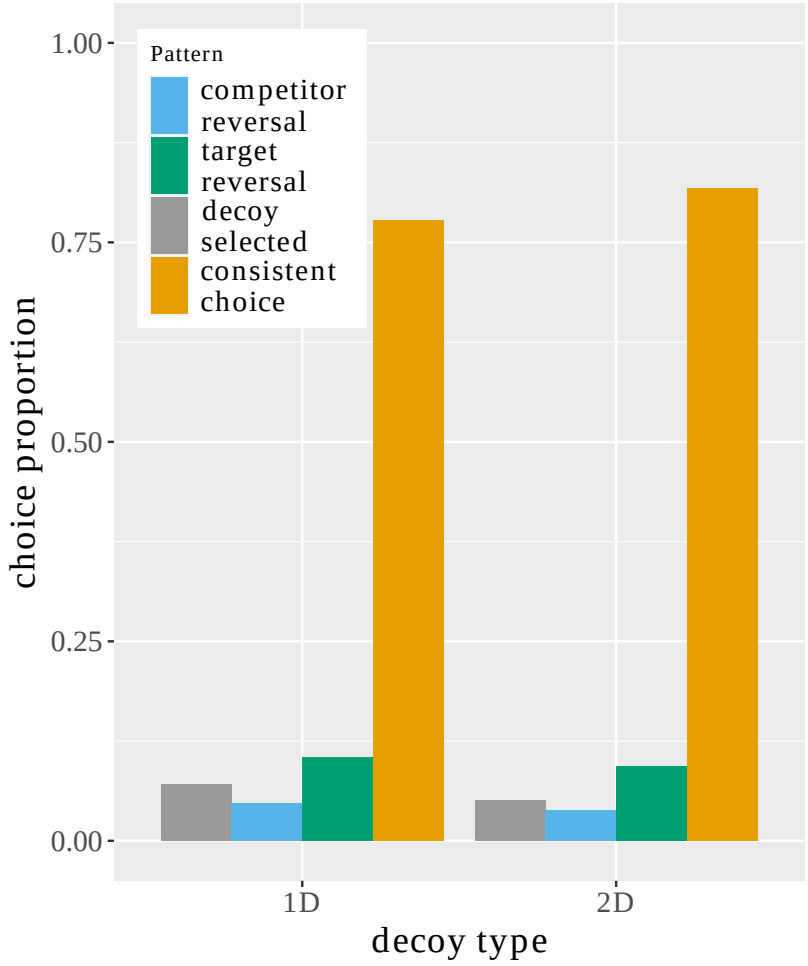


Figure A.11: Aggregated response patterns for all items (excluding *firing an employee*) in Experiment 3 (N = 456).

Analyses with all Items Here we include descriptive analyses and statistical analyses with the full model including the *responsibility & years* item, which we have excluded in the main paper.

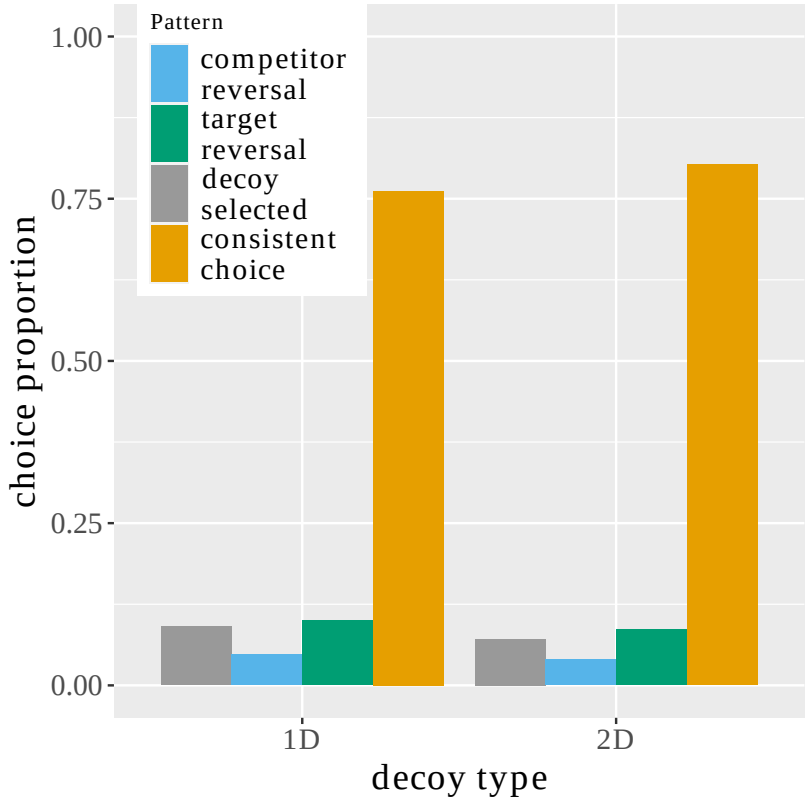


Figure A.12: Response patterns aggregated over all 8 items in Experiment 3 (N=456).

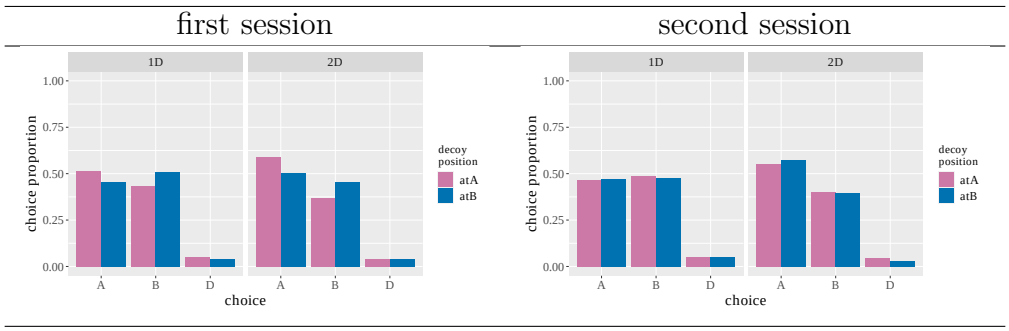
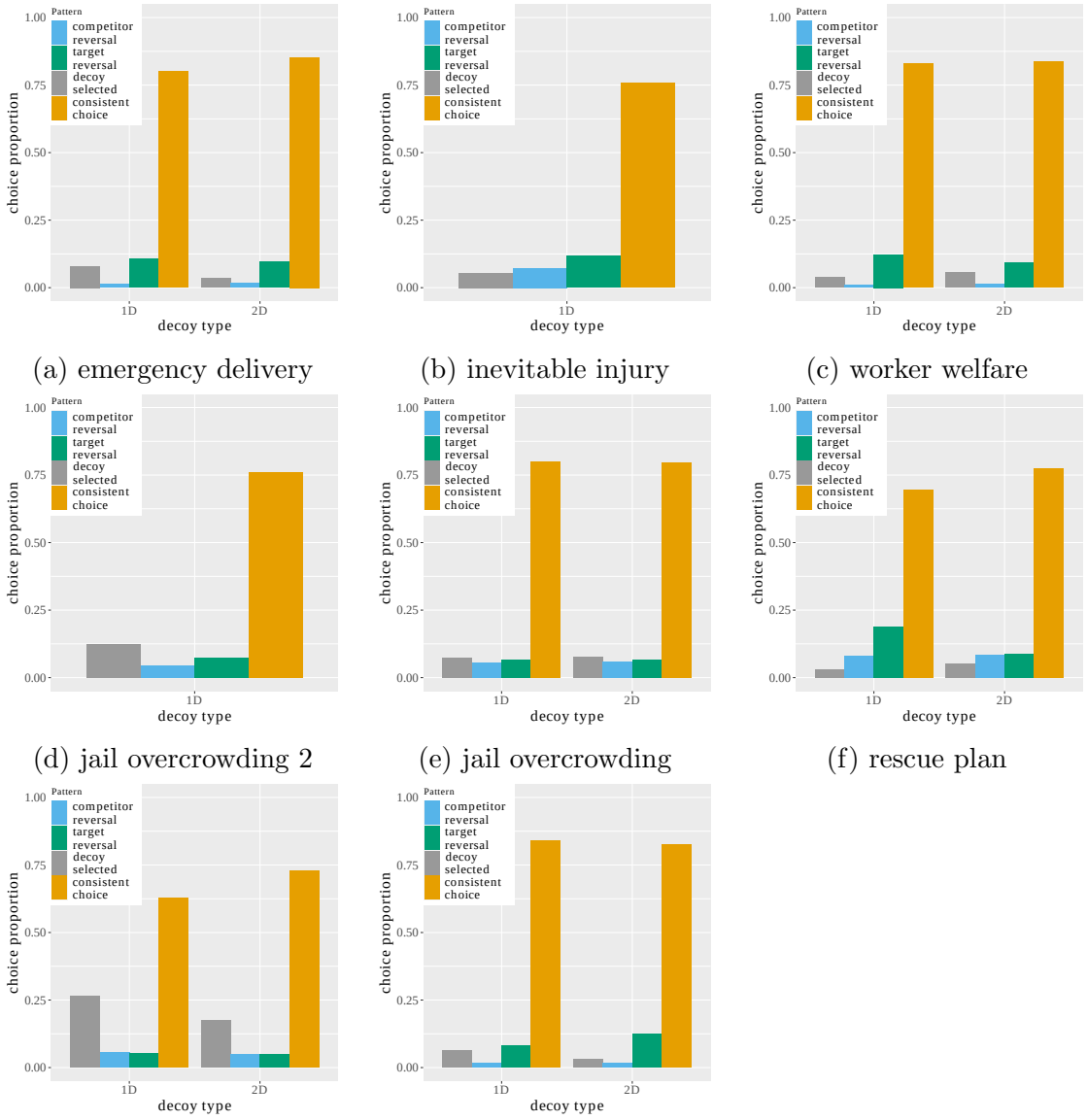


Figure A.13: Aggregated choice proportions for all 8 items during the first and second block in part 2 of Experiment 3 (N=456).

Descriptive Analysis by Item Here we present the descriptive results of choice patterns and choice proportions by item in Experiment 3.

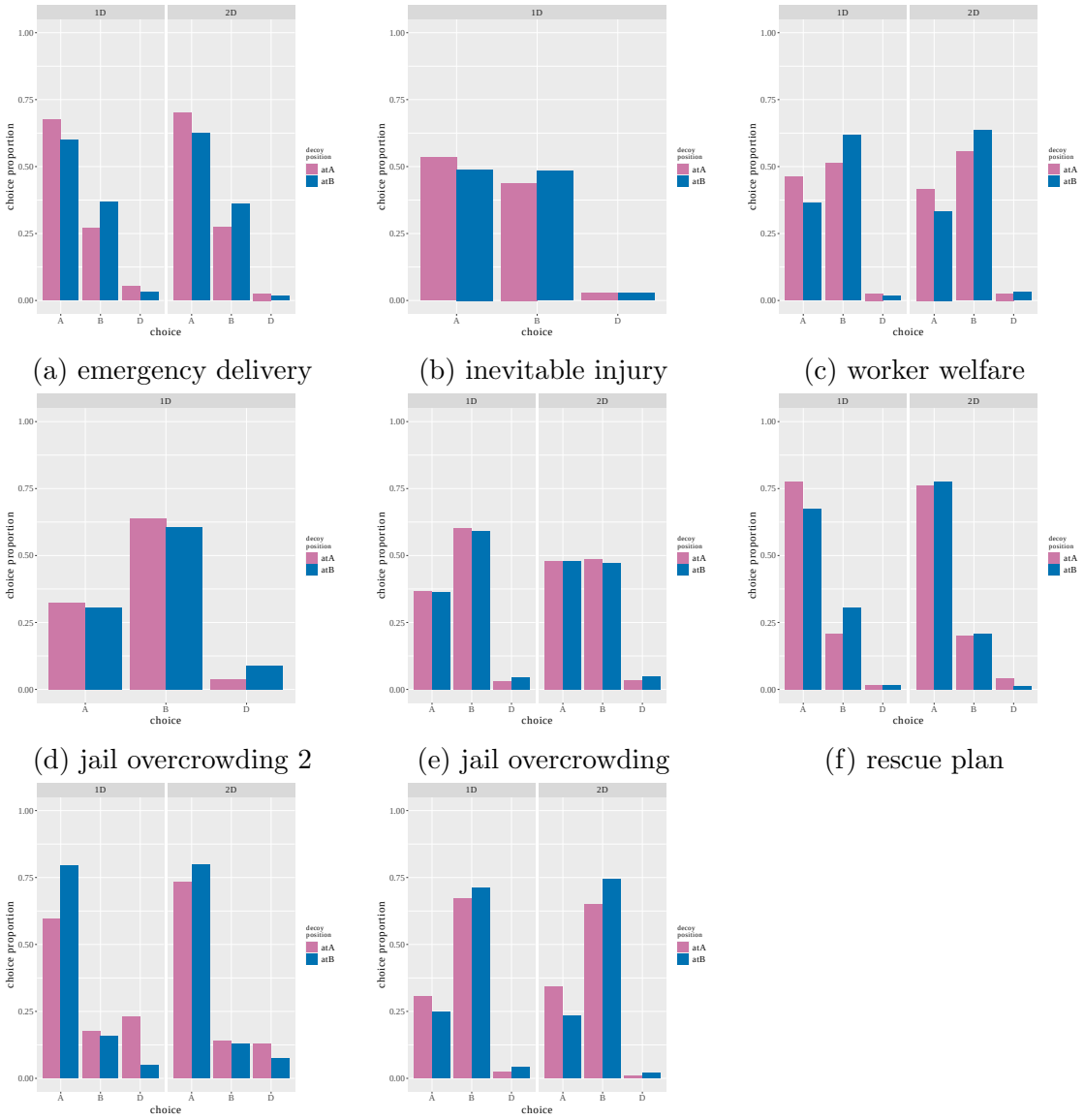


(a) emergency delivery (b) inevitable injury (c) worker welfare

(d) jail overcrowding 2 (e) jail overcrowding (f) rescue plan

(g) firing an employee (h) worker welfare 2

Figure A.14: Experiment 3: choice patterns for all 8 ethical dilemmas.



(a) emergency delivery (b) inevitable injury (c) worker welfare

(d) jail overcrowding 2 (e) jail overcrowding (f) rescue plan

(g) firing an employee (h) worker welfare 2

Figure A.15: Choice proportions for each ethical dilemma in Experiment 3.

A.4.6.2 Statistical Results

Full Model First we show the full statistical results for the full model with 7 items (excluding *responsibility & years*).

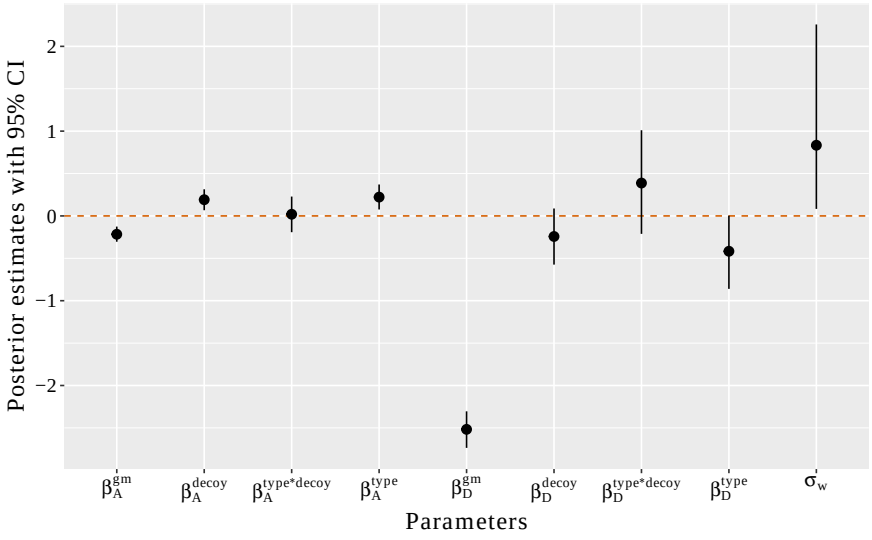
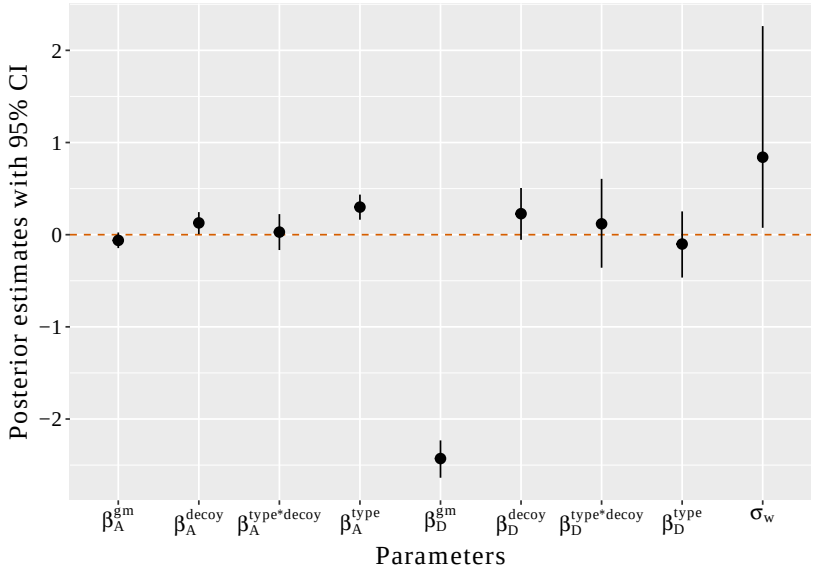


Figure A.16: Posteriors for the full model (Experiment 3).
β_A^{gm} and β_D^{gm} indicate the log odds of choosing A and D when decoy is "atB" and "1D". β_A^{decoy} and β_D^{decoy} indicate the change in log odds of choosing A & D over B when decoy is changed from "atB" to "atA". β_A^{type} and β_D^{type} indicate the change in logodds of choosing A & D over B when decoy is changed to 2D. $\beta_A^{type*decoy}$ and $\beta_D^{type*decoy}$ indicate the interaction of decoy position and type effect on the change in log odds of choosing A & D over B. σ_w estimates the **sd** of the item variations' distribution.

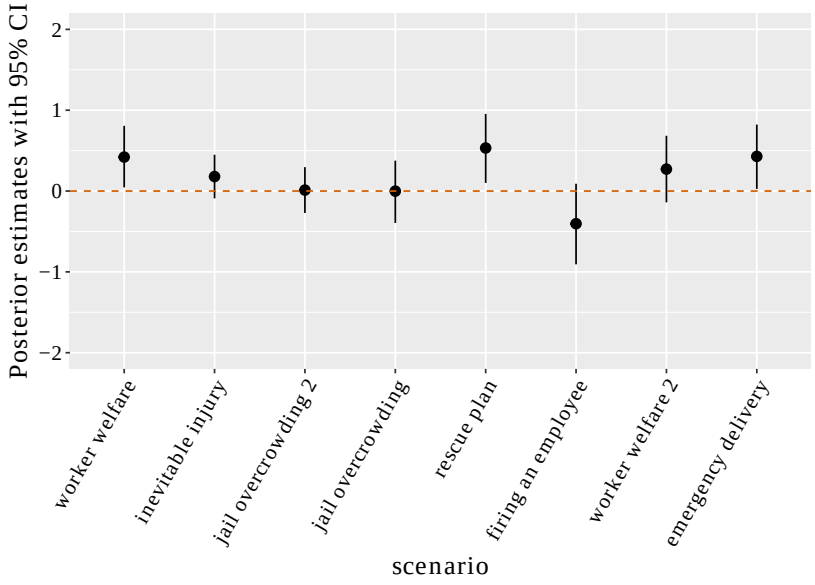
Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	6393.88	-2.52	0.11	-2.74	-2.51	-2.30
β_A^{gm}	1.00	6431.44	-0.22	0.05	-0.30	-0.22	-0.13
β_D^{decoy}	1.00	6382.67	-0.24	0.17	-0.57	-0.24	0.09
β_A^{decoy}	1.00	6122.17	0.19	0.06	0.07	0.19	0.31
β_D^{type}	1.00	6185.96	-0.42	0.22	-0.86	-0.41	0.00
β_A^{type}	1.00	6152.68	0.22	0.08	0.08	0.22	0.37
$\beta_D^{type*decoy}$	1.00	6321.02	0.39	0.31	-0.21	0.38	1.01
$\beta_A^{type*decoy}$	1.00	6212.90	0.02	0.11	-0.19	0.02	0.23
σ_w	1.00	768.18	0.83	0.59	0.08	0.71	2.26

Table A.12: Posterior mean and 95% CIs for the full model parameters (Experiment 3).

Below we show the full statistical results for the full model and simple model including *responsibility & years* item (Figure A.17).



a. Posteriors for the full model.



b. Posteriors for β_A^{decoy} for all scenarios.

Figure A.17: Posterior estimates for means and 95% CIs of all parameters in the full model and those for the main parameter of interest, β_A^{decoy}, for each scenario in the simpler model (Experiment 3, including *responsibility & years* item).

Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	6356.29	-2.43	0.10	-2.64	-2.43	-2.23
β_A^{gm}	1.00	5463.52	-0.06	0.04	-0.15	-0.06	0.02
β_D^{decoy}	1.00	6445.88	0.23	0.14	-0.06	0.23	0.51
β_A^{decoy}	1.00	5372.78	0.13	0.06	0.01	0.13	0.24
β_D^{type}	1.00	5928.12	-0.10	0.18	-0.47	-0.10	0.25
β_A^{type}	1.00	5481.40	0.30	0.07	0.16	0.30	0.43
$\beta_D^{type*decoy}$	1.00	5993.30	0.12	0.25	-0.36	0.11	0.61
$\beta_A^{type*decoy}$	1.00	5612.28	0.03	0.10	-0.17	0.03	0.22
σ_w	1.01	826.37	0.84	0.60	0.07	0.71	2.26

Table A.13: Posterior mean and 95% CIs for the full model parameters (Experiment 3, including *responsibility & years* item).

Full Model Estimating Effects of Instruction Following the same model setup, the full analysis model that estimates the effect of instruction is:

$$\text{categorical } P(Y_{ijkl} = m) = \beta_m^{gm} + \beta_m^{type} X_{ik} + \beta_m^{decoy} X_{ij} + \beta_m^{instr} X_{il} + \beta_m^{interaction} X_{ij} X_{ik} X_{il} \tag{A.3}$$

In the baseline condition where decoy is at B, one-dimensional, and no instruction is given, the equation is simply the log odds for the baseline category:

$$\text{categorical } P(Y_{ijkl} = m) = \beta_m^{gm}$$

.

In Figure A.18 and Table A.14 below, we present the posterior estimates and 95% CIs for the parameters in this model, and we cannot observe any effect of whether instruction on dominance is given or not. In this analysis, all 8 items were included.

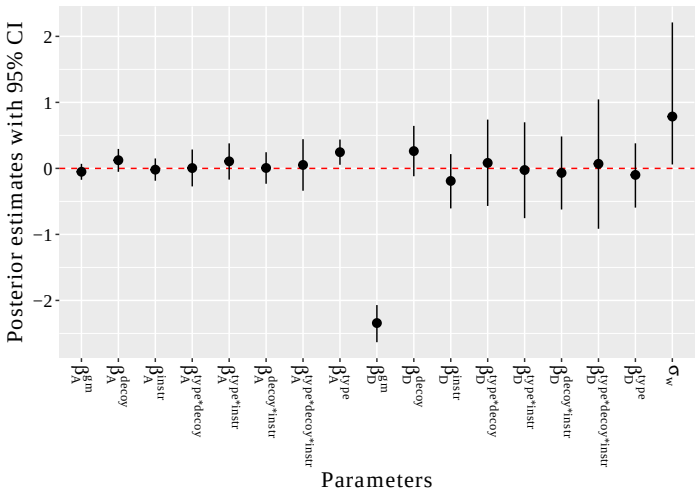


Figure A.18: Posterior estimates for means and 95% CIs of all parameters (Experiment 3, all 8 items included).

Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	4652.01	-2.34	0.14	-2.63	-2.34	-2.07
β_A^{gm}	1.00	5476.76	-0.05	0.06	-0.17	-0.05	0.07
β_D^{decoy}	1.00	4578.80	0.26	0.19	-0.12	0.26	0.64
β_A^{gm}	1.00	5588.49	0.12	0.09	-0.05	0.12	0.29
β_D^{type}	1.00	4508.25	-0.10	0.25	-0.59	-0.09	0.38
β_A^{type}	1.00	5420.15	0.25	0.10	0.05	0.24	0.44
β_D^{instr}	1.00	4817.09	-0.19	0.21	-0.61	-0.19	0.22
β_A^{instr}	1.00	5525.40	-0.02	0.09	-0.19	-0.02	0.15
$\beta_D^{type*decoy}$	1.00	4617.72	0.08	0.33	-0.57	0.08	0.74
$\beta_A^{type*decoy}$	1.00	5355.72	0.01	0.14	-0.27	0.00	0.29
$\beta_D^{type*instr}$	1.00	5058.02	-0.02	0.37	-0.75	-0.03	0.70
$\beta_A^{type*instr}$	1.00	5626.59	0.11	0.14	-0.17	0.11	0.38
$\beta_D^{decoy*instr}$	1.00	4607.28	-0.07	0.28	-0.62	-0.07	0.48
$\beta_A^{decoy*instr}$	1.00	5457.34	0.01	0.12	-0.23	0.01	0.24
$\beta_D^{type*decoy*instr}$	1.00	5107.27	0.07	0.50	-0.91	0.07	1.05
$\beta_A^{type*decoy*instr}$	1.00	5384.01	0.05	0.20	-0.34	0.05	0.44
σ_w	1.01	890.79	0.79	0.59	0.06	0.66	2.21

Table A.14: Posteriors for the full analysis model including parameters estimating the effect of instructions.

Interactions: $\beta_A^{type*decoy}$ indicates the interaction between decoy position and decoy type; $\beta_A^{type*instr}$ indicates the interaction between decoy type and instruction; $\beta_A^{decoy*instr}$ indicates the interaction between decoy position and instruction; $\beta_A^{type*decoy*instr}$ indicates the three-way interaction. All interactions are inconclusive as their 95% CIs include 0.

Simple Model Table A.15 below shows the full statistical results of the simple model applied to each item individually.

Scenario	Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
emergency delivery	β_D^{gm}	1.00	4912.47	-2.52	0.40	-3.36	-2.49	-1.81
	β_A^{gm}	1.00	5675.58	0.49	0.14	0.23	0.49	0.76
	β_D^{decoy}	1.00	4972.94	0.85	0.51	-0.13	0.84	1.89
	β_A^{decoy}	1.00	5281.71	0.43	0.20	0.03	0.43	0.82
	β_D^{type}	1.00	5028.99	-0.63	0.66	-1.95	-0.63	0.62
	β_A^{type}	1.00	5230.20	0.06	0.19	-0.32	0.07	0.44
	$\beta_D^{type*decoy}$	1.00	5103.23	-0.14	0.84	-1.77	-0.14	1.50
	$\beta_A^{type*decoy}$	1.00	5075.42	-0.03	0.28	-0.58	-0.04	0.53
responsibility & years	β_D^{gm}	1.00	3586.80	-1.20	0.35	-1.92	-1.18	-0.55
	β_A^{gm}	1.00	4099.60	1.63	0.18	1.29	1.63	2.00
	β_D^{decoy}	1.00	3059.20	1.45	0.40	0.69	1.44	2.28
	β_A^{decoy}	1.00	3959.04	-0.40	0.25	-0.91	-0.40	0.09
	β_D^{type}	1.00	3558.91	0.64	0.46	-0.24	0.64	1.57
	β_A^{type}	1.00	4263.76	0.21	0.27	-0.32	0.21	0.75
	$\beta_D^{type*decoy}$	1.00	3379.38	-1.00	0.56	-2.12	-1.00	0.09
	$\beta_A^{type*decoy}$	1.00	4201.36	0.22	0.37	-0.51	0.22	0.96
worker welfare 2	β_D^{gm}	1.00	5725.93	-2.92	0.35	-3.64	-2.91	-2.29
	β_A^{gm}	1.00	5757.03	-1.06	0.16	-1.37	-1.05	-0.76
	β_D^{decoy}	1.00	5716.62	-0.58	0.59	-1.77	-0.57	0.52
	β_A^{decoy}	1.00	5790.26	0.27	0.21	-0.14	0.27	0.68
	β_D^{type}	1.00	5849.87	-0.72	0.58	-1.89	-0.70	0.38
	β_A^{type}	1.00	5796.39	-0.11	0.22	-0.55	-0.11	0.33
	$\beta_D^{type*decoy}$	1.00	6095.62	-0.35	1.06	-2.53	-0.31	1.66
	$\beta_A^{type*decoy}$	1.00	5829.21	0.25	0.30	-0.33	0.25	0.84
worker welfare	β_D^{gm}	1.00	4015.51	-3.62	0.51	-4.73	-3.58	-2.72
	β_A^{gm}	1.00	4662.10	-0.53	0.14	-0.81	-0.53	-0.26
	β_D^{decoy}	1.00	4305.08	0.55	0.66	-0.74	0.55	1.87
	β_A^{decoy}	1.00	4648.17	0.42	0.19	0.05	0.42	0.81
	β_D^{type}	1.00	4201.99	0.51	0.64	-0.72	0.50	1.83
	β_A^{type}	1.00	4627.35	-0.12	0.20	-0.51	-0.12	0.28
	$\beta_D^{type*decoy}$	1.00	4426.40	-0.57	0.88	-2.34	-0.56	1.15
	$\beta_A^{type*decoy}$	1.00	4709.15	-0.06	0.28	-0.60	-0.06	0.48
jail over-crowding	β_D^{gm}	1.00	5542.58	-2.63	0.33	-3.32	-2.62	-2.04
	β_A^{gm}	1.00	5182.19	-0.49	0.14	-0.77	-0.49	-0.22
	β_D^{decoy}	1.00	5613.81	-0.40	0.52	-1.44	-0.40	0.60
	β_A^{decoy}	1.00	5253.68	-0.00	0.20	-0.39	0.00	0.38

Continued on next page

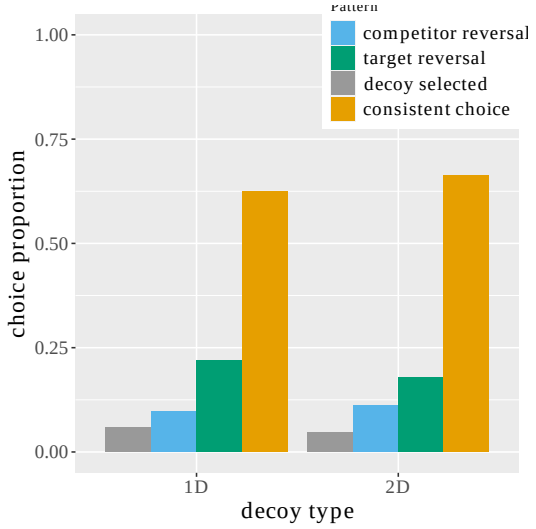
Continued from previous page

Scenario	Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
	β_D^{type}	1.00	5575.31	0.30	0.46	-0.59	0.29	1.21
	β_A^{type}	1.00	5063.08	0.50	0.19	0.12	0.50	0.88
	$\beta_D^{type*decoy}$	1.00	5788.67	0.04	0.71	-1.36	0.05	1.45
	$\beta_A^{type*decoy}$	1.00	5201.39	-0.03	0.27	-0.57	-0.03	0.51
rescue plan	β_D^{gm}	1.00	4956.11	-2.96	0.52	-4.06	-2.94	-2.03
	β_A^{gm}	1.00	5061.93	0.79	0.14	0.51	0.79	1.08
	β_D^{decoy}	1.00	4932.31	0.38	0.74	-1.09	0.39	1.80
	β_A^{decoy}	1.00	5136.61	0.53	0.22	0.10	0.53	0.95
	β_D^{type}	1.00	4857.66	0.06	0.81	-1.57	0.07	1.62
	β_A^{type}	1.00	4681.02	0.53	0.22	0.10	0.53	0.96
	$\beta_D^{type*decoy}$	1.00	4673.66	0.86	1.03	-1.10	0.85	2.93
	$\beta_A^{type*decoy}$	1.00	4874.43	-0.51	0.32	-1.15	-0.51	0.13
jail over-crowding 2	β_D^{gm}	1.00	6308.33	-1.92	0.17	-2.25	-1.91	-1.60
	β_A^{gm}	1.00	5891.53	-0.69	0.10	-0.90	-0.69	-0.49
	β_D^{decoy}	1.00	6554.10	-0.95	0.31	-1.56	-0.95	-0.37
	β_A^{decoy}	1.00	5921.02	0.01	0.15	-0.27	0.01	0.30
inevitable injury	β_D^{gm}	1.00	5724.02	-2.85	0.29	-3.45	-2.84	-2.32
	β_A^{gm}	1.00	5896.67	0.02	0.10	-0.17	0.02	0.20
	β_D^{decoy}	1.00	5605.70	0.08	0.41	-0.73	0.08	0.89
	β_A^{decoy}	1.00	5592.61	0.18	0.14	-0.09	0.18	0.45

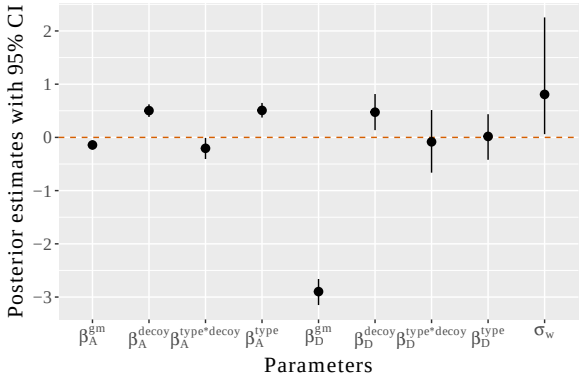
Table A.15: Complete results for simpler model applied to each scenario in Experiment 3.

A.4.7 Combined Results

Analyses of Shared Items from Experiment 2 and 3 Here we present descriptive and statistical results of the shared items.



a. Full response patterns aggregated over shared items from both Experiment 2 and 3.



b. Posteriors from the full model analysis of shared items from Experiment 2 and 3: β_A^{gm} and β_D^{gm} indicate the log odds of choosing A and D when decoy is "atB" and "1D". β_A^{decoy} and β_D^{decoy} indicate the change in log odds of choosing A & D over B when decoy is changed from "atB" to "atA". β_A^{type} and β_D^{type} indicate the change in logodds of choosing A & D over B when decoy is changed to 2D. $\beta_A^{type*decoy}$ and $\beta_D^{type*decoy}$ indicate the interaction of decoy position and type effect on the change in log odds of choosing A & D over B. σ_w estimates the **sd** of the item variations' distribution.

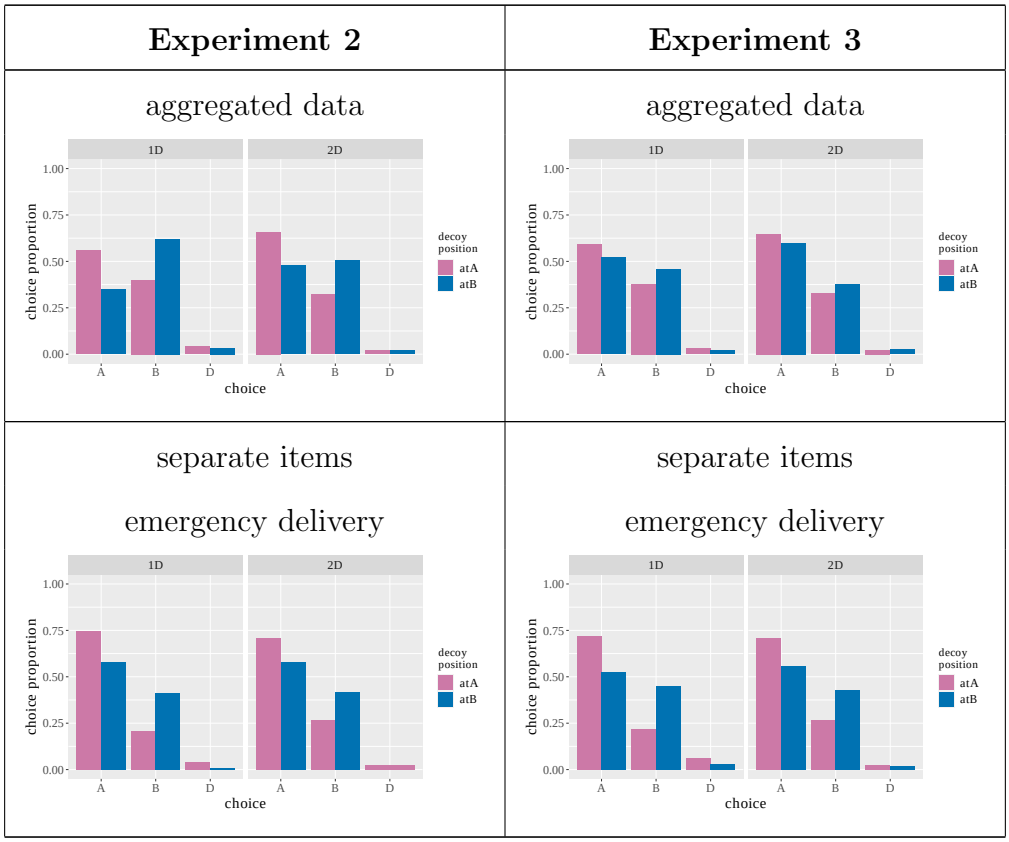
Figure A.19: Descriptive and statistical results for shared items aggregated in Experiment 2 and 3 combined data (N=931).

Parameters	Rhat	n_eff	mean	sd	2.5%	50%	97.5%
β_D^{gm}	1.00	5476.68	-2.90	0.13	-3.15	-2.90	-2.66
β_A^{gm}	1.00	5352.07	-0.14	0.04	-0.23	-0.14	-0.06
β_D^{decoy}	1.00	5280.01	0.47	0.17	0.14	0.47	0.81
β_A^{decoy}	1.00	5733.58	0.50	0.06	0.38	0.50	0.62
β_D^{type}	1.00	5806.07	0.02	0.22	-0.42	0.02	0.44
β_A^{type}	1.00	5402.03	0.51	0.07	0.37	0.51	0.65
$\beta_D^{type*decoy}$	1.00	5884.75	-0.08	0.30	-0.66	-0.09	0.51
$\beta_A^{type*decoy}$	1.00	5435.22	-0.21	0.10	-0.41	-0.20	-0.01
σ_w	1.00	1329.03	0.81	0.60	0.06	0.67	2.25

Table A.16: Posterior estimates for mean and 95% CIs for combined shared scenarios in Experiment 2 and 3.

A.4.7.1 First Occurrences of Scenarios

Here we present the results of choice proportions for the first time when a scenario is presented. Table A.17 below contains each scenario that is **shared** in experiment 2 & 3.



Continued on next page

Continued from previous page

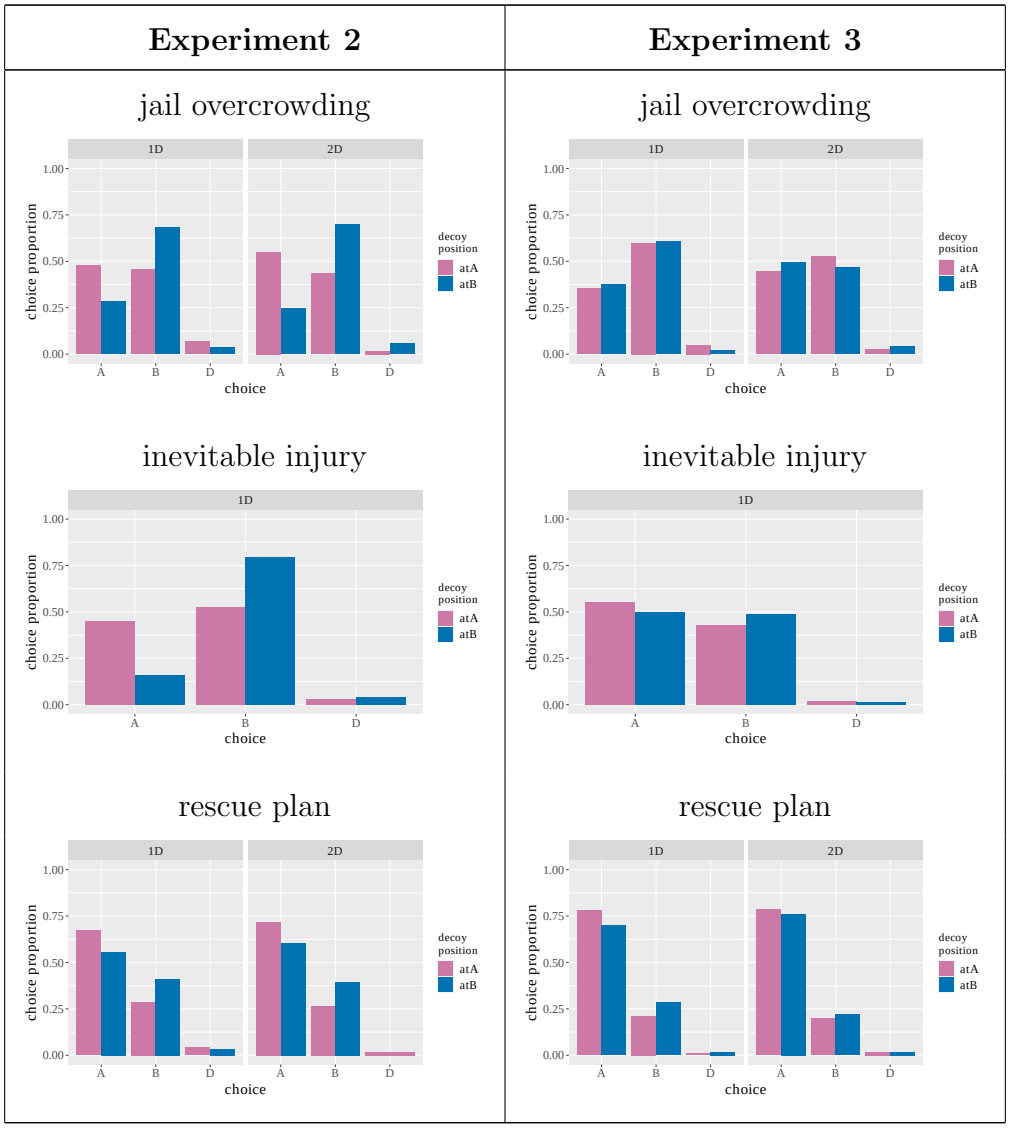
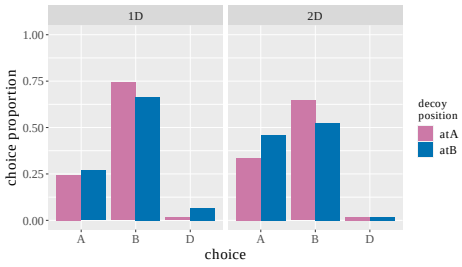
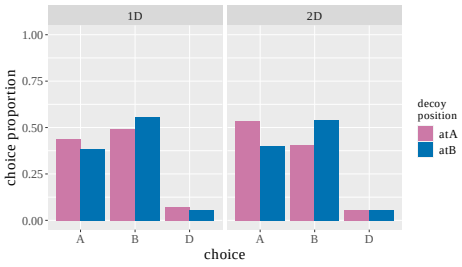
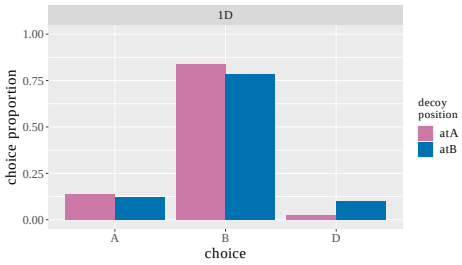
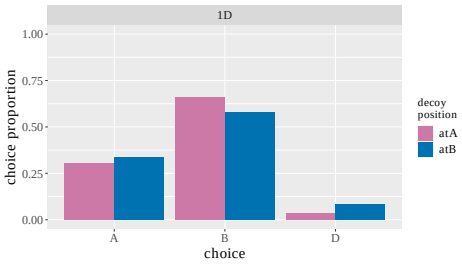
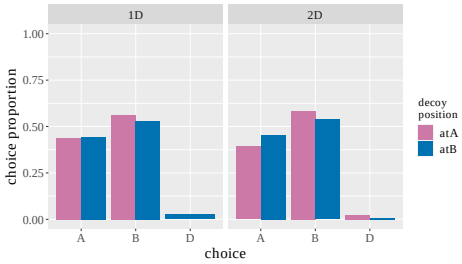
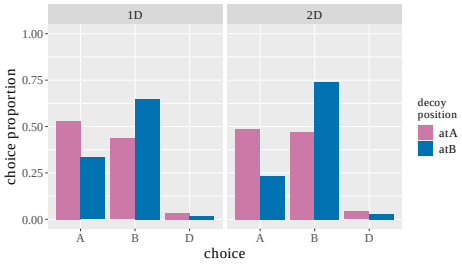
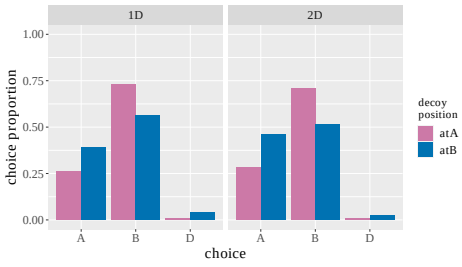
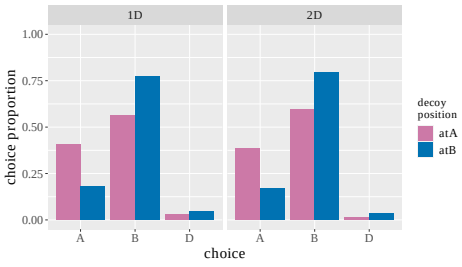


Table A.17: Choice proportions in the first occurrences of each shared scenario in Experiment 2 & 3.

Here we present the results of choice proportions for the first time when a scenario is presented. Table A.18 below contains the scenarios that were **different** in experiment 2 & 3, including the *rescue a survivor* item in Experiment 2 and the *responsibility & years* item in Experiment 3.

Experiment 2	Experiment 3
aggregated data	aggregated data
	
separate items	separate items
jail overcrowding 2	jail overcrowding 2
	
worker welfare	worker welfare
	
worker welfare 2	worker welfare 2
	
rescue a survivor	

Continued on next page

Continued from previous page

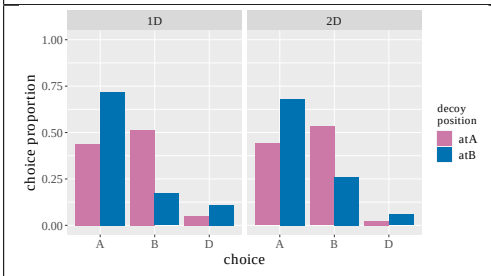
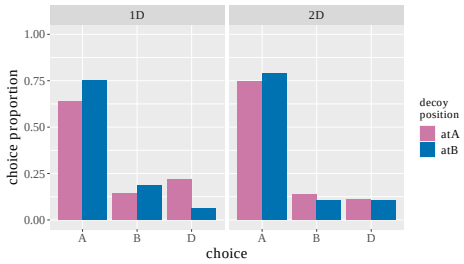
Experiment 2	Experiment 3
	firing an employee 

Table A.18: Choice proportions in the first occurrences of each revised scenario in Experiment 2 & 3.

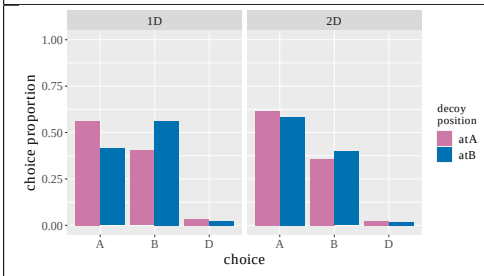
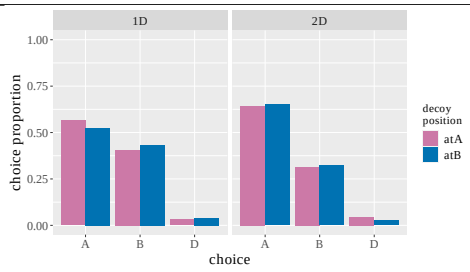
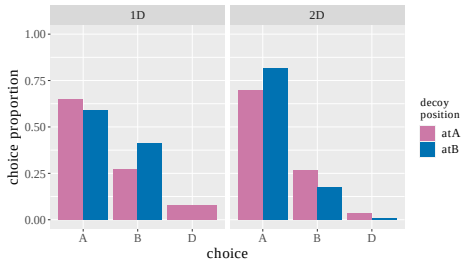
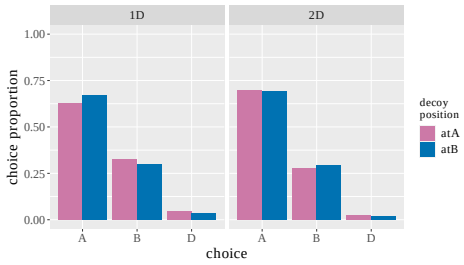
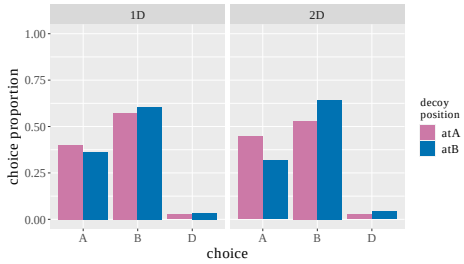
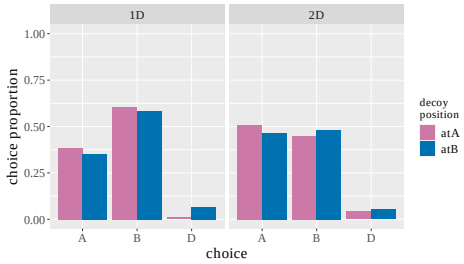
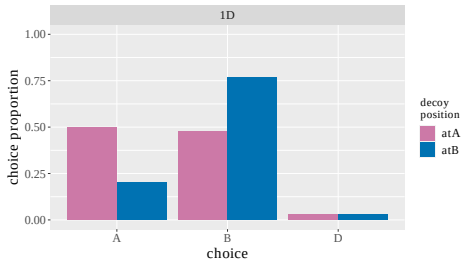
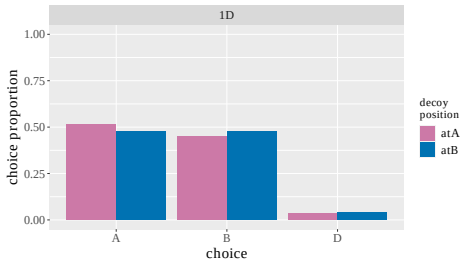
A.4.7.2 Second Occurrences of Scenarios

Here we present the results of choice proportions for the second time when a scenario is presented. Table A.19 below contains the scenarios that were **shared** in experiment 2 & 3.

Experiment 2	Experiment 3
aggregated data	aggregated data

Continued on next page

Continued from previous page

Experiment 2	**Experiment 3**
	
separate items	separate items
emergency delivery	emergency delivery
	
jail overcrowding	jail overcrowding
	
inevitable injury	inevitable injury
	
rescue plan	rescue plan

Continued on next page

Continued from previous page

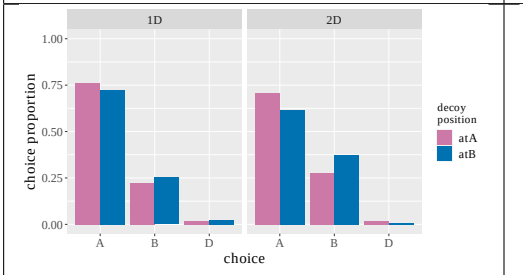
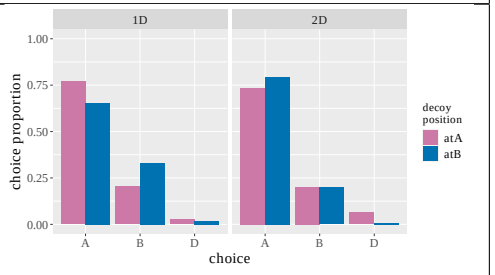
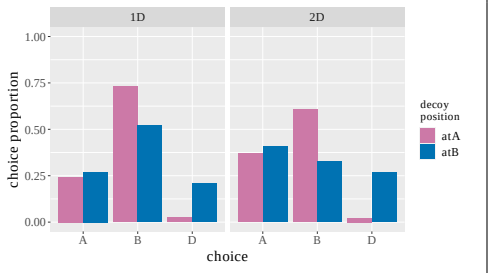
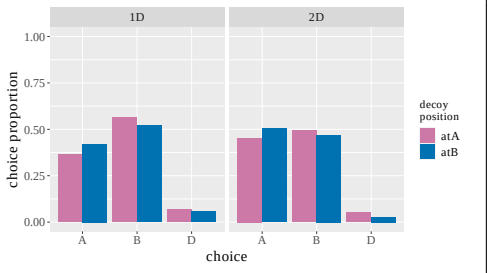
Experiment 2	**Experiment 3**
	

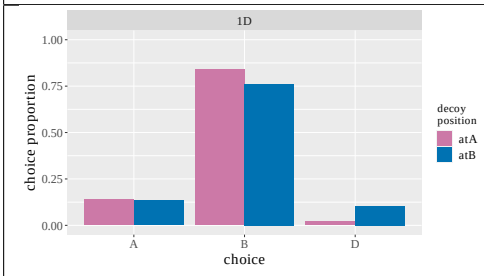
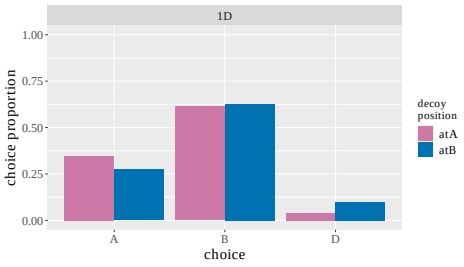
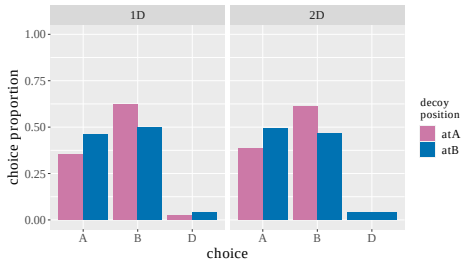
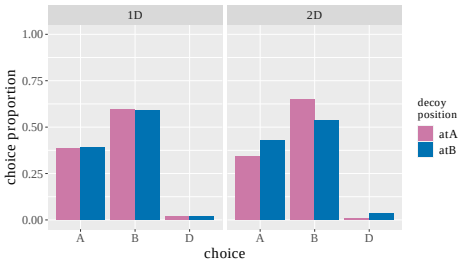
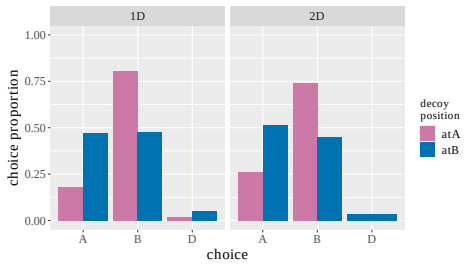
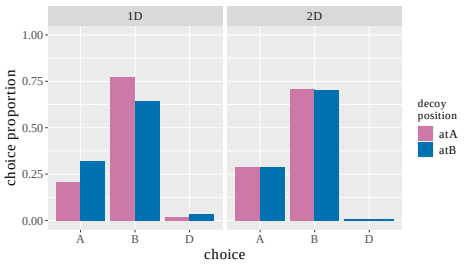
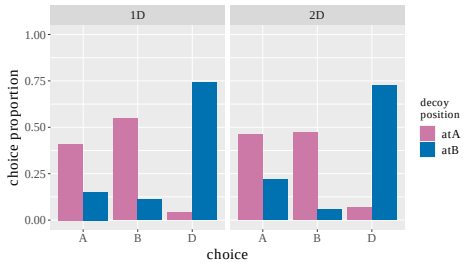
Table A.19: Choice proportions in the second occurrences of each shared scenario in Experiment2 & 3.

Here we present the results of choice proportions for the second time when a scenario is presented. Table A.20 below contains the scenarios that were **different** in experiment 2 & 3, including the *rescue a survivor* item in Experiment 2 and the *responsibility & years* item in Experiment 3.

Experiment 2	**Experiment 3**
aggregated data	aggregated data
	
separate items	separate items
jail overcrowding 2	jail overcrowding 2

Continued on next page

Continued from previous page

Experiment 2	Experiment 3
	
worker welfare	worker welfare
	
worker welfare 2	worker welfare 2
	
rescue a survivor	
	firing an employee

Continued on next page

Continued from previous page

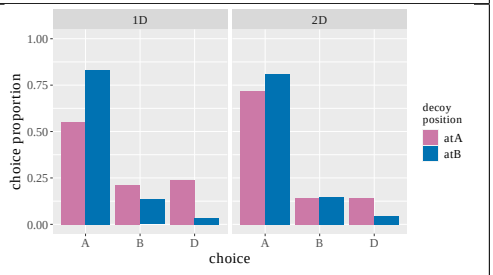
Experiment 2	Experiment 3
	

Table A.20: Choice proportions in the second occurrences of each revised scenario in Experiment2 & 3.

APPENDIX B

Supplemental Materials for the Generative Model of Response Patterns in Ethical Decisions

B.1 Simulation Results

Here we present the simulated and empirical results for "consistent choice" selection rates and "competitor reversal" rates.

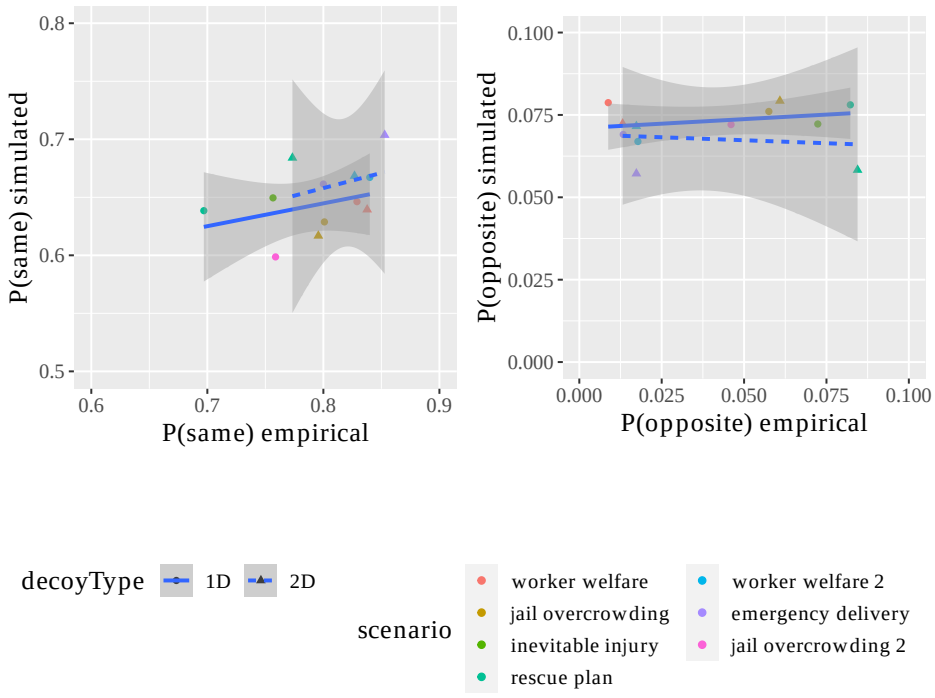


Figure B.1: Simulated and empirical "consistent choice" selection rates (left) and "competitor reversal" rates (right) for each scenario and decoy type.
Generally, simulated same selection rates were lower than those in empirical data and simulated "competitor reversal" selection rates were higher than those in empirical data (Items are from Experiment 3, excluding the *responsibility & years* item).

Below we present the complete results including the *responsibility & years* item in Figure B.2.

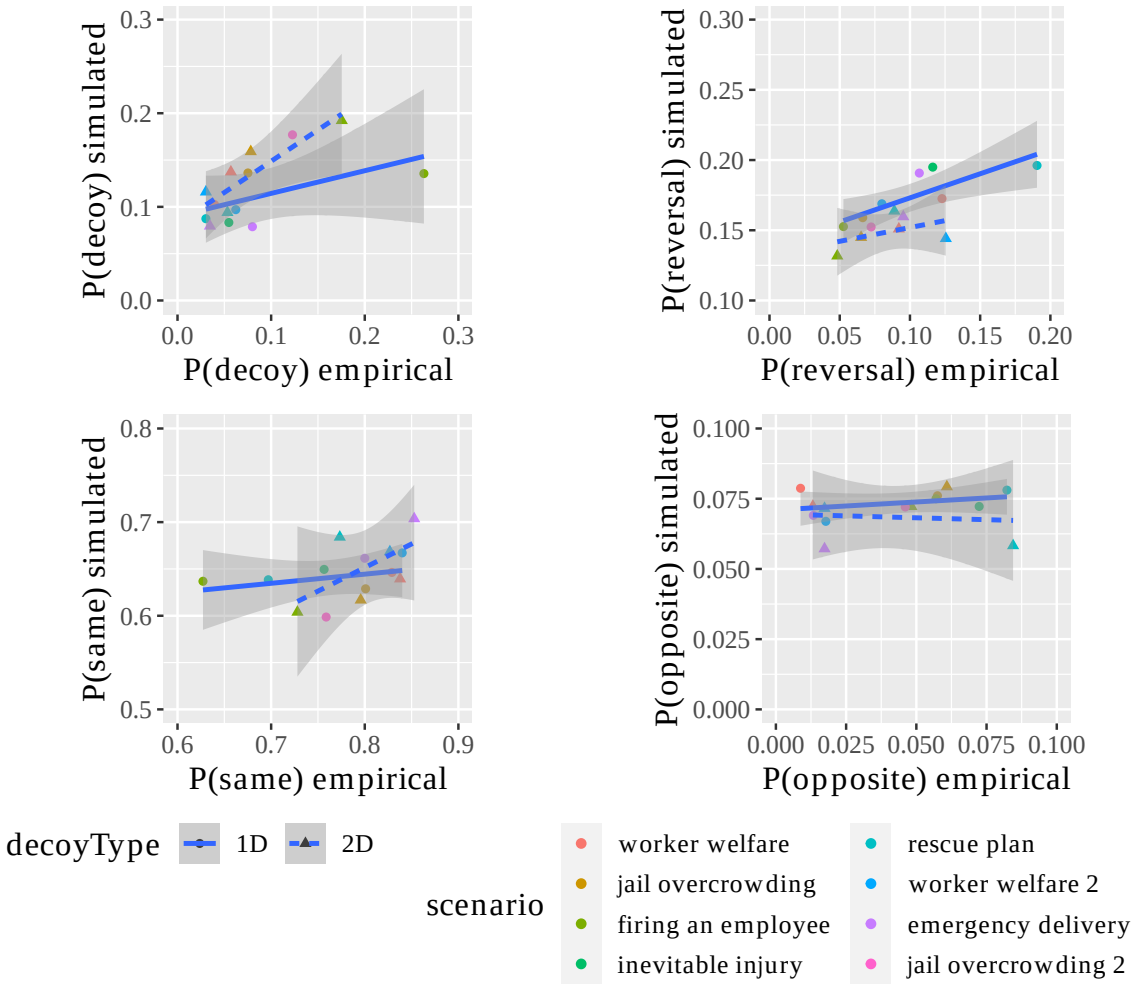
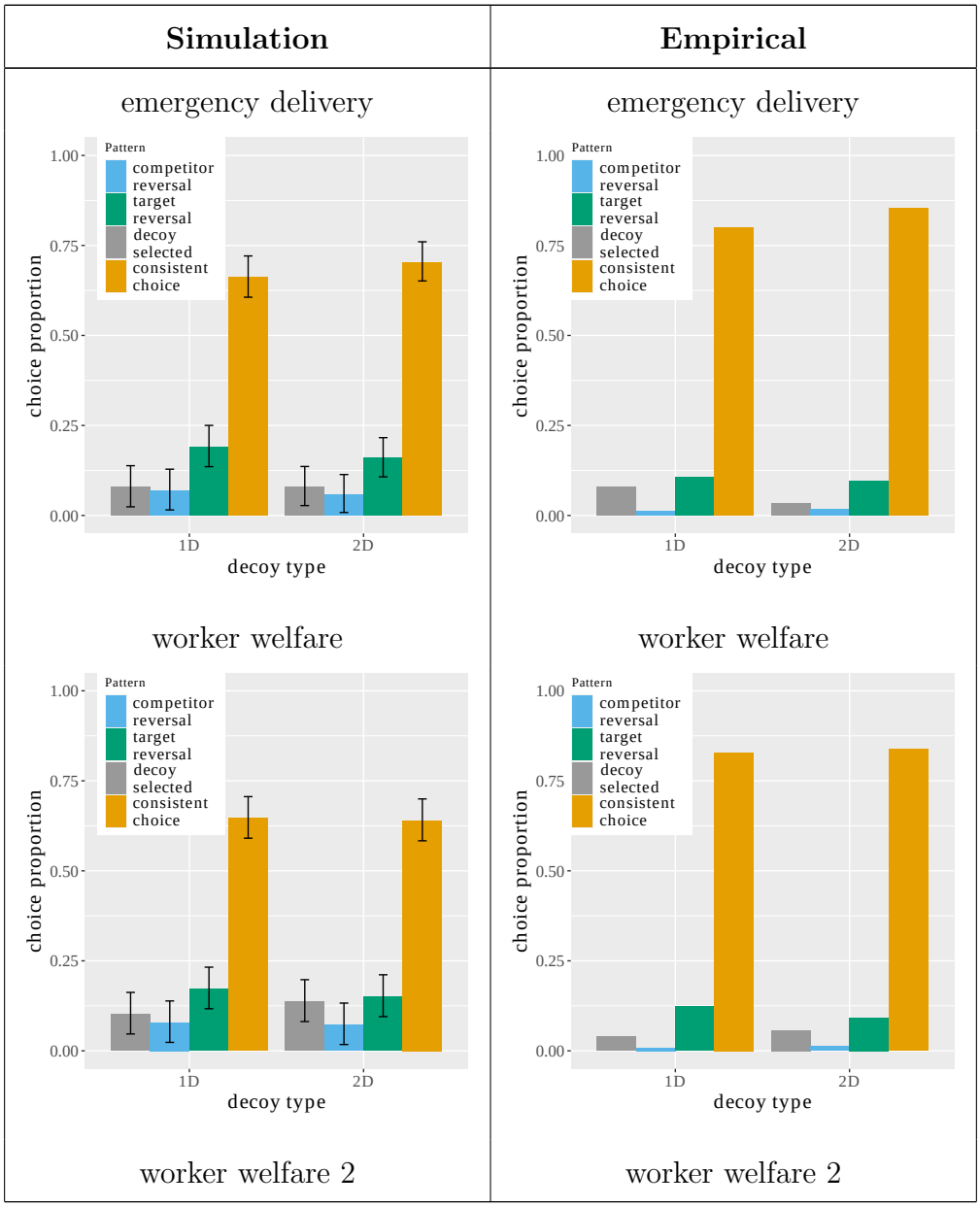


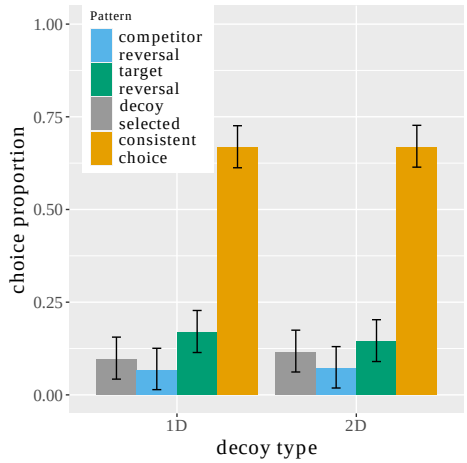
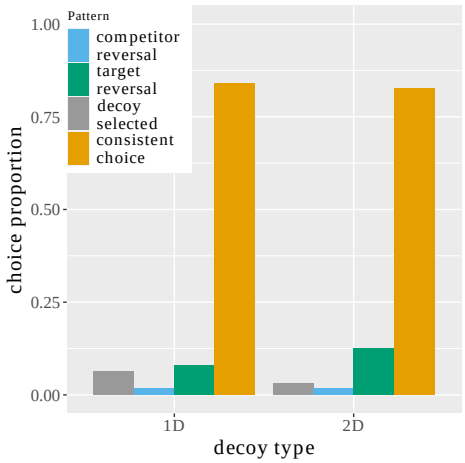
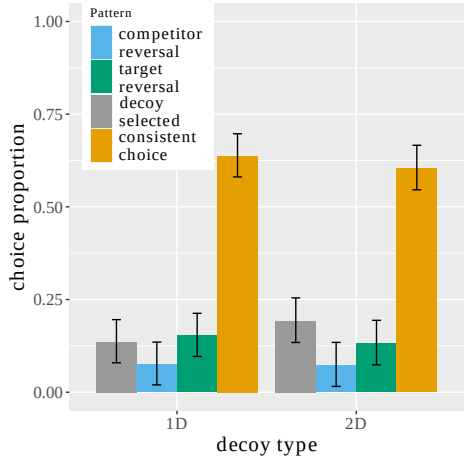
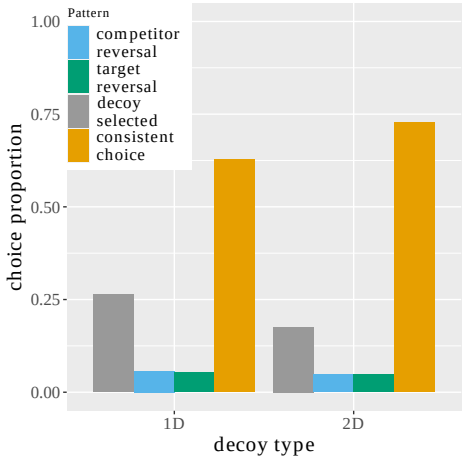
Figure B.2: Simulated and Experiment 3's empirical decoy selection rates (upper left), choice reversal rates (upper right), same-option selection rates (lower left), and opposite selection rates (lower right) for each scenario and decoy type (all 8 items included).

Simulated response patterns for each item in Experiment 3.



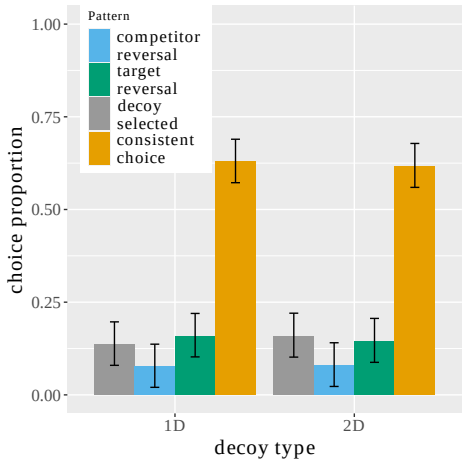
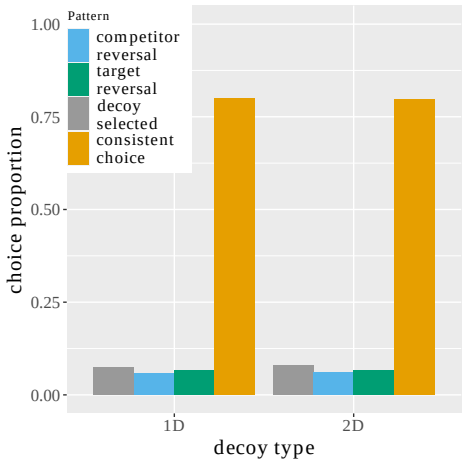
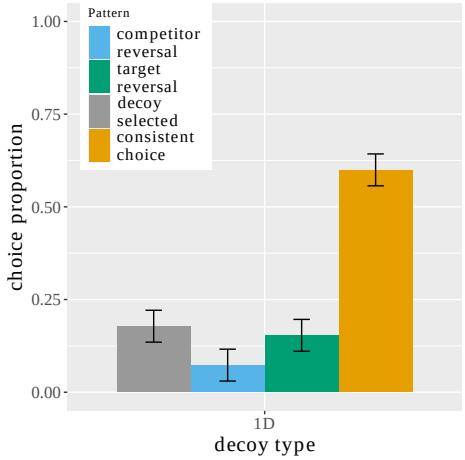
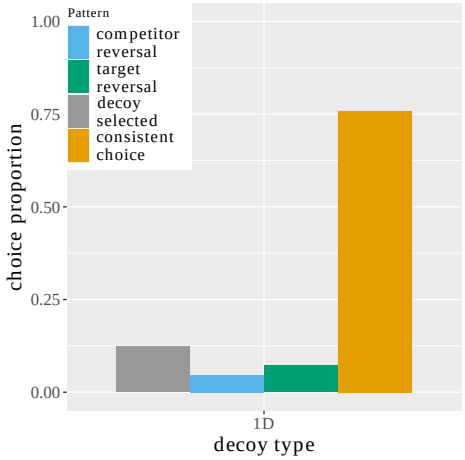
Continued on next page

Continued from previous page

Simulation	Empirical
	
firing an employee	firing an employee
	
jail overcrowding	jail overcrowding

Continued on next page

Continued from previous page

Simulation	Empirical
	
jail overcrowding 2	jail overcrowding 2
	
inevitable injury	inevitable injury

Continued on next page

Continued from previous page

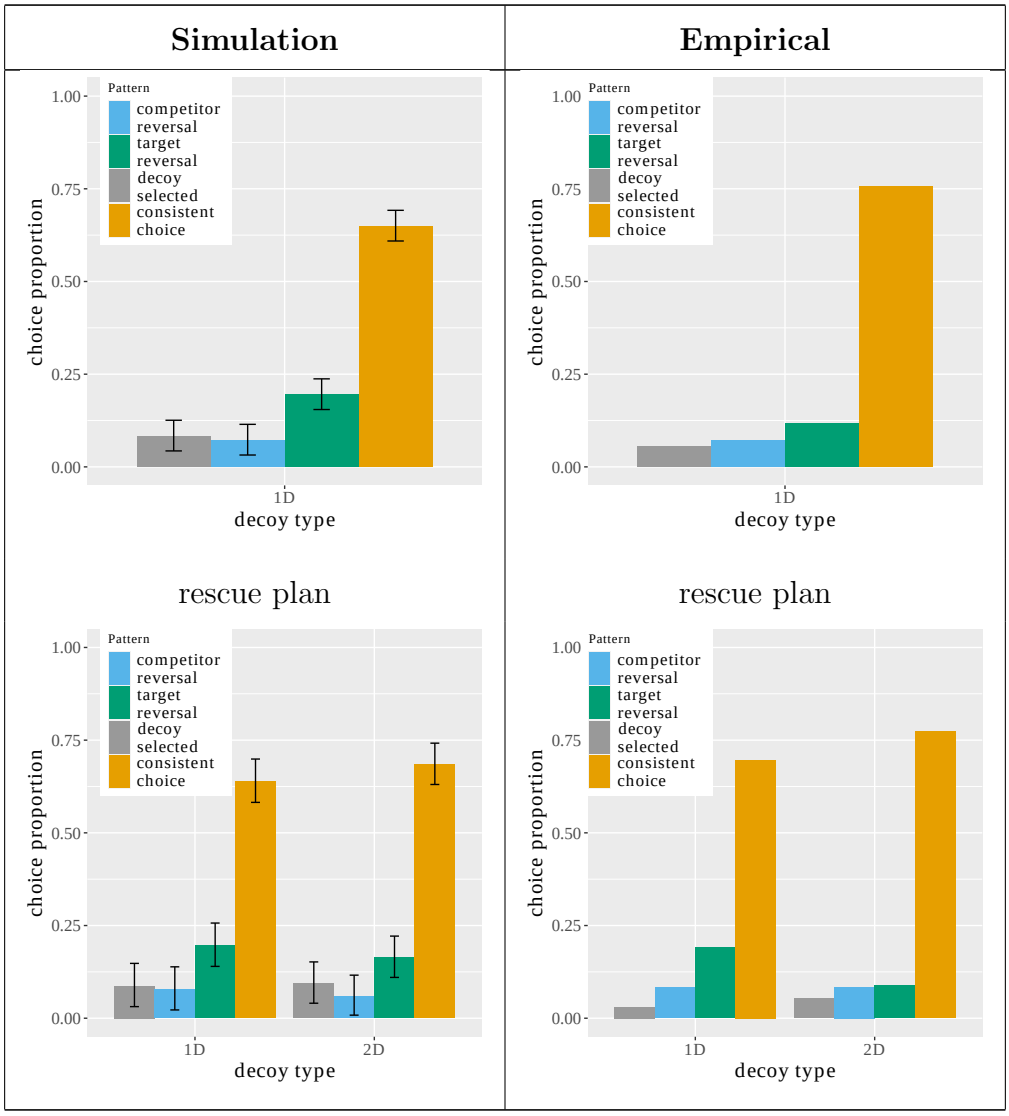
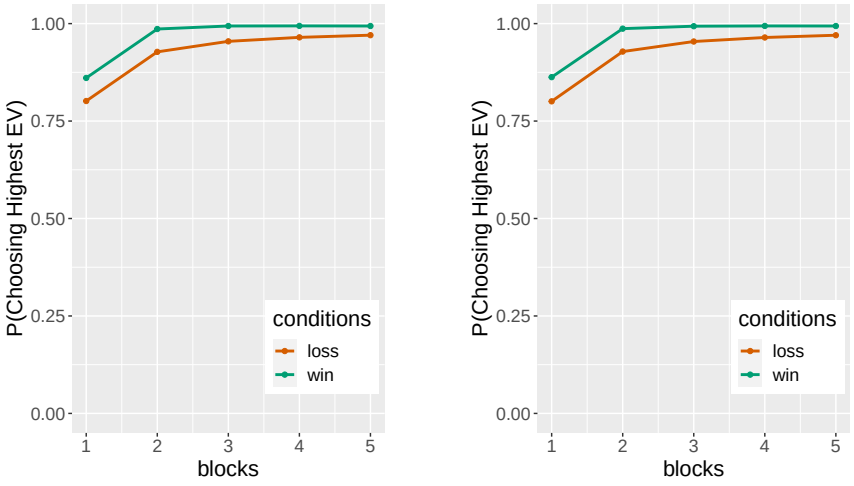


Table B.1: Simulated and empirical data for each dilemma in Experiment 3.

APPENDIX C

Supplemental Materials for the Individual Differences in the Value Learning Task

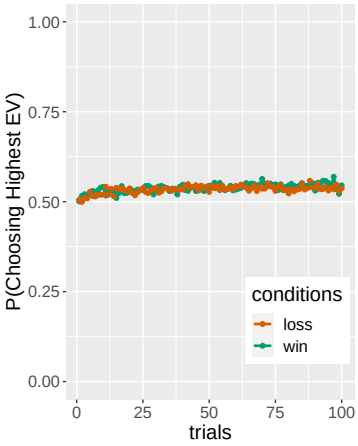


(a) Model, *Nearly Equal Learners* (N = 95; optimal parameters)

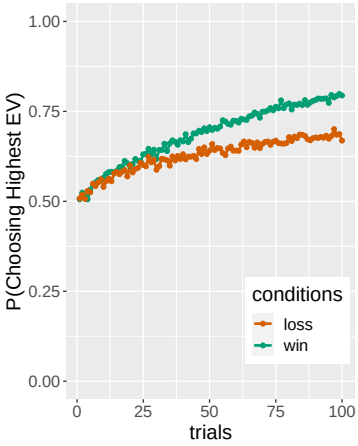
(b) Model, *Unequal Learners* (N = 96; optimal parameters)

Figure C.1: Model simulations (200 runs for each participant) of the VLT with the exact experiences of human participants — using optimal parameters.

(a) Model, all *Poor Performers* (N=48)



(b) Model, *Poor Performers*, "neither" group (N=23)



(c) Model, *Poor Performers*, "win-only" group (N=17)

(d) Model, *Poor Performers*, "loss-only" group (N=8)

Figure C.2: Model simulation for *Poor Performers* and its sub-groups. (a). Model simulation for *Poor Performers* (each=200 simulations). (b). Model simulation for *Poor Performers* who learned neither stimuli. The correct selection rates stay around 50% for both stimuli. (c) & (d). Model simulations for *Poor Performers* who learned only win stimuli and those who learned only loss stimuli. The simulations do not reflect when loss stimuli are learned better win stimuli.

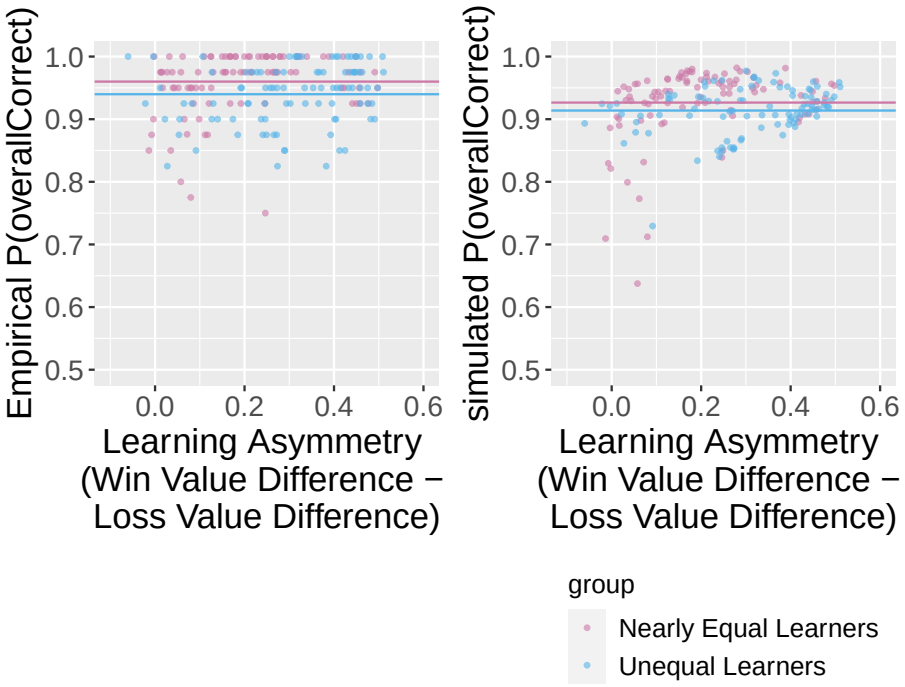


Figure C.3: Learning asymmetry and overall performances by human subjects (left; N = 191) and model simulation with best-fit parameters (right).
Vertical lines indicate the means of overall performance by each subject group. Learning asymmetry is given by the *differences between the value differences of win stimuli and the value differences of loss stimuli*: larger absolute differences indicates larger asymmetry between wins and losses. Overall performance is the mean P(correct) of wins and losses in the last block. Our model predictions are consistent with empirical data: *Nearly Equal Learners* generally have lower learning asymmetry and higher overall performance.

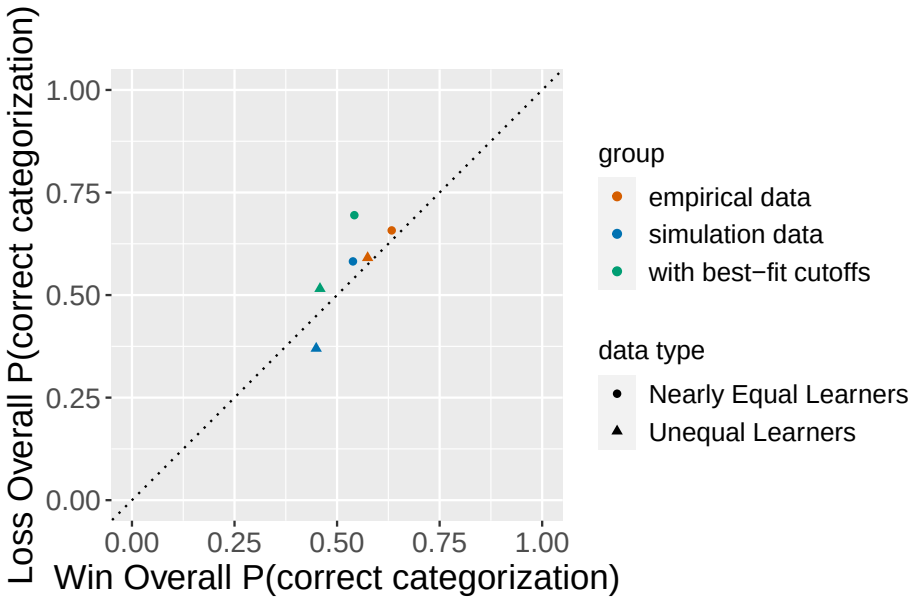


Figure C.4: Overall correct categorization for win- and loss-stimuli by *Nearly Equal Learners* and *Unequal Learners* in empirical data, simulated data, and simulations with best-fit cutoffs.
Our simulated results show the observed win-loss interaction in the human data, including the surprising finding that accuracy in categorizing outcomes of loss-stimuli is slightly better than win-stimuli. This finding does not appear in results simulated with sampled cutoffs, but it appears in results simulated with best-fit cutoffs.

BIBLIOGRAPHY

Aberg, K., Müller, J., & Schwartz, S. (2017). Trial-by-trial modulation of associative memory formation by reward prediction error and reward anticipation as revealed by a biologically plausible computational model. *Frontiers in Human Neuroscience*, *11*. doi: 10.3389/fnhum.2017.00056

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

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. doi: 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. doi: 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. doi: 10.1007/978-94-011-2192-7_2

Baron, J., & Ritov, I. (2004, 07). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*, 74-85. doi: 10.1016/j.obhdp.2004.03.003

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.

Brosch, T., & Sander, D. (2013). Neurocognitive mechanisms underlying value-based decision-making: from core values to economic value. *Frontiers in human neuroscience*, *7*, 398.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018, May). A resource-rational analysis of human planning. In *Proceedings of the 40th annual conference of the cognitive science society.* (Frederick Callaway and Falk Lieder contributed equally to this publication.) doi: 10.13140/RG.2.2.15636.40326

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. doi: 10.1073/pnas.1408988111

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision making, affect, and learning: Attention and performance xxiii.* Oxford University Press. doi: 10.1093/acprof:oso/9780199600434.003.0001

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–1711.

Della, L. C., & Chelazzi, L. (2009). Learning to attend and to ignore is a matter of gains and losses. *Psychological science*, *20*(6), 778-84. doi: 10.1111/j.1467-9280.2009.02360.x

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, *21*(5), 575–597.

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

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annual review of psychology*, *68*, 101–128.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science*, *2*(3), 528-554. doi: 10.1111/j.1756-8765.2010.01094.x

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox*. Cambridge: The MIT Press.

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)

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, *21*(5), 493–518.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517–523.

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*. doi: 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. (1983a). Market boundaries and product choice: Illustrating attraction and substitution effects. *Journal of consumer research*, *10*(1), 31–44.

Huber, J., & Puto, C. (1983b, 06). Market Boundaries and Product Choice: Illustrating Attraction and Substitution Effects. *Journal of Consumer Research*, *10*(1), 31-44. Retrieved from https://doi.org/10.1086/208943 doi: 10.1086/208943

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. doi: 10.1126/science.1091721

Kahneman, D. (2003, December). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449-1475. Retrieved from https://www.aeaweb.org/articles?id=10.1257/000282803322655392 doi: 10.1257/000282803322655392

Kahneman, D., & Frederick, S. (2002, 01). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*, 49–81. doi: 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. doi: 10.1023/A:1007710408413

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kant, I. (1998). *Kant: Religion within the boundaries of mere reason: And other writings*. Cambridge University Press.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., & Rahwan, I. (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. doi: https://doi.org/10.1016/S0749-5978(02)00518-6

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.

Lewis, R., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*. doi: 10.1111/tops.12086

Lieder, F., & Griffiths, T. L. (2018). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Psychological Review*, *85*(4), 249–277.

Liew, S. X., Howe, P. D., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic bulletin & review*, *23*(5), 1639–1646.

Lin, Z., Cabrera-Haro, L. E., & Reuter-Lorenz, P. A. (2020). Asymmetrical learning and memory for acquired gain versus loss associations. *Cognition*, *202*, 104318. doi: https://doi.org/10.1016/j.cognition.2020.104318

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, 05). Influence of uncertainty on framed decision-making with moral dilemma. *PLoS One*, *13*(5).

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, *431*(7010), 760–767.

Nadurak, V. (2018). Two types of heuristics in moral decision making. *Filosifija sociologija.*, *29*(3), 141-149.

Nadurak, V. (2020, 01). Why moral heuristics can lead to mistaken moral judgments. *Kriterion (Austria)*, *34*, 99-113.

Nagel, T. (2012). *Mortal questions.* (Vol. Canto edition). 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. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1057740895704247` doi: https://doi.org/10.1207/s15327663jcp0401_04

Painter, D. R., Kritikos, A., & Raymond, J. E. (2014). Value learning modulates goal-directed actions. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1166-1175. (PMID: 24224537) doi: 10.1080/17470218.2013.848913

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. (2012, 07). Testing the effect of time pressure on asymmetric dominance and compromise decoys in choice. *Judgment and decision making*, *7*, 513-523.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012, Sep 20). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-30.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature reviews neuroscience*, *9*(7), 545–556.

Raymond, J. E., & O'Brien, J. L. (2009). Selective visual attention and motivation: The consequences of value learning in an attentional blink task. *Psychological Science*, *20*(8), 981 - 988.

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. doi: 10.1111/bioe.12707

Rothkirch, M., Tonn, J., Köhler, S. J., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain: A Journal of Neurology*, *140*(4), 1147 - 1157.

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* (p. 31-40). London: Earthscan.

Russell, S., & Subramanian, D. (1995, 04). Provably bounded optimal agents. *J. Artif. Intell. Res. (JAIR)*, *2*, 575-609. doi: 10.1613/jair.133

Savage, L. J. (1972). *The foundations of statistics.* Courier Corporation.

Simon, H. A. (1955, 02). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99-118. doi: 10.2307/1884852

Simon, H. A. (1978). Rationality as process and as product of thought. *The American economic review*, *68*(2), 1–16.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research*, *16*(2), 158–174.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82.

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.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175-200. Retrieved from `http://www.tqmp.org/RegularArticles/vol12-3/p175/p175.pdf` doi: 10.20982/tqmp.12.3.p175

Stan Development Team. (2017). Rstan: the r interface to stan. r package version 2.16.2. [Computer software manual].

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. doi: 10.1017/S0140525X05000099

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Todd, P., Czerlinski, J., Davis, J., Gigerenzer, G., Goldstein, D., Goodie, A., ... Miller, G. (1999). *Simple heuristics that make us smart*.

Trueblood, J. S. (2012). Multialternative context effects obtained using an inference task. *Psychonomic bulletin & review*, *19*(5), 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. doi: 10.1177/0956797612464241

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological review*, *79*(4), 281.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

van Baar Jeroen, M., Chang, L. J., & Sanfey, A. G. (2019, 12). The computational and neural substrates of moral strategies in social decision-making. *Nature Com-*

munications, *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). *Moral luck: Philosophical papers 1973-1980*. Cambridge University Press.

Wollschlaeger, L. M., & Diederich, A. (2020). Similarity, attraction, and compromise effects: Original findings, recent empirical observations, and computational cognitive process models. *The American Journal of Psychology*, *133*(1), 1–30.

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. doi: 10.1111/tops.12382