Models of Morality

Value Integration and Inference

Alexandra Surdina

A thesis submitted to the University of Warwick in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Psychology

Department of Psychology, University of Warwick

March 2023

Contents

List of Figures

List of Tables

Glossary of Terms

agent A person or entity that acts or has the capacity to act ("*Agent*", n.d.).

deontology A theory in normative ethics that considers an action is right or wrong to be determined by a set of rules, duties or obligations.

EU Expected utility theory. The paradigm of choosing actions based on their *utility*, which is understood as the weighted expected values of their outcomes—defined in Section 1.2.1.

GARP Generalised Axiom of Revealed Preference—introduced in Definition 3.5.

LLM Large language model. Language models are computer programs that capture patterns in natural language data by predicting probabilities of words that follow sequences of words. Large language models use deep learning and neural networks and are trained on very large text data sets.

MFQ-30 Moral Foundations Questionnaire, a set of questions used to determine how an individual scores on each MFT foundation.

MFT Moral Foundations Theory, a framework in social psychology. MFT describes moral judgments in terms of a set of *foundations*, that is, values shared across different cultures to a varying degree. This base of morality includes in its originally proposed form includes the foundations of harm/care, fairness/reciprocity, loyalty/ingroup, authority/respect, and purity/sanctity.

Type Moral foundations can be grouped into two types: The first type, known as the *individualising* type, that consists of the foundations of harm and fairness; and

the second, usually called the *binding* type, which includes loyalty, authority, and purity. In this thesis, individualising morality is sometimes denoted as 'type 1' and binding morality as 'type 2'. It is worth noting that this notion of type applies to foundations/values (e.g. 'Harm is a *type 1*-foundation'), not people ('someone is a *type 1*-person').

utilitarianism A view in normative ethics centered around the basis that whether an action is right or wrong depends on the utility of its outcome. The concept of *utility* is usually understood in terms of happiness or well-being of individuals, or as the opposite of pain and suffering.

Acknowledgements

First and foremost, I would like thank my main supervisor, Adam Sanborn, for giving me the freedom and assistance to pursue my ideas in depth, and for kindness and support when things did not go as planned. This work would not exist without you.

Thank you to my second supervisor, Jim Smith, for warm advice, afternoons filled with tea and ideas, and ceaseless encouragement.

I would like to express my gratitude to the Leverhulme Trust for supporting this research as part of the interdisciplinary *Bridges Programme* at Warwick, and to Thomas Hills, without whom *Bridges* would not have existed.

My sincere gratitude goes to the Psychology department at Warwick, including all of its support staff. Thank you all for inspiring me, disagreeing with me, giving me chances to learn by teaching others, and for helping me navigate research under what we would probably call an 'adverse circumstances' condition. Special thanks go to Anu Realo, Derrick Watson, Elliot Ludvig, and Tom Freeman; and to David Smyth over at ITS for IT support. Heartfelt thanks also to my fellow PhD students, particularly my former office mates, for sharing copious amounts of commiseration, tea, and triumph, all of them being things best consumed together; with special thanks to Alina Gutoreva, Jess Whittlestone, Jianquiao Zhu, Mengran Wang, Toria Collard, and Sarah Olin, all of whom helped me stay as sane as possible.

I have people to thank for inspiring me before this journey even started. I would like to thank Dave Lagnado and Neil Bramley for supporting my very first steps in self-directed academic research at UCL, and Hannes Leitgeb for sparking an interest in applying

mathematical methods to human reasoning. I am also grateful to Klaas-Enno Stephan and Karl Friston for a summer school that tipped the scales on my decision to study people rather than abstract algebra, to Linus Mattauch for introducing me to moral psychology more than a decade back, and to Andreas Stuhlmüller for sparking my interest in cognitive science and the study of rationality all these years ago.

I have fond memories of party discussions during which we turned windows into ad-hoc whiteboards, at all universities I got to attend. Thank you to the people who go to this kind of party.

Many thanks to Timm Schorsch-Trautwein and all members of the HCD team at Deutsche Bahn for having my back at work during the write-up phase of this dissertation.

I am deeply thankful to the examiners, Anne Hsu and Gordon Brown, for engaging with my ideas and challenging them. Their input was essential in shaping the final version of this work.

Thank you to all the people who provided feedback on parts of this thesis, including Dawn Drescher, Rike Müller-Werkmeister, Tosca Lechner, Georg Sauerwein, and Ute Stumpf. And thank you to David Lorch for his unwavering support of various kinds, including the intellectual, moral, feline, logistic, and culinary.

Finally, I would like to thank my parents for always supporting me in the pursuit of whatever I was curious about at any given time, setting the tone for the future.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data collection and analysis) was carried out by the author.

Inclusion of Published Work

Chapter 2 of this thesis has been published by the author:

> Alexandra Surdina and Adam Sanborn. Temporal variability in moral value judgement. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci)*, 2017.

A paper based on Chapter 3 is in preparation for submission:

> Alexandra Surdina and Adam Sanborn. Max or mix? Cause preference integration in charitable giving. Manuscript in preparation, 2023.

Abstract

Our sense of what is right and wrong plays a main role in all important decisions we make. In the context of human societies, shared sets of such values and norms encourage cooperative behaviour and help groups thrive. This thesis seeks to understand the dynamics and interplay of individual human moral values, looking at moral decision-making through the lens of mathematical modelling.

Using data collected in a first experiment which captures multiple sequential snapshots of people's values with measurements based on Moral Foundations Theory, a set of models captures different possible patterns of temporal consistency in people's values. A Bayesian model comparison shows that people's values over time resemble samples from two distinct, stochastic processes. For a subsequent experiment, a small set of charities is constructed to be as diverse as possible with respect to their agreement with these two types of moral values. Data is gathered using a donation allocation task among pairs of charities from this set, and shows that while some people tend to allocate their donation fully to their top choice, presumably maximising some combined version of their underlying values, others opt to allocate nonzero amounts to both options, thereby choosing a mix of the different available value combinations. In a follow-up study, a fixed cost associated with donation splitting is introduced for some participants, allowing a comparison of the extent to which it discourages donation splitting. A selection of models of each individual's decisions shows that while people substantially differ in their preference to split their choice, that tendency persists even when doing so comes at a cost. Moreover, their underlying moral values influence not only the allocated total amount, but also the choice pattern they exhibit: The noisy values which encompass our beliefs of what is right and wrong have an effect on the way that these values are combined when a decision is made.

Linking properties of human moral decisions to artificial intelligence research on value alignment, the thesis concludes with a proposed research agenda, introducing the design of an interactive game framework that can be used to study intention attribution in human and artificial agents.

Introduction

> *To bring about the rule of righteousness in the land, to destroy the wicked and the evil-doers; so that the strong should not harm the weak.*
>
> — The Code of Hammurabi

In 1901, a French archeological expedition came across an ancient artefact ingrained with mysterious symbols. These writings, carved in stone, are not an example of ancient poetry or tales of forgotten magic—instead, they tell us what to do.

The object is known to contain the Code of Hammurabi, a Mesopotamian king, and is now believed to stem from the 18th century B.C. It is a legal text with normative rules on a range of topics concerning society, such as adultery, commerce, or crime. Among the laws, you can find *'If a man destroy the eye of another man, they shall destroy his eye'* (§196); *'If one break a man's bone, they shall break his bone'* (§197); and *'If a man strike another man in a quarrel and wound him, he shall swear: "I struck him without intent," and he shall be responsible for the physician'* (§207).[1]

The Code is an old answer to specific versions of a probably even older question: How ought people to act? How can we tell right from wrong?

This document's age and content serve as proof that morality was deemed a topic of relevance thousands of years ago. Today, we differentiate between different disciplines when dealing with morality, each equipped with its own perspective; the fields of law, philosophy, and psychology all use different sets of methods to achieve different goals,

[1] Excerpted from the translation by Harper (1904).

unlike ancient scholars of the past whose work often does not neatly fit into the categories of the present. Although the Code of Hammurabi was likely meant to serve as a legal document that formalises conventions, along with the consequences violating them, it also provides information on what the authors considered to be relevant for moral judgement. For example, from the rules above we can infer that the agreed-upon opinion was that the intention of an agent should be considered when judging an action's outcome: If someone was injured, it's 'eye for an eye', unless it was done unintentionally, in which case the man is only responsible for the other man's doctor's fees.

This thesis studies the psychology of morality through the lens of statistical modelling, with the general aim of improving our understanding of how moral decisions are made. There will also be very little poetry or magic in it; but unlike the Code, it will not tell you how to act, either.

1.1 Thesis overview

The dissertation is composed of five chapters.

Chapter 1 provides the context for subsequent research: First, an overview of theories of rational decision-making is presented, linking it to the study of morality. Then, a short overview over ethical theories is given, followed by a brief history of moral psychology.

Chapter 2 looks into the dynamics of moral values and choices. One potential explanation for seemingly inconsistent behaviour are values that assume different states which change over time, similar to emotional states. Another is that morality is multidimensional and while the underlying values remain constant, decision outcomes vary, potentially because the way values are integrated into making a choice does not. How stable are values?

Building upon the results of the previous chapter, **Chapter 3** examines the way we

combine different, occasionally conflicting values into specific decision outcomes. Do we pay attention to individual attributes, or do we attempt to integrate our weighted preferences?

Chapter 4 focuses on the problem of aligning artificial intelligence to our moral values. It first looks at the implications of previous results; and, second, proposes research directions that address the same process as the chapters above from another perspective: How do we, by observing the decision outcomes of humans and artificial agents, decide what the people's underlying values must be?

The dissertation is concluded by **Chapter 5** which summarises contributions and limitations of the research in previous chapters.

Quotes are included at the beginning of some chapters and sections to set the tone, often in pairs. Their inclusion is not a statement of endorsement for their claims.

1.2 Rational decisions

> *Happiness is possible only to a rational man, the man who desires nothing but rational goals, seeks nothing but rational values and finds his joy in nothing but rational actions.*
>
> — Ayn Rand, Atlas Shrugged

> *Are you so unobservant as not to have found out that sanity and happiness are an impossible combination?*
>
> — Mark Twain, The Mysterious Stranger

Making decisions is hard. It is nearly impossible to be completely informed about every available choice option, whether when choosing which refrigerator to purchase, which

college subject to study, or whether or not to adopt a kitten from the local shelter, or invest in a company by buying its stock. You will not be able to gather and process full information about all available refrigerator models because it would overwhelm you: Each manufacturer offers a variety of exhaustingly similar products with only slightly different features, with fact sheets that list hundreds of device parameters, far exceeding the ten or so items people can keep in mind simultaneously (Cowan, 2010; G. A. Miller, 1956). In other cases, information might not be readily available: In the case of the shelter cat, you might not know their specific age or anything about any past trauma, while for-profit companies do not generally make all internal plans on upcoming changes available openly. Or the information might be available in principle, but not within the amount of time that you have: Researching potential college courses will yield many subjective reports with conflicting information and recommendations, and digging through them might take longer than until the respective applications deadlines. And when in need to purchase a device to replace a recently broken refrigerator, a customer may will not be willing to invest months of research to, say, interview previous owners of a particular model.

Nonetheless, imagine you were able to gather full information, and keep it all in mind. Even knowing everything there is to know about the past and present of a kitchen appliance, pet, field of study, or publicly traded company, it is generally not possible to predict the outcome of choosing each option with certainty. A feline companion might behave differently in your home than it did in any other home it has been in. A company might switch to a new majority owner and change course when it comes to their main products and strategy; and even if it nothing changes for the company itself, its stock will be subject to random fluctuations inherent to the stock market. A field of study that was particularly hyped up may lead to an overabundance of graduates, ruining future

candidates' chances on the job market.[2] Or perhaps you end up making a decision you regret because your very own set of preferences is contradictory. You might have wished for a large fridge that is energy-efficient and inexpensive, despite the fact that the size of a refrigerator correlates positively with both its price and the amount of energy it consumes. You wanted a pet kitten, but you also wanted to own a non-scratched sofa. Aside from information and outcome uncertainty, you might further be influenced by your emotional state, or physical well-being: If your fridge just broke down in the middle of the warmest month of the year, or you went fridge shopping while hungry, you might opt for a larger model than you would otherwise.

A limited amount of available information, limited computational resources and time, uncertain outcomes, conflicting preferences, and the emotional state of the agent are among the many obstacles that it so difficult to make good decisions.

1.2.1 Rationality as utility maximisation

> *The definition of rationality has been much debated, but there is general agreement that rational choices should satisfy some elementary requirements of consistency and choice. (...) People systematically violate the requirements, and we trace these violations to the psychological principles that govern the perception of decision problems and the evaluation of options.*
>
> — Tversky and Kahneman (1981)

One proposed approach to formalising rationality in an uncertain environment is known as *expected utility theory* (EU).[3] In its essence, it proposes to rank choice options by their

[2] Or it might not, as some predicted trends may continue for generations (Haug, 1972).

[3] More precisely, expected utility *theories*, which are all share the expected utility hypothesis. These theories also allow for transformations of outcome probabilites in the formula below: $u(x) = \sum_{i=1}^{n} u(\overline{x}_i)F(p_i)$. Different EU theories mainly differ in which transformations F they allow (P. J. Schoemaker, 1982).

subjective expected value to the agent, where the expected value of a choice obtained by computing an average of the personal utility of the respective outcomes, weighted by their respective probability of occurrence (P. J. Schoemaker, 1982).

$$u(x) = \sum_{i=1}^{n} u(\overline{x}_i)p_i$$

Here, $u(x)$ denotes the agent's subjective utility u of choosing x, with each possible outcome $\overline{x}_i$ happening with probability p_i. A rational agent, in the context of EU, chooses actions that maximise expected utility.

Giving the limitations we are operating under, it should not be entirely surprising that humans are not, in general, rational agents. We are known to prefer avoiding losses to acquiring gains (Tversky & Kahneman, 1992); we selectively seek out evidence that confirms our hypotheses rather than evidence that helps us be more accurately informed (Nickerson, 1998); we tend to prefer smaller, immediate rewards to larger, delayed rewards (Kirby & Herrnstein, 1995). And, to complicate matters further, while we are able to notice it when others make such mistakes, we seem to be less able to recognise our own failures (Pronin, Lin & Ross, 2002). Cognitive biases such as these show that, whether or not applied rationality in the sense of expected utility maximisation is possible in principle, we can be pretty sure that humans do not do it (P. J. H. Schoemaker, 2013).

Expected utility theory, thus, is not a very good descriptive theory of human decision-making under uncertainty (see Kahneman, Slovic and Tversky (1982) and many others); yet, in principle, there is not much that can be said against striving to obtain the very outcome that is best by definition—if what is "best" is well-defined—hence it fares less badly as a normative theory (Briggs, 2019; Fishburn, 1981).

The concept of maximising expected utility requires having a model of the world and an

ability to predict expected consequences of actions, and a way to choose between actions in a way that maximises the expected benefit of each outcome. This amounts to solving an optimisation problem that may be difficult to solve for agents who have a limited working memory and, as a result, cannot compare large numbers of options or option attributes simultaneously.

Decisions under uncertainty have been a somewhat popular subject in psychology—perhaps unsurprisingly, as we exist in a world that can be considered non-deterministic for practical purposes in which a state of complete knowledge can never be obtained by us. We know that solving the specific optimisation problem at hand may be impractical given real-world limitations such as incomplete information or limited computational resources. In practice, as we are bound by such constraints, heuristics and workarounds may be preferable to attempting to solve the problem exactly; an idea known as 'bounded rationality' (Simon, 1990). And yet: While that kind of constraint-dependent rationality may be optimal under the given circumstances, in the absence of these restrictions, if we could indeed pick the action leading to the best outcome in terms of utility, we would probably opt to do so.

1.2.2 Bounded rationality

In an old physics joke, a theoretical physicist is tasked with improving milk production of dairy cattle. At the blackboard, the physicist begins to answer: 'First, let us consider a spherical cow in a vacuum …'.

A perfectly rational, utility-maximising human being might be psychology's equivalent to a biologist's perspective on frictionless space cows. Just like spherical bovine models make simplified assumptions that are implausible from the point of view of biology, the idea of a perfectly rational human utility maximiser ignores restrictions we know

to exist: Humans have limited computational resources and the computational cost of finding optimal solutions can be prohibitively expensive. Also, just as we can observe the shape of cows, we can observe people's behaviour and see that they deviate from utility maximisation: People's preference for immediate over delayed rewards, or our asymmetry between losses and gains, are examples of typical human failures to solve that kind of optimisation problem, known as "cognitive biases".

Even a mere approximation of expected utility maximisation would conceptually not be possible without the ability to evaluate the extent to which an outcome is desirable (MacAskill & Ord, 2020).

What is the utility we would like to maximise in a human-like, bounded fashion? What people deem to be *good* is not defined within the framework of rational choice itself and thus has to depend on evaluations outside of it. To find out more about any utility function humans may have, we need to consider human values.

1.2.3 Are values rational?

> *Science, by itself, cannot supply us with an ethic. It can show us how to achieve a given end, and it may show us that some ends cannot be achieved. But among ends that can be achieved our choice must be decided by other than purely scientific considerations. If a man were to say, "I hate the human race, and I think it would be a good thing if it were exterminated," we could say, "Well, my dear sir, let us begin the process with you." But this is hardly argument, and no amount of science could prove such a man mistaken.*
>
> — B. Russell (1950)

Rationality is not directly applicable to decisions directly concerning utility itself. When

we try to make decisions about what 'best' is, 'act in a way that most likely chooses the best option' is not a very helpful instruction.

Any attempt to increase utility—even a biased, non-optimal one—requires a way of evaluating said utility. In a sense, the range of rationality does not cover the domain of morality; utility maximisation requires a utility function. But is there any extent to which the concept of rationality applies to moral values?

Rationality alone does not determine utilities and might thus not be a sufficient basis for morality.

If the utility function has any effect on itself, it would be that it enforces consistency of values. Otherwise, the corresponding expected utilities would violate the axioms of probability and this would allow a 'Dutch Book Argument'-esque exploit (Vineberg, 2022) of these inconsistencies.

Different societies establish different sets of moral values (Graham et al., 2011), which, along with other arguments (Greene, 2002), points towards morality not being objective. Further, there appear to be different dimensions to morality (Graham et al., 2013). If values are multidimensional, there has to be a cognitive process mapping the set of values to multidimensional utilities of different actions and a subsequent function that forms a preference, such that choosing an action over another is the outcome of this function. And if they are subjective, then unlike rationality, people's utility functions must be specific to each individual.

So if there is no objective utility to adhere to, where do people's utilities have to come from?

1.3 Morality

> *In treating the principles of morals there are two questions to be considered. First, wherein does virtue consist—[…] [a]nd secondly, […] how and by what means does it come to pass, that the mind prefers one tenor of conduct to another; denominates the one right and the other wrong; considers the one as the object of approbation, honour, and reward, and the other of blame, censure, and punishment?*
>
> — Adam Smith, The Theory of Moral Sentiments

This thesis aims to shed some light on how people make moral decisions—in Adam Smith's words, 'how and by what means does it come to pass, that the mind […] denominates the one [conduct] right and the other wrong'. But to provide some context that helps distinguish the 'what' from the 'how', let us briefly take a look at the history of research that addressed Smith's first question.

1.3.1 Ethics and moral philosophy

Moral philosophy has seen different theories that aim at answering the question: How ought people act? This section will give an overview of some of them.

1.3.1.1 Consequentialism, deontology, and virtue ethics

The normative theory of consequentialism considers the moral value of actions to be a result of their outcomes. It argues that whether or not an action is morally right depends solely on the action's consequences, without taking e.g. the agent's intentions into consideration. Utilitarianism, one flavour of consequentialism introduced by Jeremy Bentham, considers the *utility* of the consequences to be the relevant metric.

A contrasting approach is known as deontology. Like consequentialism, it regards morality as something inherent to people's choices, in contrast to other approaches judging morality as a property of the kind of person someone is. Unlike consequentialism, it argues that the moral value of a choice is not determined by its consequences but by its adherence to a moral norm (Alexander & Moore, 2021). From a deontological point of view, some actions are always inherently morally wrong, independently of their consequences. Actions might be *less* or *more* wrong than from a purely consequantialist point of view—in the Code of Hammurabi (Section 1), intentionally hurting someone is presumably deemed more wrong than unintentional injury, and deserves a harsher punishment.

Deontological approaches can take the metaethical stance that morality is objective ('moral realism'). The existence of an objective moral reality implies that moral claims can be objectively true or false (Sayre-McCord, 2021). Such theories may stem from a subtly (or less subtly) imperialist, Western-centric point of view accompanied by the genuine belief that 'other cultures are less developed than us'; a more modest approach might be the (still moral realist) view that none of the cultures we know, including ours, have discovered the True Morality yet. Still, deontology does not necessarily imply moral realism. It is possible that the moral value of an action is independent of its consequences, but the moral value inherent in the action is culturally constructed, or a result of an individual's intuition, or commanded by God. Greene (2008) and Haidt (2001) argue that deontological reasoning is, for the most part, a post-hoc rationalisation of an individual's intuitive moral response.

Virtue ethics is the position that morality does not stem from properties of actions, but from characteristics of the agents themselves: virtues and vices, practical wisdom, moral character. Virtues are positive character traits that ultimately determine how moral a person is. To do what is right is not achieved by choosing actions based on their actual

or intended outcomes, or based on rules, but being the right kind of person—e.g. honest, charitable, or wise (Hursthouse & Pettigrove, 2022).

1.3.1.2 Moral psychology is not ethics, and ethics is not moral psychology

The study of how people make moral decisions is a different approach from normative ethics. While the philosophical study of morality is concerned with ethics itself ("how should we act?"), the psychological perspective constitutes a descriptive take of how humans tackle normative problems ("how do we as humans figure out what we ought to do?").

The field of normative ethics is not particularly interested in which cognitive processes are involved, or people's reasons for ethical disagreements. Instead, it is a prescriptive theory that aims to tell us what it is we ought to do, and which cannot directly be subjected to empirical testing.[4] The relevance of social context is, somewhat characteristically for almost any subfield of philosophy,[5] a matter of debate.[6]

While this thesis is concerned with descriptive models of people's moral decision-making, it is worth noting that some formal approaches to normative ethics also have been undertaken. For example, quantified utilitarianism is trying to enable comparisons between any possible courses of action, e.g. by focusing on one specific utility domain and choosing a 'currency' for it—such as life expectancy measured in disability-adjusted life years (DALY) in medical ethics (Persad, Wertheimer & Emanuel, 2009), or happiness measured by asking people to rate how happy they are on a given scale (Hills & Argyle, 2002)

[4] To any experimental philosophers reading this and preparing to disagree, I would argue that surveys in experimental moral philosophy are used to answer descriptive questions about people's normative positions.

[5] Bourget and Chalmers (2014) interviewed 931 philosophers and found a wide spread of positions on nearly every philosophical question they included.

[6] More generally, the question of whether not an ethical theory can be true is a matter of debate; Bourget and Chalmers (2014) found that while a majority tends towards moral realism—a view that there are objective moral truths—there was only a slight average preference for utilitarianism or deontology.

in ethical hedonism; although whether or not suffering and happiness can be mapped to a one-dimensional scale has been a matter of debate for a while now (Griffin, 1979; MacAskill, 2019b). This is a similar approach to measuring morality in monetary costs; there would probably be an exchange rate. Yet, this does not seem to be the way people make such decisions: If I asked you, the reader, how much money I'd have to pay you to stab the hand of a friend—or, say, kill a small rodent—would a number immediately come to mind? Thinking about the lives of mice in monetary terms seemed to elicit a non-moral evaluation of the choice option, with the decision setting (market) further influencing the amounts people chose (Falk & Szech, 2013), rather than measuring and revealing the underlying moral value of a mouse's life in monetary terms. While this does not mean that the utility calculus framework is not a useful tool when trying to figure out what to do, as one might use it as a basis for a choice and be happy with the result, it does imply that the cognitive processes involved work differently, involving intuition rather than explicit reasoning (Haidt & Joseph, 2004) (arguably, if we were doing the cost calculus implicitly anyway, thinking it through as an aid would merely produce identical results more slowly, making it perhaps a less useful tool).

1.3.2 Game Theory

Another framework that can be used to model social decision-making comes from game theory. Game theory[7] is an approach in economics which models particular social interactions in situations in which the parties involved make decisions and each party's reward depends on the actions the parties take (Osborne et al., 2004). Each possible outcome has a "utility" or "payoff", represented as a real number, for each agent, inducing preferences for some outcomes—namely those with a higher utility—over others. One expansion

[7]Game theory's notion of a 'game' is wider than the colloquial concept—here, a game does not imply entertainment.

on the classical framework of game theory, known as Evolutionary Game Theory, adds an iterative element as well as a survival selection mechanism. This version of Game Theory is able to describe various processes we can observe in other domains from biology to traffic (Hammerstein & Selten, 1994; Zhang, Su, Peng & Yao, 2010), and can explain some aspects of morality (altruism) (Harms & Skyrms, 2008), but relies on a number of assumptions and idealisations to do so (Levy, 2011), partially due to the specific kind of scenarios that are studied, e.g. resource allocation between self and other among the players. Also, evolutionary game theory usually considers a one-dimensional payoff function,[8] and moral values do not seem to neatly map to a one-dimensional reward, as we will continue to see throughout this thesis.

1.3.3 A very short history of moral psychology

1.3.3.1 Moral development

Historically, moral psychology arose within the area of developmental psychology, with early works considering morality to be an acquired ability. In Lawrence Kohlberg's 'stages of moral development' framework, moral 'ability' could be grouped into six stages of development, with an individual's morality maturing with age, although not every person is able to reach the highest stage.

In the first two of Kohlberg's stages, summarised as the *'pre-conventional level'* (Kohlberg, 1976), morality is first understood in terms of punishment or reward (Stage 1), and then as satisfying one's own needs, from which a pragmatic version of reciprocity follows: 'you scratch my back and I'll scratch yours', not loyalty, gratitude, or justice (Stage 2). The

[8] In behavioural experiments, monetary rewards are often used, possibly because this used to be rarely the case in psychology experiments before interdisciplinary research in psychology and economics gained in popularity (V. L. Smith & Walker, 1993).

second, *'conventional level'*, encompasses conforming to society's customs to obtain the approval of others as the determining factor of good behaviour (Stage 3), which then turns into adhering to an explicit version of this with authority and fixed rules determining goodness (Stage 4). In the final, *'post-conventional level'*, what is right is defined by a society-oriented view which takes legal perspectives into account, but recognises that laws can be changed and improved (Stage 5), succeeded by the ultimate, Kantian morality, characterised by adherence to a self-chosen, consistent, universal ethical principle (Stage 6). While the final two stages allow for subjective preferences and opinions outside of moral considerations, morality itself is considered objective. Disagreements on moral issues imply at least one of the disagreeing parties being in the wrong.[9]

Kohlberg's framework successfully collects different flavours of morality evaluations under one umbrella; but at the same time, it lacks explanatory value, since any behaviour can be explained away by an individual not having reached the ultimate stage which completes moral development.

1.3.3.2 Morality and cognition

The empirical study of cognitive processes associated with morality became particularly popular—or, as Bloom (2012) writes, "barely mainstream"—about a decade ago. In a review of literature on morality in experimental psychology, Ellemers, Van Der Toorn, Paunov and Van Leeuwen (2019) find that while the number of social psychology publications almost tripled between 1981 and 2014, the number of those among the set concerning morality rose by a factor of ten—a more pronounced rise.

The central assertion of Moral Foundations Theory (Graham, Haidt & Nosek, 2009) is

[9] It is worth noting that Kohlberg considers morality universal and explains its variety across cultures by asserting that morality in traditional, non-Westernised societies deviates from what the Western world considers moral because the former are less developed (Edwards, 1986).

that there is a universal set of underlying values which guide people's moral decisions. These values, called *moral foundations*, are shared between people and across cultures and constitute a basis for human moral judgement. The established list of moral foundations consists of *Harm/Care*, *Fairness/Reciprocity*, *Loyalty/Ingroup*, *Authority/Respect*, and *Purity/Sanctity*. The extent to which each of foundation matters varies from person to person, as well as across different cultural and political contexts—Western societies tend to put emphasis on Harm and Fairness, valuing *Loyalty/Ingroup* and *Purity/Sanctity* to a lesser degree compared to non-Western cultures (Graham et al., 2011); and somewhat similar differences were observed between liberals and conservatives in the United States (Graham et al., 2009). The list of established moral foundations within MFT is unlikely to be exhaustive—for example, Graham et al. (2013) listed *Liberty/Oppression*, *Efficiency/Waste*, *Ownership/Theft*, and *Honesty/Deception* as promising candidates for future additions.

The model of Dyadic morality (Schein & Gray, 2018) argues that all moral judgements occur through perceived harm. This, however, fails to explain e.g. instances of the dictator game in which people are harming themselves by depriving them of a reward in order to punish unfair behaviour, indicating that at the very least, harm and fairness are not functionally identical. The authors argue that since the MFT score for fairness has a strong positive correlation ($r = .72$) to the score for harm, those two cannot be considered distinct dimensions caused by separate cognitive mechanisms, and write that *'[...] while MFT may be a convenient taxonomy of overlapping values, it certainly does not capture a set of innate moral 'taste buds' or cognitive mechanisms'.* (p.22), because taste buds are receptors activated by separate chemicals in food, using different physiological mechanisms.

However, staying within the taste bud analogy, not all people taste things equally strongly. Some people, known as 'supertasters', are able to detect more subtle differences in the concentration of taste-inducing substances than the average person (Hayes & Keast, 2011).

This ability does not seem to be limited to one taste quality; for example, people who are particularly sensitive to bitterness appear to be more sensitive to sourness (Prescott, Soo, Campbell & Roberts, 2004). If we now asked a set of people across taste abilities to rate the extent of each taste dimension in a food sample, we would expect to find a positive correlation between bitterness and sourness ratings. This should not make us question the distinctness of bitterness and sourness receptors.

It is quite likely that the MFT questionnaire does not present a set of separate, uncorrelated dimensions of moral components (think: a multidimensional Cartesian coordinate grid); but neither does it claim to do so. One explanation for this correlation is the fundamental difficulty to think of examples to use in experiments that only affect one moral dimension, without a better understanding of the actual cognitive mechanisms involved in moral judgement. We are able to isolate chemicals that we know to affect specific kinds of taste receptors, such as glutamate inosinate and guanylate, so we can make a taste sample that only tastes umami; but e.g. in Western cuisine, things that taste umami usually also taste salty, so without the background knowledge in chemistry, we may not be able to tell apart umami and saltiness as unrelated properties.

Another property of morality worth mentioning is that in some aspects, there is an asymmetry when it comes to valence of moral judgements, assigning blame differently than praise (Pizarro, Uhlmann & Salovey, 2003). Botzer, Gu and Weninger (2022) examine text passages posted to the /r/AmITheAsshole subreddit, an online community forum where people present stories that are then morally judged by other users. They observe that the score of a post correlates with whether or not the judgement valence is positive. Since a user's text is posted online first, and the overall judgement occurs second, the authors conclude that if there is a causal connection between moral valence and popularity, *'posts with positive moral valence result in higher scores than posts with negative moral*

valence' (Botzer et al., 2022). But this observation could also be interpreted differently: Perhaps posts that are funny and well written are more likely to receive upvotes (a popularity metric), but also are more likely to elicit a positive moral evaluation from the community, because the way a scenario is phrased has an effect on moral judgement (Shou, Olney, Smithson & Song, 2020).

Ever since Kohlberg used thought experiments in the study of morality—interviewing people about e.g. scenarios with a spouse's severe illness and the morality of theft (Kohlberg & Kramer, 1969)—thought experiments became a tool of choice in the field. Borrowing scenarios from philosophy but using psychology's empirical methods allowed to draw conclusions based on empirical evidence (Doris, Stich, Phillips & Walmsley, 2020). One popular kind of thought experiments are known as *moral dilemmas*: hypothetical decision problems which are designed to elicit internal conflict about which choice is the right one (McConnell, 2022). A particularly well-known moral dilemma which has been used extensively in experimental research on morality is known as the 'Trolley Problem'.

1.3.3.3 The Trolley Problem

> *Michael: Why don't you just tell me the right answer?*
> *Chidi Anagonye: Well, that's what's so great about the trolley problem, is that there is no right answer.*
>
> — The Good Place, Season 2 Episode 5

Imagine a runaway tram speeding down a track; further down the track, there are five workers conducting repairs. The driver can intervene by pulling a lever and thereby diverting the tram onto a different track where only one person is standing, killing that person. If the driver does nothing, the five workers die. Should the driver divert the

trolley? The first known source mentioning the 'tram' dilemma appears to be an essay by Foot (1967), with Thomson (1984) elaborating on it further as well as giving it the name it is now known by. Aside from the main scenario described above, different variations are presented—e.g. instead of the lever, the protagonist has the choice of whether or not to push a fat man onto the tracks, stopping the train, saving the workers but killing the falling man. It has also been used by Greene (2002) to study morality empirically in psychology and neuroscience, where it was subsequently extensively studied, leading to its reputation as the *'Drosophila of ethical inquiry'* (Railton, 2020), and helping it obtain cult status in popular culture.[10]

1.3.3.4 Problems with the Trolley Problem

> *Michael: Well, obviously, the dilemma is clear. How do you kill all six people?*
>
> — The Good Place, Season 2 Episode 5

The trolley problem has been criticised as a completely unrealistic scenario in the past, although this line of criticism significantly quietened once self-driving cars entered the realm of immediate possibility (Awad et al., 2018). Still, some significant drawbacks of using it as the main example to study moral cognition remain, and motivated the decision not to utilise it for experiments within this thesis.

The trolley problem stems from philosophy, and as such, does not depict a realistic scenario even for autonomous vehicles because in the dilemma, all outcomes are known with absolute certainty; there are probabilistic versions but in those, any and all risks are quantified and also known in their entirety. It was intended as a thought experiment to understand and perhaps challenge intuitions, not designed as a representative example

[10] At the time of writing, the 'Trolley Problem Memes' group on Facebook has 8,215 members. The problem also made it into popular TV shows such as 'The Good Place'.

to be used for the empirical study of morality. Not every moral issue can be framed as a version of the trolley problem without sacrificing plausibility. Adding to this, in several versions of the problem, the available outcomes are neither clearly "morally good" nor "morally bad" since the hypothetical situation is constructed to create a dilemma in the first place. Yet, people have to come up with an answer within the experiment. As a result, people's choices may be a result of a suboptimal solution attempt of the not-optimally-solvable problem at hand, rather than a direct reflection of their underlying moral values, and may not be generalisable to non-dilemma situations.

More generally, studying thought experiments is not enough to understand morality because people's actual choices are known to deviate from their thoughts on hypothetical scenarios (Blasi, 1980), including mouse-based (Bostyn, Sevenhant & Roets, 2018) or VR (Patil, Cogoni, Zangrando, Chittaro & Silani, 2014) implementations of the trolley problem. These approaches might capture people's intentions, which is not the same as their actions (Sheeran & Webb, 2016).

One applied area in which these questions in moral psychology have practical relevance due to need for a better understanding of human values is the design and operation of artificial intelligence systems.

1.3.4 AI value alignment

The question of how we can ensure artificial intelligence will be aligned to human values is an open problem (Christian, 2020).

One popular view that is also currently represented in current legal regulation efforts is that AI systems cannot be considered moral agents, and therefore are fundamentally incapable of making moral or immoral decisions, serving merely as an extension that

executes moral decisions made by humans. Therefore, it is not necessary to understand how morality works; we only have to agree upon the right thing, and design AI systems in a way that ensures its implementation while humans remain in control ("human-in-the-loop").

Humanity's lack of agreement on what is right and wrong has been a source of conflict and war throughout history for millenia (Atran & Ginges, 2012). Transferring human-level moral decision-making in groups to systems with superior-to-human abilities in other aspects unchanged (including working memory, speed of computation, and access to knowledge) may cause larger-scale problems. Moreover, groups of people are known to occasionally make moral decisions that individuals in the group would not have made on their own because people tend to conform to ingroup norms (Jiang, Marcowski, Ryazanov & Winkielman, 2023; Pryor, Perfors & Howe, 2019). Imitating human agents may lead to perpetuation of the same undesirable patterns.

Concrete applications of research in moral psychology—including the results of previous chapters of this dissertation—to the problem of AI value alignment will be the focus of the fourth chapter of the dissertation.

Noisy Morals

> *How does one become stronger? By deciding slowly; and by holding firmly to the decision once it is made. Everything else follows of itself.*
>
> — Friedrich Nietzsche, The Will to Power

> *The snake which cannot cast its skin has to die. As well the minds which are prevented from changing their opinions; they cease to be mind.*
>
> — Friedrich Nietzsche, The Dawn of Day

2.1 Introduction

Morality is a vital part of who we are. A person's moral beliefs are tied into their identity (Aquino, Freeman, Reed II, Lim & Felps, 2009; Aquino & Reed II, 2002)—humans believe that if their moral values changed, they would change (Heiphetz, Strohminger & Young, 2016). Are people's intuitions about this correct? Are our moral values consistent over time? This chapter aims to determine whether people's moral values stay constant over time, looking at time intervals with lengths of relevance to everyday decision-making tasks.

Since moral beliefs tend to be associated with a person's sense of identity, we should expect people's underlying moral values to largely endure over short time periods. Yet, there have been many recent explorations of moral inconsistency. These have included manipulations of two kinds—manipulations of response timing, or manipulations by

exposure to new information or decisions. In terms of timing, we now know that time-limited decisions appear to be more altruistic (Rand, Greene & Nowak, 2012) and that choices can be influenced by forcing decisions at a specific point in time (Pärnamets et al., 2015), indicating that the actual decision outcome is time-sensitive. Regarding information or decisions, dishonest behaviour increases future dishonesty (Engelmann & Fehr, 2016; Garrett, Lazzaro, Ariely & Sharot, 2016). At the same time, a morally good action makes a subsequent morally bad action more appealing and vice versa, effects known as moral cleansing and moral licensing (Merritt, Effron & Monin, 2010; Sachdeva, Iliev & Medin, 2009) that stand in direct conflict with dishonest behaviour repeating itself in the future. Moreover, exposure to a moral dilemma leads to belief revision in moral decisions that persists for multiple hours (Horne, Powell & Hummel, 2015).

The fact that changes in external circumstances can influence the outcomes of moral decisions is hardly surprising assuming morality evolved as an adaptive strategy (Machery & Mallon, 2010). Likewise, viewing moral judgment as a decision process, we would expect the effects of changed response timing on general decision-making (McClelland, 1979; Usher & McClelland, 2001) to transfer into the moral domain. But in the absense of such manipulations, are our moral judgments fundamentally noisy? Outside of the moral domain, there is evidence in decision making research that people's decisions vary stochastically even in cases where external conditions remain constant (Mosteller & Nogee, 1951). We are interested in exploring whether there is a corresponding moral variability beyond the actual decision process: are our moral values different from moment to moment, even in the absence of new information or manipulations of response timing?

Moral Foundations Theory (MFT) provides a way to look at this. It is based on a dominant model of morality, the social intuitionist model, according to which moral choices are made primarily intuitively and then justified post hoc (Haidt, 2001). MFT maps out the

moral domain in terms of six fundamental hidden parameters that appear to capture an individual's moral judgment (Graham et al., 2009), enabling us to distinguish between conservative and liberal political profiles on the basis of an agent's foundation weights. This idea that there are foundational categories that guide intuitive moral judgement has the potential to explain people's tendency to disagree on moral issues, and predict future moral judgement based on the individual scores. If we can find a systematic structure in the stochastic changes of different foundation scores beyond merely a layer of noise, this would point towards moral variability, rather than just motor variability or variability in how the response scale is used.

In line with the aforementioned results indicating temporal consistency, moral foundation scores appear stable over longer time periods; Graham et al. (2011) tested participants again after approximately a month and found that their moral foundation scores exhibited test-retest reliability. Yet, effects such as moral licensing and moral cleansing—where the outcome of an individual's moral decision influences subsequent moral decisions, even decisions made by others in their ingroup (Kouchaki, 2011), over the course of single experimental sessions and thus shorter timescales—suggest the possibility of an interaction between moral foundations. Moreover, the list of known moral foundations is likely incomplete—a view shared by moral foundations theorists (Haidt & Joseph, 2011).

Viewing moral decisions as a sampling process from a distribution that represents an agent's moral values, we can use the framework provided by MFT to investigate hidden parameters which predict an individual's moral variability. Conversely, observing within-subject variability over time can help us understand to which extent individual moral variability reflects between-individual variability that has been used to support the existence of MFT (Graham et al., 2011). Are we all sometimes a little bit more conservative and sometimes a little bit more liberal in our moral judgments and values?

In this chapter,[1] we aim to discuss the extent to which randomness plays a role in moral judgment over time by collecting responses to the moral foundations questionnaire delivered repeatedly. We subsequently fit a set of models to the data and compare them. If the variability we observe stems merely from randomness in the decision process, we expect variation in individual responses to be explained by a single noise-generating process. We find evidence for at least two separate stochastic processes associated with different sets of moral foundations, indicating the existence of inherent variability in moral values.

2.2 Method

2.2.1 Participants

The participant pool consisted of 80 psychology undergraduate students (mean age 19 years, 90% female). 14 participants were excluded from the analysis due to wrong responses on the two 'catch' trials, as done by Graham et al. (2011).

2.2.2 Materials

The original moral foundations questionnaire (MFQ30) asks participants to respond using a 1–6 scale; to enhance precision and avoid subjects simply recalling previous answers, the participants in our task had to use a slider bar to indicate their responses instead:

[1] This content of this chapter was published in a paper (Surdina & Sanborn, 2017) presented at CogSci 2017.

Figure 2.1: Slider bar provided for participants to indicate their respective responses.

In addition, our version of the questionnaire contained four further questions (see Table A.1 in the appendix). Those were chosen so as not to correspond in any obvious way to the five foundations measured in the MFQ30, nor to the recent addition of the liberty foundation (Graham et al., 2013; Haidt, 2012). We added these questions because we wanted the same number of presumably neutral trials as the number of foundation-related questions—the MFQ30 includes six questions for each foundation but only two neutral 'catch' items.

2.2.3 Procedure

The questionnaire was presented six times in randomised order, with a word search task before the last two trials. In each trial, one of the two question types was displayed (see Table A.1 in the appendix, left and right side, respectively), along with one of the statements for that question type. Randomisation was implemented so that each statement was shown to the participant exactly once in each block: The set of questions within each block was shuffled, and presented within the block in randomised order, so no regular pattern in the order of foundations would occur. After four blocks, a word search task[2] was shown for 6 minutes to provide a timed break:[3] Participants had to find and mark words in a 18x18 letter square filled with a selection of words and random letters, based on

[2]We removed words such as 'excellent' from the task to reduce the likelihood that word valence in this task would influence future participant responses. This quiz block was followed by three more blocks.

[3]Due to an off-by-one error in our code, the first statement from the block after the word search task was erroneously displayed before the word search task. We excluded this error trial from the analysis.

the WordFind.js library (Scheidel, 2012). With the exception of the timed word search task, participants provided responses at their own pace. The experiment took approximately 20-25 minutes to complete.

2.3 Results

Since participant responses are indicated using slider bars, foundation scores change between the blocks (participants will be unable to recall the exact position of the slider for previous trials). But beyond the expected variation resulting from differences in participant's slider operation accuracy, is there a relationship between these changes in different moral foundation scores?

2.3.1 Means

As found by Graham et al. (2011), we anticipated and found our psychology undergraduate subject pool in the UK to remain largely at the liberal end of the U.S. political spectrum. Welch's t-test shows that the differences between the means for harm and fairness ($p=.16$) and for loyalty and authority ($p=.44$) are not significant. All other pairs of means indeed differ significantly ($p<.001$). In particular, the first two foundation means differ significantly from the last three, with higher subject scores for harm ($M = 72.9$, $SD = 24.6$) and fairness ($M = 70.3$, $SD = 23.6$) and lower scores for loyalty ($M = 51.1$, $SD = 27.1$), authority ($M = 48.8$, $SD = 26.1$) and purity ($M = 42.1$, $SD = 28.8$). The between-subject standard deviation is notably larger than the within-subject standard deviation (see Figure 2.3), supporting the MFT framework for examining between-subject differences.

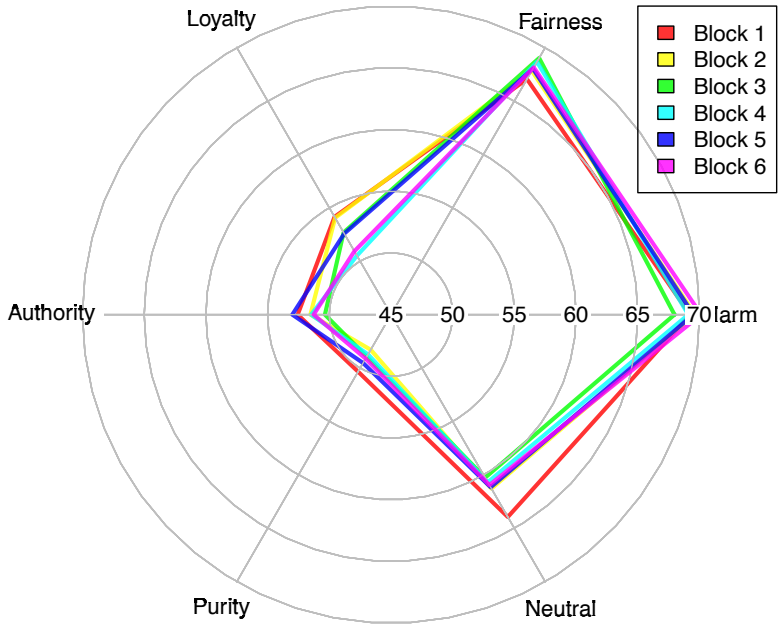


Figure 2.2: Spider plot of means for each foundation and block. Average participant scores were larger for the individualising foundations harm and fairness than for the binding foundations loyalty, authority and purity.

A within-subjects ANOVA[4] showed a main effect of both foundation ($F(5,325) = 72.67$, $p<.001$) and block ($F(5,325)=6.26$, $p<.001$) on average slider bar values, as well as an interaction between foundation and block ($F(25,1625)=1.764$, $p=0.011$). But we are mainly interested in changes in the absence of new information, and Figure 2.5 suggests that the very first block in which the whole questionnaire was new to the participant qualitatively differs from the others. Notably, from Figure 2.2 we can also tell that participant responses,

[4]It should be noted that in this ANOVA, we are treating the block number as a factor variable rather than a numeric variable due to the non-linear relationship between block number and participant response; including the block number as a numeric variable yields qualitatively the same results.

on average, do not appear to be moving closer to the centre of the scale block by block. This suggests that the effect we can see is more likely due to a loss of novelty, rather than a gradual loss of attention.

Excluding the first block from the analysis indeed makes the effects of block ($F(4,260)=6.26$, $p=.47$) and the interaction effect between foundation and block ($F(20,1300) = 1.201$, $p=.24$) in the ANOVA above no longer significant: While moral foundation scores differ between Block 1 and the other blocks, for the later blocks alone, this is no longer true.

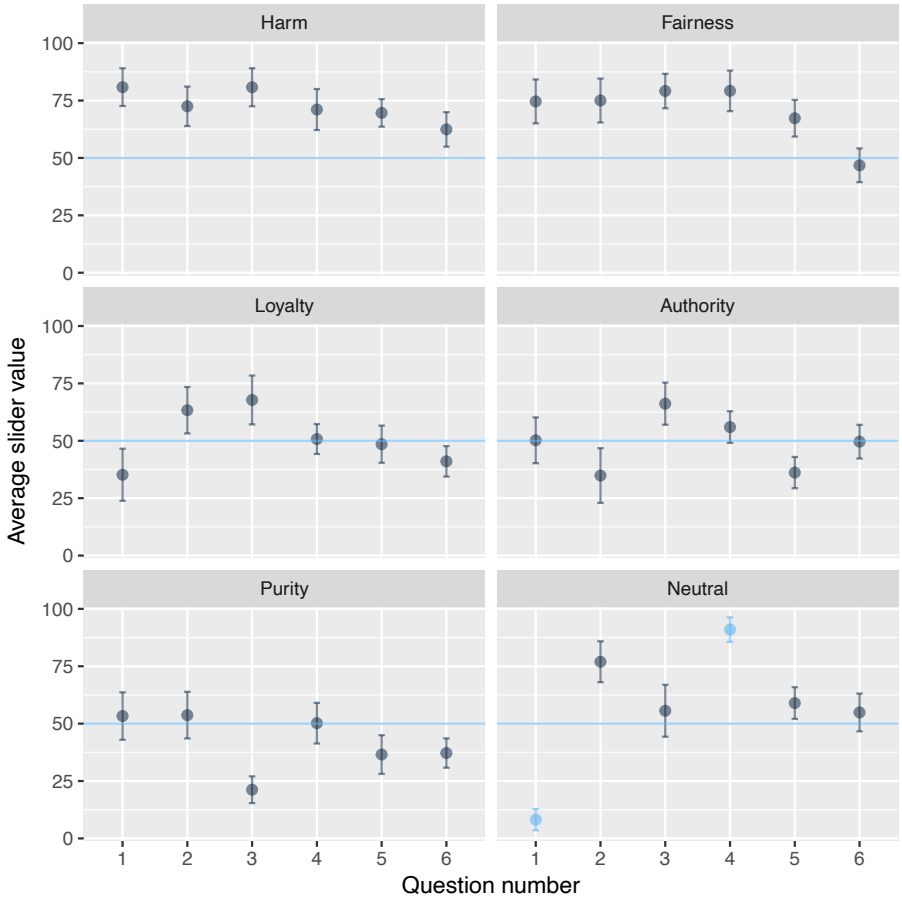


Figure 2.3: Average slider value for each response, and the average of within-subject standard deviations. The catch trials and the baseline level are marked in blue.

2.3.2 Variability over time

We also expected that within-foundation variance, i.e. the variance between participant responses to the sets of questions for each respective foundation, would decrease over time: As time passes, people's certainty which choice they will make will increase as they get more familiar with the questionnaire. Moreover, we thought we might be able

to observe a shift towards more extreme values for each question over time—as people become increasingly familiar with the set of questions they will encounter, there would be less need to for caution about new options which are more or less morally upsetting than the previous maximum or minimum, respectively.

We computed residual slider values by subtracting the within-subject mean for each foundation from the slider values for each trial. The two hypotheses above can be rephrased as: The slider residual variance for each participant and block decreases as a reflection of the increase in certainty; and the average absolute residual value increases over time as a result of the decision drifting towards the extremes.

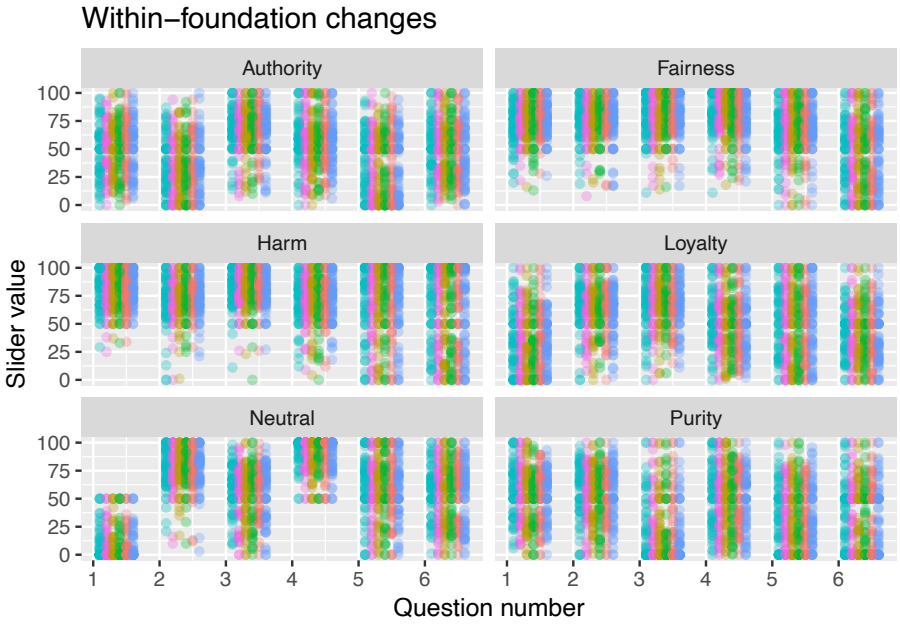


Figure 2.4: Changes over time, by foundation. The colours represent the different blocks. No visual indication for alternating patterns (resembling licensing or cleansing effects) of individual foundation scores.

In fact, we found no significant effect of block number on foundation variance: Again, an

ANOVA only yields significant results for the variance hypothesis (F(5,325)=18.71, p<.001) and the absolute residual hypothesis (F(5,325)=47.4, p<.001) if we are taking the very first block into account—here, a slight decrease after the first block can be spotted (see Figure 2.5). If we are looking at only the other blocks, we do not find any significant change in the variance (F(4,260)=1.90, p=.11), nor in the absolute slider residual (F(4,260)=1.22, p=.30).

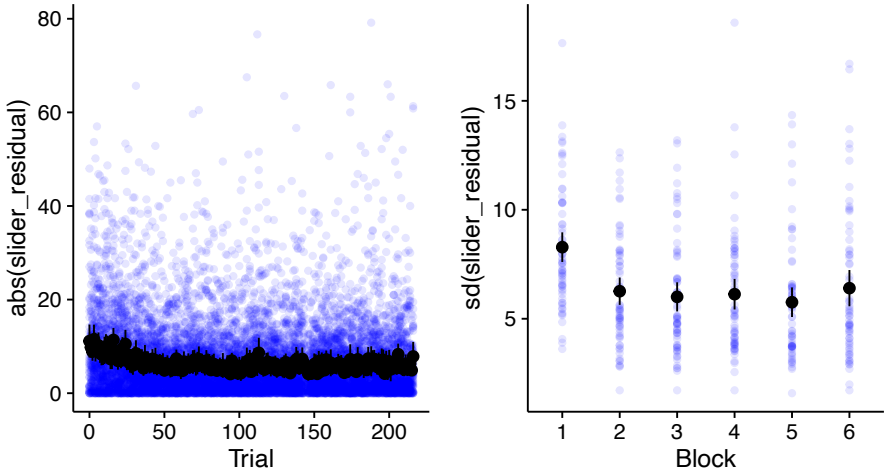


Figure 2.5: Absolute value and within-subject standard deviation of slider residual over time.

2.3.3 Between-foundation variability

One hypothesis is that changes in moral foundations that are opposed with respect to their representation on the political spectrum, such as harm and purity, will balance each other out—that is, they are negatively correlated (Fig.~2.6a). Each person may have a constant morality 'budget', and thus an increase in a moral foundation score will inevitably be accompanied by a decrease in others. This would imply that people's position on the liberal-conservative spectrum might not be fixed. Another hypothesis is that changes in

opposing moral foundations are positively correlated (Figure 2.6b). This would for instance be the case if people's moral profile was indeed fixed, and the sampled moral foundation scores are scaled by a time-dependent factor. Alternatively, changes in different moral foundations may not be correlated at all.

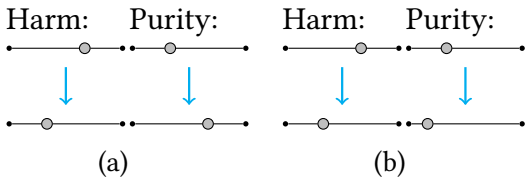


Figure 2.6: Relative changes in foundation scores we would expect to see if changes within different foundations exhibit negative or positive correlation.

We did not find evidence for any of these relationships between participant scores for different foundations. On the contrary, the changes in foundation scores over time were not particularly large.

To test for interactions between changes in different foundation scores, we modelled the data using mixed effects models with a full covariance matrix and a diagonal covariance matrix, respectively. We created dummy coded variables for each foundation. Since we did not detect any notable change in the means after the first block, we now focused on the variability and removed the influence of the means entirely by modelling slider residuals: we calculated the mean slider value for each question for each participant, and subtracted it from the raw slider values. We used residuals for each question rather than for each foundation score because of the differences in responses to the different questions within each foundation (see Figure 2.4). Furthermore, we excluded the first block in which all information had been newly introduced from the analysis. We fitted two models to the data: First, a model including a full covariance matrix and thus allowing for interactions

between the different foundations, and second, a model with a diagonal covariance matrix reflecting the assumption that sampling occurs for each foundation individually.

As a baseline model, we used a model assuming a random slider residual for each participant and block, sampled from the same distribution for each foundation (random noise model). The models for the slider residual y_{ijkl} of Participant i in Block j for a question or statement l relating to Foundation k are:

$$y_{ijkl} = u_{ij} + u_{ijk} + \epsilon_{ijkl}, \tag{M1-M3}$$

with $u_{ij} \sim \mathcal{N}(0, \sigma)$, and

$$u_{ijk} = 0 \tag{M1}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \\ u_{ij3} \\ u_{ij4} \\ u_{ij5} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{bmatrix}\right) \tag{M2}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \\ u_{ij3} \\ u_{ij4} \\ u_{ij5} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & 0 & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 & 0 \\ 0 & 0 & \sigma_{33} & 0 & 0 \\ 0 & 0 & 0 & \sigma_{44} & 0 \\ 0 & 0 & 0 & 0 & \sigma_{55} \end{bmatrix}\right) \tag{M3}$$

The model M2 ($\chi^2(10) = 26.98$, $p = .003$) differs significantly from the baseline model M1. M2, which has a full covariance matrix, shows an interesting pattern of dependencies within the two different foundation types:

Foundation	Harm	Fair	Loya	Auth
Fair	1			
Loya	-0.86	-0.86		
Auth	-0.95	-0.95	0.95	
Puri	-0.7	-0.7	0.64	0.80

Responses for harm and fairness appear to be positively correlated with each other and negatively correlated with responses for the other foundations, and vice versa. This would be less surprising if it was merely capturing a between-participant relationship between foundation scores. Note however that this model describes the slider *residuals* which add up to zero for each foundation and participant—yet, this model implies that participants who drag the slider bar a bit further to the right for harm-related questions *than in the last block* will do a similar thing with the fairness-question slider, but the opposite with sliders on loyalty, authority and purity trials.

Comparing the models M1 to M3 to each other suggests that M2 (BIC $= 71058$) has a higher BIC value than M1 (BIC $= 70960$) and M3 (BIC $= 70993$). However, these results do not suffice to dismiss the idea behind M2 and M3 in favour of M1: Attempting to fit these models using a succession of increasingly Bayesian R packages for generalised linear mixed-effects models (`lme4`, `blme`, `brms`) yielded singular fits every time, possibly due to interdependence among the individual foundation values—thus hinting at a lower dimensionality.

Is there some overlap between which property of morality harm and fairness on the one hand and loyalty, authority and purity on the other hand are measuring? Since the mean foundation scores for harm and fairness, and the scores for loyalty, authority, and purity seem similar to each other (see Figure 2.2), we introduced alternative models that only distinguish between these two groups instead of the individual foundations.

To find out if we could successfully fit a set of models by looking at two dimensions instead of five, we fitted a set of linear mixed effects models to the data. As an alternative to our approach above, we dummy-coded two foundation *types* (the *individualising* foundations harm and fairness, and the *binding* foundations loyalty, authority, and purity (Graham et al., 2009)). Again, we fitted a full covariance model and a diagonal covariance model to the data, adding the two models below to our list of candidate models. They are describing the slider residual y_{ijml} of Participant i in Block j for a question l of Foundation type m:

$$y_{ijml} = u_{ij} + u_{ijm} + \epsilon_{ijml}, \tag{M4-M5}$$

with

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}\right) \tag{M4}$$

$$\begin{pmatrix} u_{ij1} \\ u_{ij2} \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & 0 \\ 0 & \sigma_{22} \end{bmatrix}\right) \tag{M5}$$

We find that out of these, the model M5 differs significantly from the baseline model ($\chi^2(2) = 27.85, p < .001$). Comparing M4 (BIC=70960) and M5 (BIC=70951) to the models above suggests that M5 is preferable to M1 and M4. Thus, it appears that from a model comparison perspective, the main distinction in the moral foundation framework lies in the two different foundation types rather than the individual foundations, and that at this level of description, between-foundation correlations do not play a prominent role.

The two-type model of response residuals implies that people's response noise is not merely one-dimensional. This indicates that more than one noisy process is involved

when an individual produces a response.

2.4 Discussion

We found that people showed moral variability even in the absence of new information or time pressure. This moral variability is distinguishable from response variability because we found two random processes that were associated with different sets of moral foundations. The evidence for MFT is based on an analysis of between-individual responses to the MFQ (Graham et al., 2011), and much of this may actually be due to the within-individual variability that we have found. This within-individual variability may also be what allows timing interventions to have an effect (Pärnamets et al., 2015), and might potentially even allow to influence the outcomes of value-related decisions in cases where the preferences are different, but close (such as election results).

While for our data set a simpler two-type model was preferable to the more complex model including five moral foundations, we hesitate to draw general conclusions about the number of moral foundations due to the small size and relative cultural homogeneity of our subject pool. Yet, our brief glimpse at candidates for additional foundations suggests the possibility of a wider underlying structure of which MFT has captured but a part.

A common criticism of MFT is that the known moral foundations are unlikely to capture moral judgment in its entirety (Suhler & Churchland, 2011). We had expected our added questions to be rated similarly irrelevant to morality as the more conservative moral foundations in our liberal subject pool. Somewhat surprisingly, the responses to our added, 'neutral' foundation appear to be less neutral overall. We chose the four additional statements in the neutral foundation because we suspected that they might turn out to be morally relevant. Figure 2.3 suggests that questions 2 and 5 in particular (see Table A.1

and Figure 2.4) indeed resonate with our participants' values. While the act of lying may arguably be related to the purity scale, it is remarkably more morally relevant than any of the purity questions. This particularly utilitarian view on having children also appears to lie outside of the given scales.

The two-dimensional structure of response noise hints at the existence of two processes, each of which generates its own noise, presenting evidence for two different types of 'moral taste buds' (Section 1.3.3.2): One type, present in all people, sensitive to individual-centric values; the other type, not universally present, responsive to values that bind us to social groups.

This leads to interesting open questions that reach beyond refining and expanding MFT. While we observe a range of scores within these two flavours of morality, we do not yet understand the actual decision process: What are their respective roles in choices we make? How does moral variability translate into decision outcomes? And, more specifically: How are different moral values integrated in a decision between options that are morally relevant for more than one moral foundation, or options that are uncertain? Are moral foundation scores—or moral type scores—combined in a linear fashion as the MFT scoring mechanism (Graham, Haidt & Nosek, 2008) suggests?

This line of investigation will be the focus of the following chapter, in which we take a look at a specific kind of value-based choice: giving out money to do good.

Integration of Moral Values in Charitable Giving

3.1 Introduction

We ended the last chapter having confirmed that there are multiple, independent moral values which influence people's moral cognition—in agreement with one of the main assertions of Moral Foundations Theory—and the finding that they are stochastic in nature, thus bringing up the question how these possibly conflicting values are combined when a decision is made. This question will be the focus of this chapter.

Different moral values do not always suggest the same course of action. In artificial scenarios such as the trolley problem, the respective intuitions culminate in a choice between two distinct outcomes; but in less isolated settings, the space of possible actions is wider.

Many moral decisions are not binary. A high-school student who wants to choose the right career path is not just facing a video game-like choice between two floating clickable banners saying 'become a psychotherapist and save ten people's lives' and 'go into medical research, and kill fifty people but save a thousand with a new drug'. A non-specialist performing first aid doesn't get to choose between exactly the two options of 'do nothing for a 43.7% risk of a wound getting infected' and 'disinfect the wound with a 1% risk of permanently damaging the patient's limb'.

One way a decision between differently evaluated outcomes might be made is by choosing the highest-value option alongside each moral dimension ('max'). Essentially, the agent would choose an attribute to pay attention to, and opt for the least-wrong or most-right action regarding that attribute. For instance, a first-aid practitioner may primarily care about the risk of causing harm, and choose the action that reduces the risk of doing so (not touching any person who is not in immediate danger).

Another approach would be attempting to combine the conflicting values to some degree, making a moral trade-off ('mix'). Guzmán, Barbato, Sznycer and Cosmides (2022) observed trade-off responses combining utilitarian and deontological moral intuitions in dilemmas specifically constructed to allow for such in-between responses; in the context of Moral Foundations Theory, a trade-off would constitute a compromise not between deontology and utilitarianism, or emotion and reason (Greene, 2023), but between the different Moral Foundations (or groups thereof).

Based on Moral Foundations Theory, by its very way of computing scores in a linear fashion, we might expect a linear evaluation process of some sort whenever possible. That is, choice options are evaluated with respect to each individual moral value, whereupon the linear combinations of these assessments are computed, and e.g. their length used as a metric. A first-aid practitioner might decide on a compromise between sticking to protocol and wanting to help, and help the injured person to go wash their wound.

This chapter focuses on the way conflicting moral preferences play a role when a decision is made. We will study the integration of different values in applied moral choice using the domain of charitable giving—the kind of charitable giving which occurs for intrinsic reasons, where an individual's moral values or what the individual deems worthy to support determines the extent of their intrinsic motivation to donate; we chose this domain because in addition to being suitable for empirically investigating our theoretical

question of multiple-value integration, learning more about people's donation behaviour also has practical value.

3.2 Charitable giving

If I can stop one heart from breaking,
I shall not live in vain;
If I can ease one life the aching,
Or cool one pain,
Or help one fainting robin
Unto his nest again,
I shall not live in vain.

—Emily Dickinson

People give to charity for various reasons. Some are altruists who feel a genuine desire to help others; others don't feel that emotional pull but instead experience a sense of obligation: their personal values tell them to reduce the suffering of others if they can, not because they intrinsically want to, but because they themselves hold the belief this is what people should do.

Thus, for any given charitable cause, there can be a set of different possible motivations to support it. An animal lover, say, might opt to donate to a local shelter, or contribute to a friend's fundraiser for an animal rights organisation (perhaps along with a supportive message on social media), or even opt to give a few dollars to a zoo to name a cockroach after their ex-partner on Valentine's day (Elassar, 2020). Broadly speaking, the spectrum of motivations for charitable giving can be grouped into three kinds: intrinsic reasons (desire to do good by supporting a given charitable cause, e.g. by giving to a pet shelter), extrinsic incentives (money or a non-monetary reward, such as a cockroach naming certificate),

and image motivation (acting in line with one's desired image, for signalling to others—but also related motifs including self-actualization, for instance by letting your social circle know about your donation to a fundraiser started by a friend). These three areas can interact in unexpected ways, such as a reduced willingness to donate when a monetary reward is introduced (Ariely, Bracha & Meier, 2009), and could also produce different patterns of behaviour that conflict with treating charitable giving as a value maximization problem.

In an idealized, 'spherical cow'-esque way, charitable giving could be viewed as the donor's attempt at solving a value maximization problem, 'How can I do the most good?'. From this perspective, the normative charitable giving strategy would be to figure out which organisation does the most of good per dollar, and support it as much as possible[1] (that is, employ the 'max'-ing strategy), and deviations from the optimal utilitarian solution would be attributed to cognitive biases.

This strategy is not a very good description of what people generally actually do. In reality, individual donation decisions are not fully determined by what a charity does or how well it does it. Affective motivations matter as well—people give to charity because it gives them warm, fuzzy feelings known as "warm glow" (Andreoni, 1990; Loewenstein & Small, 2007), because they are socially influenced by their peers (Reyniers & Bhalla, 2013), or because they perceive it to be a part of their own identity (Chapman, Masser & Louis, 2020).

Moreover, donors are known to exhibit some curious patterns. In particular, people are

[1]This is the approach advertised by the Effective Altruism movement: "Effective altruists will feel the pull of helping an identifiable child from their own nation, region, or ethnic group but will then ask themselves if that is the best thing to do. They know that saving a life is better than making a wish come true and that saving three lives is better than saving one. So they don't give to whatever cause tugs most strongly at their heartstrings. They give to the cause that will do the most good, given the abilities, time, and money they have available" (Singer, 2015).

drawn towards charities with lower overhead costs (Camerer et al., 2018; Gneezy, Keenan & Gneezy, 2014), even when differences in cost-effectiveness are otherwise accounted for by explicitly adding effectiveness information (Caviola, Faulmüller, Everett, Savulescu & Kahane, 2014). This suggests that it remains aversive beyond its use as an approximation for how well a charity works. Rather than allocating their resources to support a particular organisation, people tend to do what is known as *splitting* or *diversification* in the literature (Baron & Szymanska, 2011)—they employ the 'mix' strategy, that is, distribute their donation across multiple charities.

Previous work focused on decreasing overhead cost sensitivity to increase donation rates, comparing aversion to different types of overhead (salaries vs fundraising) (Portillo & Stinn, 2018), and incentivising donations to effective charities, for example by providing effectiveness information (Berman, Barasch, Levine & Small, 2018; Caviola, Schubert & Nemirow, 2020; Caviola, Schubert, Teperman et al., 2020). In this chapter, we focus on donation splitting behaviour and attempt to use overhead costs as an intervention: We test whether introducing fixed overhead costs associated with donation-splitting can reduce diversification in the absence of new information.

Costly costs

Charities incur a higher cost in processing a small donation compared to larger amounts. If you donate to a charity online using a credit card, a portion of your donation covers the processing fee that consists of a percentage of the transaction amount plus a flat fee (PayPal editorial staff, n.d.). If you donate $5 to charity, you lose $0.41 in processing costs and the charity gets $4.59. If you have $5 to give but two charities that you like a lot and you distribute your donation between the two, $4.29 reaches the charities. If you give $1 to a charity using PayPal, in reality, they're only getting $0.68.

Additionally, depending on the size and level of organisation within the charity, there are costs associated with processing your donation internally. Combining donations from different sources (online payments, check, wire transfer) and producing a receipt or thank you note or other forms of follow-up require additional effort that may make very small donations net-negative in terms of impact.

The financially optimal strategy in a deterministic setting with overhead cost would be to allocate one's available resources entirely to one's favourite charity as failing to do so results in a decreased overall impact.

Moral preferences

People's subjective preferences regarding a charity's cause area still have an effect even when cost effectiveness information is provided—a not unintuitive result: An individual wishing to make a donation towards cancer research in honor of someone who passed away may opt to give to a cancer research organisation, despite learning that focusing on child poverty would save more lives per dollar; while the alternative could do 'more good' from the perspective of a utilitarian 'lives saved' calculus standpoint, it remains an inferior way of supporting cancer research. Varying moral preferences would still allow us to treat moral decisions as an optimisation problem, with the caveat that each person tries to do the most what-that-individual-considers-good; that is: for different individuals, a utility function may assign different utilities to the same real-world outcome.

Caring about a charitable cause twice as much as about another will presumably be reflected in one's donation behaviour when faced with an opportunity to contribute resources to one or both of these causes. The donor may opt to give all available resources to the preferred charity; or, they may try to distribute them in a way that reflects the proportion of preference. If they naturally do the latter, then the existence of fixed

overhead costs (of processing each donation) should affect their decision. How much are they willing to pay additionally in order to distribute their donation between both charities, rather than giving it all to the charity they prefer?

Dimensions of goodness

Framing the problem of choosing a charity and making a donation as a value maximisation problem might be an oversimplification of moral choice. People have diverse sets of moral values , suggesting that they will find different charitable organisations appealing to a varying degree. Likewise, Moral Foundations Theory (MFT) finds that there are at least five separate dimensions within which moral values are evaluated; these foundations can be grouped into two types, individualizing (harm, fairness) and binding (loyalty, authority, purity). Within-type foundations are somewhat correlated, leading some researchers to argue morality is one-dimensional (Schein & Gray, 2015). In our experiment in the previous chapter, we reproduced the MFT result that there are at least two, potentially conflicting, different moral systems or values ; in this two-dimensional space, it is not quite clear how an overall value would be computed and a choice made whenever values point in conflicting directions regarding the actions they would favour.

Kahane et al. (2018) introduced the idea that there are two moral dimensions that are evaluated separately and cannot easily be combined. The dimensions they suggest are the two different valences of the harm/care dimension, that is, 'How can I do the most good?' and 'How can I do the least bad?' are considered as separate questions. While it is very possible that the processes involved in evaluating negative outcomes (sacrifice) and positive outcomes (saving lives) in the care dimension are separate, there is an analogous, perhaps even stronger case to be made for the processes of binding and individualising morality which we know to be separate. This could imply that utilitarianism requires a

trade-off to be made that is fundamentally more difficult for people who score higher on binding morality foundation scores.[2]

We suspect that in the broad grouping into moral types, people who score highly on the moral foundations of loyalty, authority and sanctity will have more of a challenge choosing between charities that might contribute towards these values and others that focus on reducing harm or increasing fairness, than people with high scores for only harm and fairness.

For example, consider a religious individual with Christian beliefs who is facing a choice between a hypothetical religious charity, 'Christians Against Poverty International,' that aims to help reduce global poverty, and 'Feed The Poor International,' another hypothetical charity that reduces poverty but does not promote religion. All else equal, they will likely prefer the first. But letting them choose between 'Holy Ghost Only International,' a religious charity which doesn't really do much beyond promoting religion on the one hand, and 'Feed The Poor International,' the charity that reduces poverty but is not religious on the other, might prove more difficult, in which case they may be more inclined to split their donation.

Two possible strategies of integrating a set of moral values, each represented to a different degree within each charity, would be picking a favourite to give everything to ('maximization'), or distributing out one's donation between both presented charities ('mixing'). Conceivably, someone might pay attention to the attribute they care about most, or the attribute that's easiest to compare (Baron & Szymanska, 2011). In the above example, this could be support of religion or reduction of poverty; in which case, the individual would make decisions based on this prominent value. Or, if the individual cares half as much

[2]In other words: a mathematical reason, of sorts, why somebody might favour a more deontological line of reasoning might be that it provides a way of reducing the complexity of an option space which, because of its two-dimensionality, does not come supplied with a natural well-ordering.

about religion as about reducing poverty, they might pay attention to that attribute in half as many decisions that they make.

Alternatively, when the decision is not a forced binary choice but allows to allocate one's resources to both options, someone might opt to spread their credits among the charities (splitting their donation) in a ratio that corresponds to the ratio of their respective preferences. This allocation might occur in a non-linear fashion, or in a hierarchical order, for instance when purity-type morality could override any utilitarian harm reduction calculus. Perhaps people want to maximise their impact in more than one domain, aiming for proportional representation of their own values ('moral compass')? Then, paying this cost might be reasonable.

In its first part, this chapter aims to find out whether people 'maximize' the amount given to their preferred charity, or 'mix' their donation by allocating a portion of the total to each positively rated charity. From previous literature, we expected that people will often tend to 'mix' to diversify their donations, but we also know that people will want to avoid overhead costs due to overhead aversion; due to this, we then proceed to test whether the introduction of fixed costs related to processing their donation leads to more 'max'-ing behaviour.

The second part of this chapter investigates the role our participants' moral values play in the way they integrate those values to make donation decisions. In particular, we look at the way those values influence their preference for 'max'-ing or 'mix'-ing.

3.3 Background

To speak about values and preferences more precisely, let us introduce a formal setting for this chapter, borrowing some notation from revealed preference theory in economics

and adapting it to our needs.

3.3.1 Revealed preferences and rationality

In economics, the Generalised Axiom of Revealed Preference (GARP) is a measure of preference consistency (Andreoni & Miller, 2002; Varian, 1982). Let us introduce the notion of a binary relation as well as some properties preference relations can have, and state GARP using a notation to adapt it to our context:

Let $X = \{x_1, \ldots, x_n\}$ be the set of all available choice options. Suppose we have an observed choice x_i. Then we can define the following:

Binary relation:

Definition 3.1 (Binary relation)**.** Let $\succ$ be a binary relation on X, that is, a set of ordered pairs (x_i, x_j) that is a subset of $X \times X$. If (x_i, x_j) is in this set, we write $x_i \succ x_j$.

- The relation $\succ$ is called **complete** if for any $x_i, x_j \in X$ with $i \neq j$ either $x_i \succ x_j$ or $x_j \succ x_i$.
- The relation $\succ$ is **transitive** if for $x_i, x_j, x_k \in X$, given $x_i \succ x_j$ and $x_j \succ x_k$, it also holds that $x_i \succ x_k$.
- The relation $\succ$ is **cyclical** if there is an $m \in \mathbb{N}$ such that there is a sequence $(x_1, \ldots, x_m),\ x_i \in X$ such that $x_1 \succ x_2$, …, $x_{m-1} \succ x_m$ and $x_m \succ x_1$. It is called **acyclical** if it is not cyclical.
- The relation $\succ$ is called **symmetric** if $x_i \succ x_j$ implies $x_j \succ x_i$, and **asymmetric** if $x_i \succ x_j$ implies $x_j \not\succ x_i$.

Definition 3.2 (Preference relation)**.** A preference relation $\succ$ on X is a binary relation on X that is complete and transitive.

Definition 3.3 (Directly Revealed Preferred)**.** Let $\succeq_R$ be a preference relation on X. x_i is called **directly revealed preferred** to x_k, written $x_i \succeq_R x_k$, if x_k was in the set of available choice options when x_i was chosen. x_i is called **strictly** directly revealed preferred to x_k if $x_i \succeq_R x_k$ and $x_k \not\succeq_R x_i$.

Definition 3.4 (Indirectly Revealed Preferred)**.** x_i is **indirectly revealed preferred** to x_k if there is a sequence of choice options $(x_1, \ldots, x_m)$ that contains x_i such that $x_j \succeq_R x_{j+1}$ for all $j \in \{1, \ldots, m\}$, and $x_m \succeq_R x_k$.

Definition 3.5 (Generalised Axiom of Revealed Preference (GARP))**.** If x_i is indirectly revealed preferred to x_k, then x_k is not strictly directly revealed preferred to x_i.

If an agent behaves according to GARP, their behaviour can be *rationalised*, that is, described using a reasonably well-behaved utility function $u : X \to \mathbb{R}$—and vice versa. In this case, $x_i \succeq_R x_k \iff u(x_i) \geq u(x_k)$ and $x_i \succ_R x_k \iff u(x_i) > u(x_k)$.

In empirical research, subjects tend to casually violate GARP when it comes to goods purchasing decisions (Février & Visser, 2004; Mattei, 2000). Diaye and Urdanivia (2009) argue that this might generally stem from a violation of the transitivity axiom, and that it is possible that by loosening the requirements somewhat and only requiring acyclicity rather than transitivity, people's behaviour can be described by a generalised utility function. In a task which involved allocation of resources to oneself (selfishness) or another participant (altruism), people did not seem to violate GARP all that much (Andreoni & Miller, 2002).

Diaye, Gardes and Starzec (2008) maintain that while GARP violations would be a useful measure of irrationality in a static system, they can also be attributed to changes in the conditions under which a choice was made. If we believe that moral values are sampled in an inherently noisy process (see Chapter 2), we may expect a non-zero percentage due to that change over time alone in those cases where a small change would make a difference.

Because our task includes a selection of stimuli that are different in their representation of individualising and binding morality, it allows us not only to look at GARP violations in the context of charitable giving, but also to investigate whether people's tendency to violate GARP differs depending on their moral values.

3.3.2 The "Square"

Equipped with this framework, let us take a more formal look at the moral decision problem. In Chapter 2, we saw that there are probably at least two distinct (clusters of) values—binding and individualising-type morality—which need to be somehow combined. Therefore, we will need $n \geq 2$ dimensions[3] to describe locations in the moral domain.

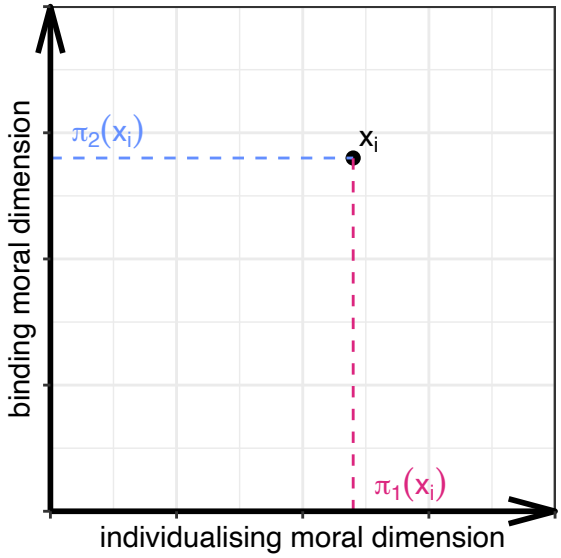


Figure 3.1: Projection mapping to obtain coordinates of choice option x_i alongside moral dimensions.

[3]If we believe that these clusters of values are entirely independent, we can assume them to constitute "moral dimensions", and by looking at the association between people's MFQ-type scores and their preferences, we can determine the location of each choice option scores along the two axes. For the sake of simplicity, let us pretend to believe that for a few more sections.

Here, we can decompose an action x_i into its components[4] using a projection map onto the j-th component (Figure 3.1):

$$\pi_j : X \to \mathbb{R}, \quad x \mapsto x^j$$

Note that the position (x_i^1, x_i^2) is a measure of the '*type 1-ness*' (extent of relevance for individualising-type morality) and '*type 2-ness*' (pertinence to binding-type values) of the action x_i. It does not determine the utility an agent might assign to this action because the extent to which each dimension matters will vary between agents.

Now, in order to make a choice, people make a moral choice by combining their different values to settle on an evaluation of each option. But how are those different aspects combined?

3.4 Experiment 1

3.4.1 Research design

3.4.1.1 Choice of stimuli

We chose the charities used as stimuli in the experiment using a pretrained word vector model: We downloaded a list of the 100 UK charities with the highest voluntary income (Charities Aid Foundation, 2017), added them to a pre-trained word vector model; we used the Google News text corpus (Mikolov, Chen, Corrado & Dean, 2013), and Python's gensim package (Řehůřek & Sojka, 2010). Then, we calculated a simple measure of the 'type1'-ness and 'type2'-ness of each charity by comptuting the average distance from each

[4] In our case, we are taking into account the two we are aware of, corresponding to the respective MFQ types ($X \cong \mathbb{R}^2$, so to say).

word vector in the charity description (excluding standard stopwords) from the names of the respective foundations ("harm," "care," "fairness," and "reciprocity" for individualising morality; "ingroup," "loyalty," "authority," "respect," "purity," and "sanctity" for binding morality). Doing this gave us a way for mapping[5] each charity onto a point $(t_1, t_2) \in \mathbb{R}^2$. Looking at the two-dimensional representations of each charity in moral space, we then manually chose four points far apart in the chart: a charity each for every combination of high/low individualising-type score and high/low binding-type score.[6]

3.4.2 Method

The experiment consisted of three parts. First, participants were asked to rate the charities presented to them: 'Please read the following list of charities and what they say about themselves. For each charity, indicate to what extent their work is aligned with your own values using the slider bar below the description.' Then, they were presented with different paired combinations of these charities, and asked to distribute a small donation ($1) between the two. Finally, they were asked to fill out the Moral Foundations Questionnaire (MFQ30).

3.4.2.1 Participants

In the first iteration of the experiment, we collected data from 100 participants on Amazon MTurk. We excluded 4 participants who either failed the sanity check (they disagreed with the statement 'it is better to do good than to do bad') or asked to be excluded.

[5]To stay in the framework introduced above, we don't explicitly know the coordinates $(x_i^1, \ldots, x_i^n)$ of x_i.

[6]It should be noted that we used this approach to choose four maximally different charities in a less arbitrary way than just picking them at will. While computational techniques using language representations may be useful to extract information on underlying ethics (Jentzsch, Schramowski, Rothkopf & Kersting, 2019), we don't think that this would be a very good way to evaluate how moral an organisation is—it would be easy to manipulate such an algorithm by generating adversarial samples.

3.4.3 Results

3.4.3.1 Rating-choice consistency

We consider a participant's ratings of each charity to be a measure of their respective strength of preference for that charity. As a first plausibility check of our results, we expected to find a positive correlation between a charity's ratings and the total amount donated to it. This seems plausible visually:

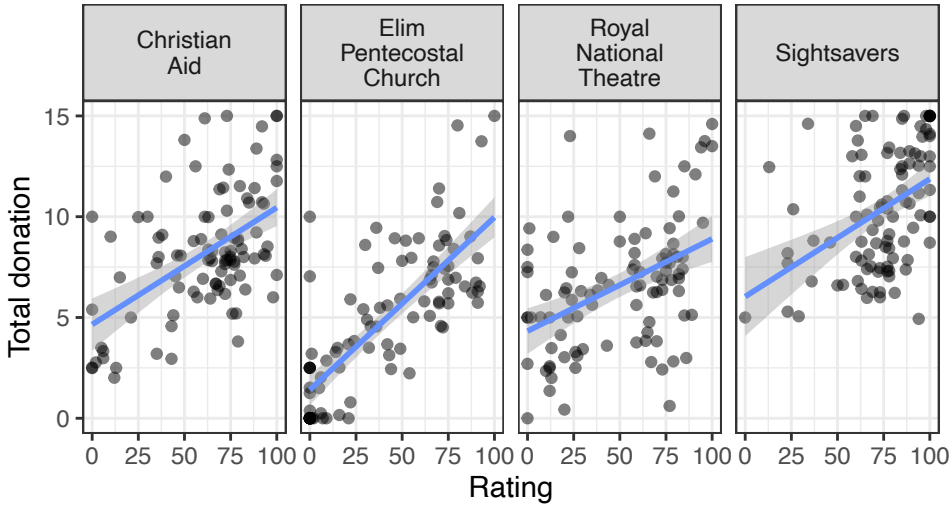


Figure 3.2: Total amount donated by each participant vs rating in Experiment 1

And indeed, we found strong evidence that the correlation coefficients are different from zero as assessed by a Bayesian test for linear correlation with the Pearson correlation for ratings and donations for each respective charity as follows in Table 3.1.

Table 3.1: Rating-choice consistency: Results of Bayesian test for linear correlation for Experiment 1

Charity	ρ	BF	p
Elim Pentecostal Church	0.85	3.1e+16	6.3e-20
Christian Aid	0.52	4.0e+05	3.6e-08
Royal National Theatre	0.42	2.5e+03	9.4e-06
Sightsavers	0.41	1.6e+03	1.5e-05

3.4.3.2 Effect of moral values

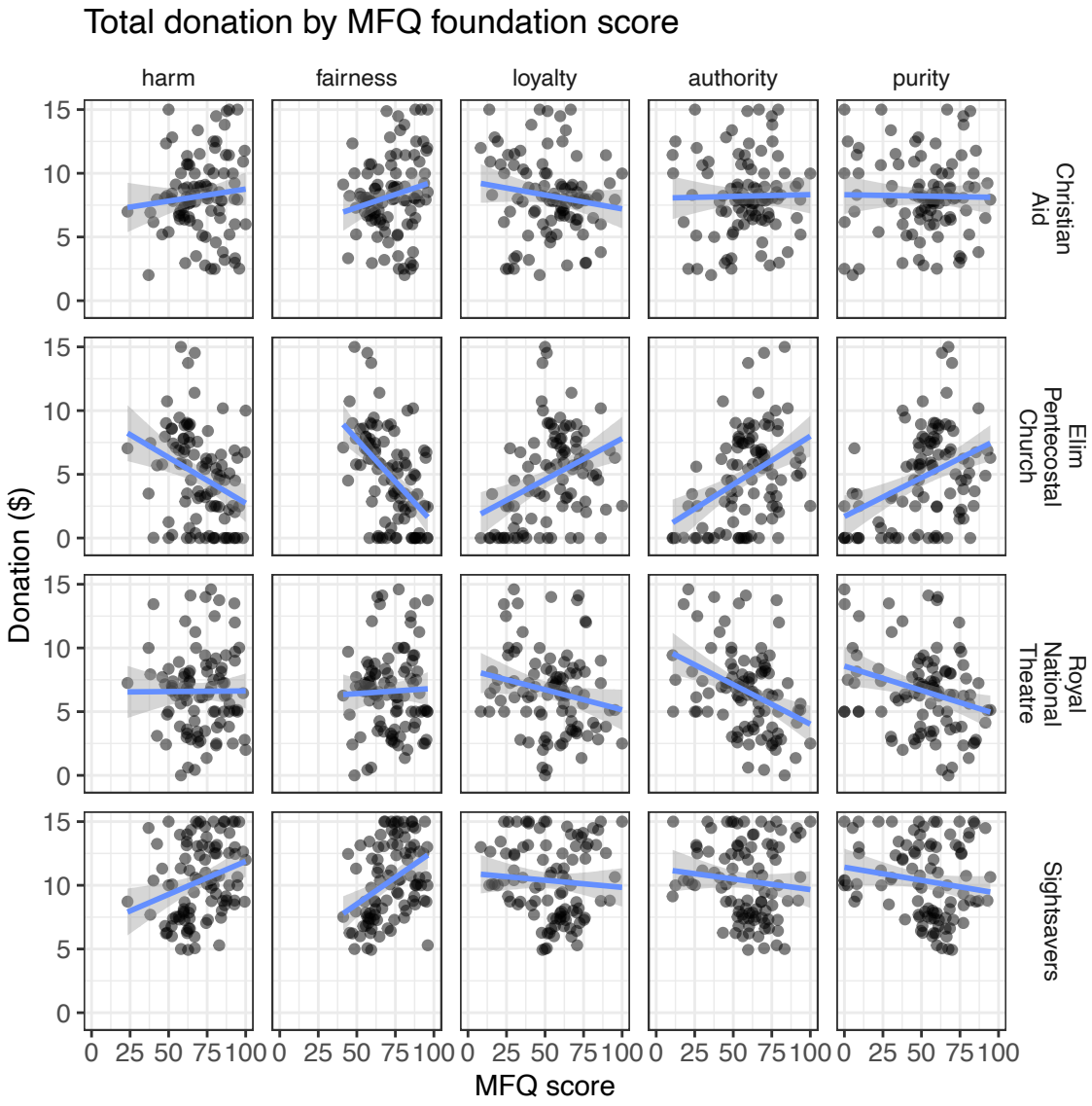


Figure 3.3: Total donation to each charity and individual MFQ foundation scores

As we can see in Figure 3.3, there appears to be a relationship between MFQ foundation scores and how much is donated to each charity throughout the experiment. Yet, we can also see that harm and fairness on the one hand, and loyalty, authority, and purity on the other hand, appear to have a similar effect, suggesting once more that for the choice we are looking at there is more of a functional difference between types than between the individual foundations. Let us therefore continue to look at the effect of moral values from the perspective of Figure 3.4.

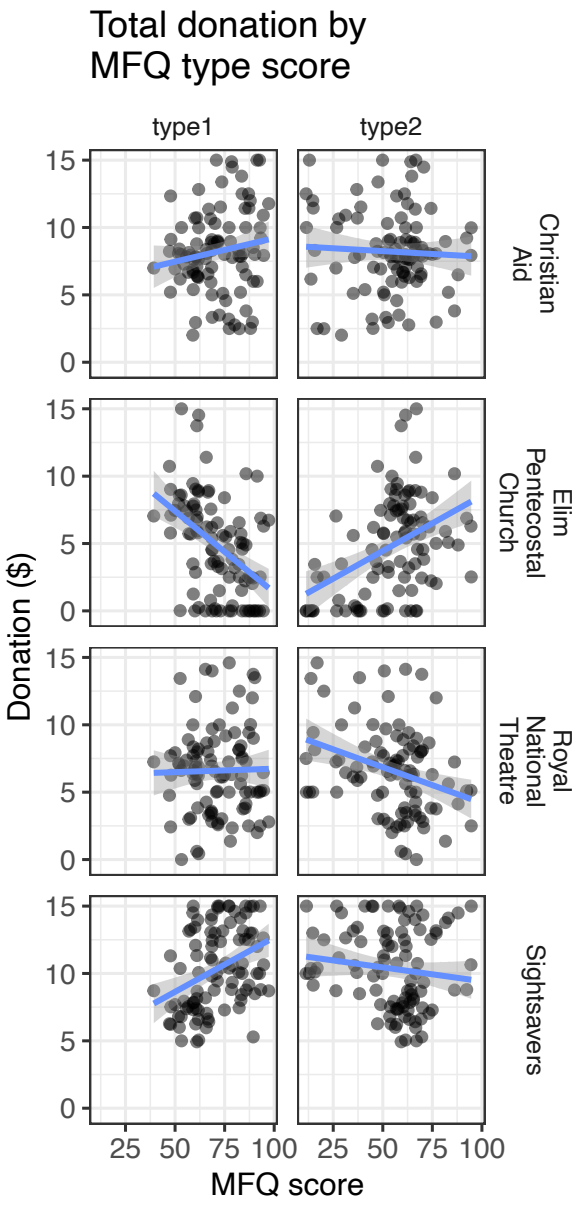


Figure 3.4: Total donation to each charity and MFQ type scores in Experiment 1

Now, do moral values exert an influence not only on the amount donated, but on whether or not a participant chooses to *max* rather than pick an in-between value? From Figure 3.5

we can tell that at first glance, the two types appears to have opposite effects on donation pattern itself: While someone with a high[7] type 1 (individualising-type morality) score, on average, will have a higher proportion of *max* choices (48.7% *max* for high individualising scores compared to 19.5% for low individualising scores), an individual with a higher binding-type score will, by and large, tend to prefer the *mix* strategy (20.7% *max* trials for high, i.e. above-median, binding-type scores compared to 43.7% for low binding-type scores).

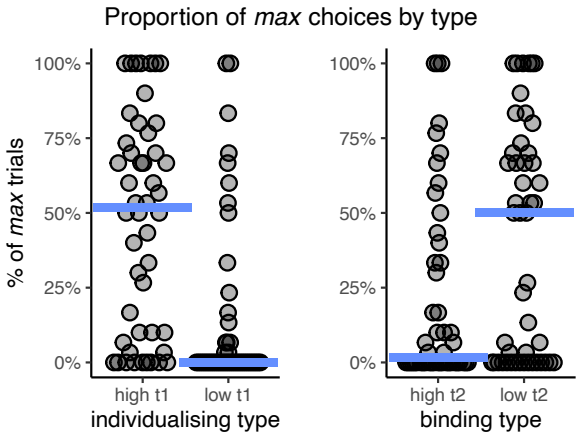


Figure 3.5: Proportion of max trials for high and low (as defined by median split) type scores in Experiment 1

[7] *high* is defined by a median split. The median was at 69.96 for type 1 and 58.5 for type 2.

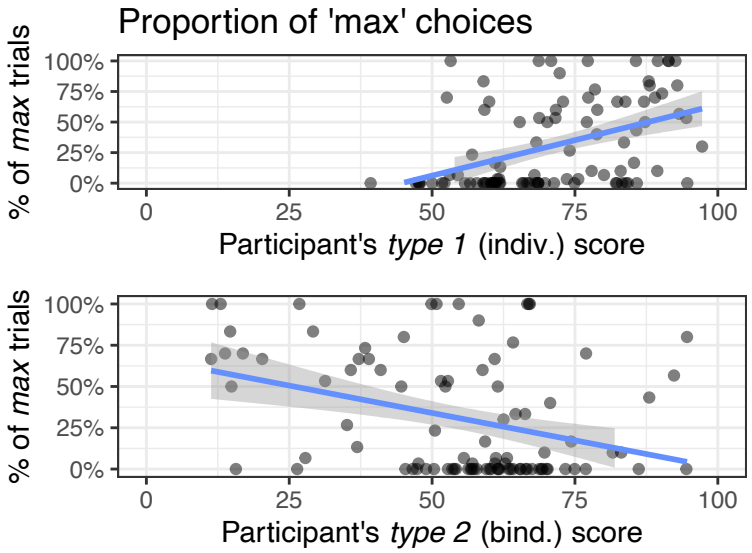
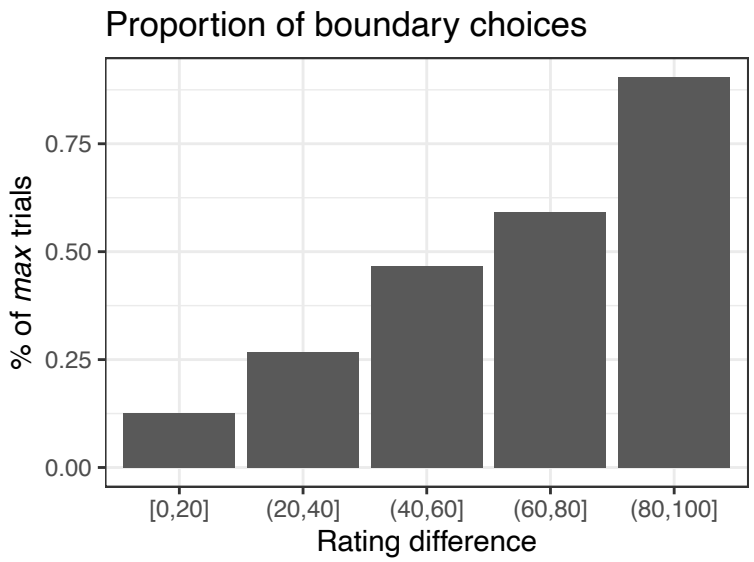


Figure 3.6: Type-dependency of 'max' choice proportion

It is worth mentioning that at this point, that difference is not entirely enough to conclude that people's *type* scores influence their donation pattern. The charities were chosen to be as different as possible alongside our two dimensions, and we saw that people's charity ratings (a measure of how much they like each charity) influenced their donation behaviour in terms of total amount. Perhaps their type scores were factoring into their ratings of each charity in a way that made the *difference in ratings* higher for some (t_1, t_2) type score combinations than for others. And those ratings differences, in turn, influenced the donation pattern—perhaps a person is more likely to 'mix' if their preference ratings for both charities do not differ by much?

A first look tells us that the proportion of 'max' trials seems to increase as the difference in ratings increases—implying, perhaps, 'the clearer a preference for one charity over another, the clearer the choice':



Values or preferences? To see if our observation can be explained by the rating difference alone, let us describe these explanations as models.

For Participant i with type scores $(\text{type}_{1,i}, \text{type}_{2,i})$, their difference in ratings is the numerical variable rating_difference $:= |\text{rating}_i(\text{charity}_1) - \text{rating}_i(\text{charity}_2)|$ for each charity pair $(\text{charity}_1, \text{charity}_2)$ and each individual choice, let us look at the binary outcome variable y_i, where

$$y_i = \begin{cases} 1 & \text{if their donation to charity}_1 \text{ or charity}_2 \text{ equals 1} \\ 0 & \text{otherwise} \end{cases}$$

.

We can describe our different possible explanations as a set of logistic regression models:

$$y_i = \beta_{i0} + \epsilon_i \tag{rd0}$$

$$y_i = \beta_{i0} + \beta_1 \cdot \text{type}_{1,i} + \beta_2 \cdot \text{type}_{2,i} + \epsilon_i \tag{rd1}$$

$$y_i = \beta_{i0} + \beta_1 \cdot \text{rating\_distance} + \epsilon_i \tag{rd2}$$

$$y_i = \beta_{i0} + \beta_1 \cdot \text{type}_{1,i} + \beta_2 \cdot \text{type}_{2,i} + \beta_3 \cdot \text{rating\_distance} + \epsilon_i \tag{rd3}$$

$$\begin{aligned} y_i = {} & \beta_{i0} + \beta_1 \cdot \text{type}_{1,i} + \beta_2 \cdot \text{type}_{2,i} + \beta_3 \cdot \text{rating\_distance} \\ & + \beta_4 \cdot \text{rating\_distance} \cdot \text{type}_{1,i} + \beta_5 \cdot \text{rating\_distance} \cdot \text{type}_{2,i} + \epsilon_i \end{aligned} \tag{rd4}$$

Our first two models both can be considered to be baseline models (the second model allows for a baseline tendency to 'max' or 'mix' that depends on moral type scores). The third model captures the hypothesis that type scores only have an effect on the tendency to max or mix by proxy because they influence the extent to which one charity is rated higher than another. The fourth and fifth models allow for types to have an additional effect, beyond rating distance alone—without and with interactions, respectively.

Out of these models,[8] we find that (rd3) performs best. We can see strong evidence compared to the baseline models ($\text{BF}_{3,0} = 9.74 \times 10^{66}$ and $\text{BF}_{3,1} = 1.51 \times 10^{61}$), and further, it outperforms the model which explains the tendency to 'max' or 'mix' in terms of rating distance ($\text{BF}_{3,2} = 1.28 \times 10^{4}$) as well as the interaction model ($\text{BF}_{3,4} = 9.69 \times 10^{4}$).

From the model comparison, we can conclude that people's moral values, as measured by their type scores, not only influences people's donations because it affects their rating of each charity, but that it has a further, additional effect on their tendency to 'mix'.

[8] All Bayesian models in this dissertation were fitted in R using the `brms` package (Bürkner, 2017). Bayes factors for model comparison were computed using `bayestestR` (Makowski, Ben-Shachar & Lüdecke, 2019).

3.5 Experiment 2

In the domain of charitable giving, mixing behaviour is of interest beyond purely theoretical reasons because in practice, it makes donations less effective due to added effort in processing. We conducted a follow-up experiment to investigate to what extent donation patterns could be manipulated by making mixing more expensive. If we can successfully incentivise participants to choose a 'max' strategy, a similar way could be applied as an intervention to decrease the costs of inefficiency in individual charitable giving, perhaps with the added benefit of encouraging people to direct their full support to the causes most important to them.

3.5.1 Method

The follow-up experiment, again, consisted of three parts; but now, participants were randomly assigned to one of four conditions which would influence the choices available to them in the second part of the experiment. In the first part, participants were asked to rate the charities presented to them. Then, they were presented with different paired combinations of these charities, and asked to distribute a small donation ($1) between the two; depending on the experimental condition, there was a varying overhead cost associated with each charity chosen. Finally, they were asked to fill out the Moral Foundations Questionnaire (MFQ30).

In the first and second condition, there was a $0.20 overhead cost associated with a donation to each charity; should the participant choose to allocate the total sum to one of the charities, the total allocated amount would be $0.80; otherwise, the total donation amount would be $0.60. The first condition allowed for the allocation of a negative 'donation': if the participant chose to give $0.05 to the first charity, that charity would

'receive' $-0.15 (subtracted from the total amount donated to the charity for the experiment overall), and the other one $0.65. The second condition didn't allow for 'negative donations' and the corresponding portion of the slider bar could not be selected. The third condition included a $0.01 overhead cost, and the fourth condition no fixed overhead.

3.5.2 Results

We recruited 400 participants on Amazon MTurk (for approximately 100 per condititon). In total, we had to exclude 142 participants who failed at least one of the attention checks (either disagreed with 'it is better to do good than to do bad,' or failed to select the right cause area for two or more of the charities after reading the description).

3.5.2.1 Confirmatory analysis

This section contains the preregistered, confirmatory analysis that lays the foundation for this study, followed by an exploration of the relationship between people's moral values and responses. The preregistration document can be found here: `https://osf.io/654kn/`

Max or mix We aimed to find out whether people tend to 'maximize' their impact by choosing their preferred charity and allocating all available resources to that one (max), or whether they prefer to spread their donation by spreading the resources among charities that appeal to them (mix).

Our overall expectation was that people likely employ max and mix strategies a part of the time, but that they respond adaptively to cost incentives in favour of 'max'-ing: When there is a fixed overhead cost associated with processing a donation for each charity, there should be more 'max' behaviour. Further, we expected participants to be sensitive to that

Table 3.2: Rating-choice consistency: Results of Bayesian test for linear correlation

Charity	ρ	BF	p
Elim Pentecostal Church	0.81	3.7e+40	2.9e-44
Christian Aid	0.57	1.5e+19	2.9e-22
Royal National Theatre	0.49	5.0e+13	1.3e-16
Sightsavers	0.36	5.3e+06	2.5e-09

cost; a higher overhead cost would lead to a higher increase in 'max'-ing because people would have a stronger incentive not to spread out their donation.

Rating-choice consistency As in Experiment 1, we are considering a participant's ratings of each charity to be a measure of their respective strength of preference for that charity. Here, too, we expected to find a positive correlation between a charity's ratings and the total amount donated to it, and found strong evidence in favour of such a relationship, as evaluated by a Bayesian test for linear correlation with the Pearson correlation for ratings and donations for each respective charity (see Table 3.2).

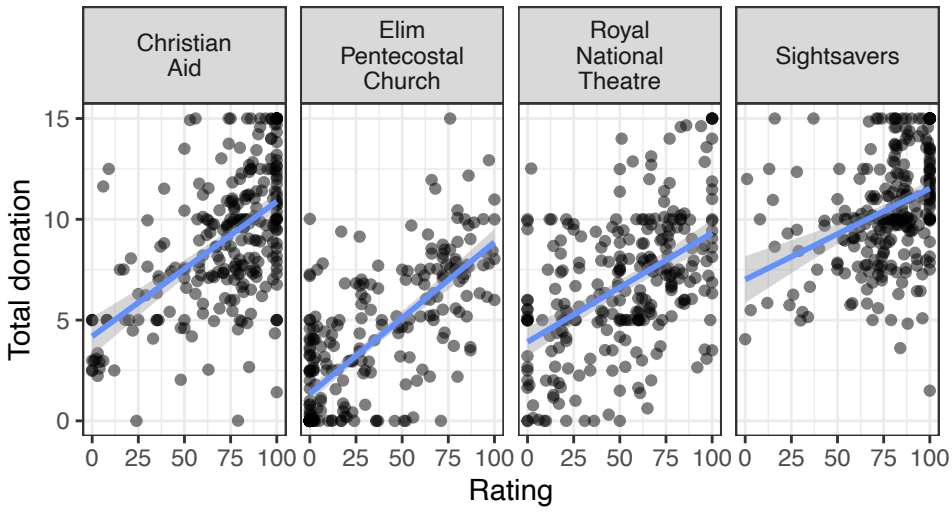


Figure 3.7: Total amount donated by each participant vs rating

Table 3.3: t-test for non-triviality of donation pattern

Condition	μ_0	BF_1	t	df	p
$0.2	0	1.7e+08	7.9	61	5.8e-11
$0.2, donations >= 0	0	1.2e+08	7.8	63	7.8e-11
$0.01	0	9.0e+12	10.6	65	7.2e-16
$0	0	3.6e+10	9.1	68	2.1e-13
$0.2	1	4.6e+13	-11.3	61	1.3e-16
$0.2, donations >= 0	1	9.3e+13	-11.4	63	6.5e-17
$0.01	1	2.3e+16	-12.8	65	2.3e-19
$0	1	1.7e+16	-12.5	68	3.1e-19

Donation patterns Our hypothesis here is that most people employ both max and mix strategies for a part of the time, i.e. the proportion of 'max' trials is larger than 0 and smaller than 1. And indeed, for each condition, this is what Bayesian t-tests show.

We observe strong evidence that the proportion of max and mix trials for each condition does neither equal to 0 nor to 1 (Table 3.3), in line with the visual presentation of results (Figure 3.8). Overall, participants chose the max strategy in about 42% of all trials. 13.8% of participants consistently chose to max, while 29.1% of participants consistently chose to mix.

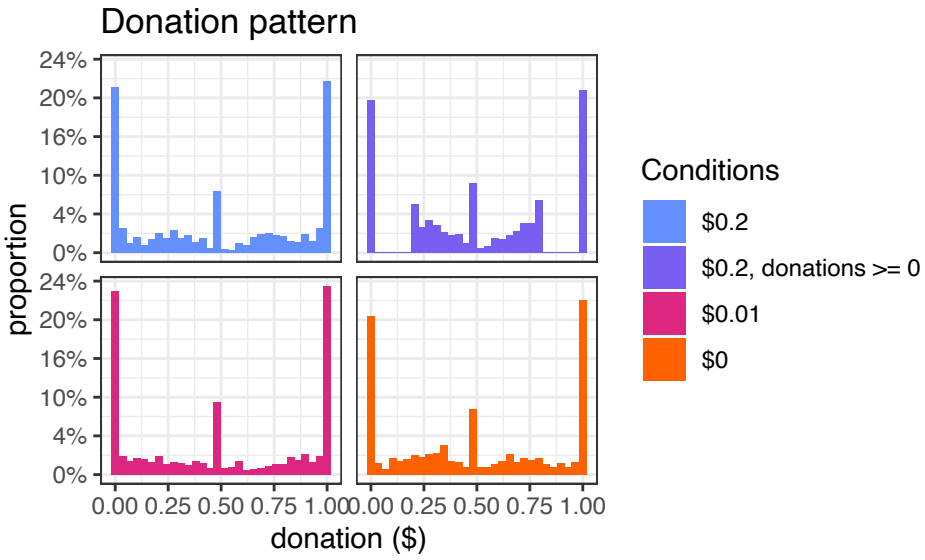


Figure 3.8: Donation patterns by condition

Effects of preference and cost We thought people would want to minimize that overall overhead—when there's a cost associated with making more than one donation people would rather collapse their choice and focus on one of the charities.

Based on the rationale that people would want to maximise their overall donation impact, we expected that people would increasingly allocate the full donation amount to one of the charities rather than distributing it between both and facing the additional fixed cost. We therefore expected that the proportion of 'max' trials in the $0.01 and both $0.2 groups would be higher than in the $0 group, respectively, as assessed by a Bayesian t-test.

We further suspected this effect might be more pronounced with a higher fixed cost. Then, in the '$0.2 cost' condition, the 'max' proportion should be higher than in the '$0.01' condition, as assessed by a Bayesian t-test.

We did not find any evidence for either of these assumptions and, instead, moderate evidence against there being a difference in proportion (see Table 3.4), for all between-

Table 3.4: Effect of fixed cost (condition): Results of two-sided t-test checking for a non-zero difference in means

Condition 1	Condition 2	μ	$BF_{1,0}$	t	df	p
$0.2	$0	0	0.19	-0.142	125	0.89
$0.2, donations >= 0	$0	0	0.19	-0.218	128	0.83
$0.01	$0	0	0.21	0.513	133	0.61
$0.2	$0.01	0	0.23	-0.627	120	0.53
$0.2, donations >= 0	$0.01	0	0.24	-0.705	122	0.48
$0.2	$0.2, donations >= 0	0	0.19	0.073	124	0.94
$0.2, donations >= 0	$0.2	0	0.19	-0.073	124	0.94

condition comparisons of the per-participant proportion of 'max' trials.

Cost-based proportion increase Cost-based proportion increase follow-up: We intended to use a Bayesian t-test comparing the 'max' proportion to the proportion of values in the $0 condition that fall into the now-prohibited interval and initial 'max' responses.

(1) In the '$0.2, no negative donations' condition, the increase in proportion of 'max' trials is higher than it would be solely by shifting the now prohibited responses from the '$0' condition to 'max' because it incentivises choosing 'max' or 'mix' even when not required by slider position limits.

(2) In the '$0.2, no negative donations' condition, the increase in proportion of 'max' trials is lower than it would be solely by shifting the now prohibited responses to 'max'; instead, some of the responses are shifted to the 'mix' area.

(3) In the '$0.2, negative donations possible' condition, a Bayesian t-test will show if the proportion of negative donations is positive.

Overall, we did not observe the effect we expected (increase in 'max' trials). Since we did not observe an increase, this part of the analysis is not necessary, and the second

hypothesis above is most promising. We found evidence supporting a stronger version of (2): All of the now-prohibited responses[9] are shifted to the mix area.

3.5.2.2 Exploratory analysis

Comparing responses within the different experimental conditions, we failed to see differences in sensitivity to overhead cost depending on people's moral values. But do binding and individualising morality scores have an effect on a participant's general tendency to max or mix?

Morals and donation patterns To examine the relationship between individualising and binding morality and donation patterns, we split our subject pool into two groups using a median split for binding morality scores. We can see the same qualitative pattern between the two groups as in Experiment 1: A higher individualising score is associated with more 'max'-ing, a higher binding score to more 'mix'-ing.

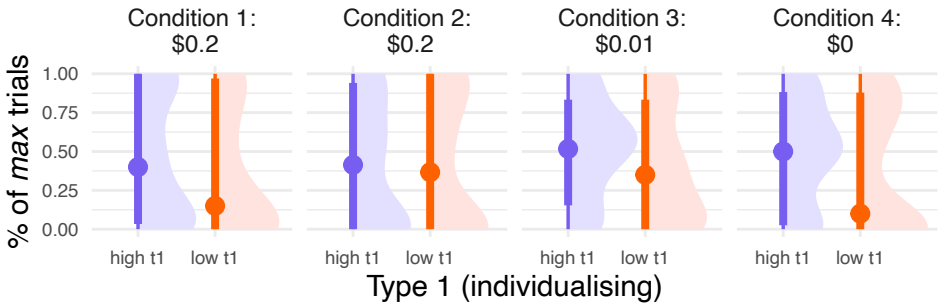


Figure 3.9: Proportion of max trials by individualising-type median split in Experiment 2. Participants who scored higher on individualising foundations showed a higher proportion of max trials, opting to allocate the entire donation to one or the other charity.

[9]Now prohibited responses in zero-cost condition (where 0 < Money1 < 0.2 or 0.8 < Money1 < 1): 16%; negative donations: 24%

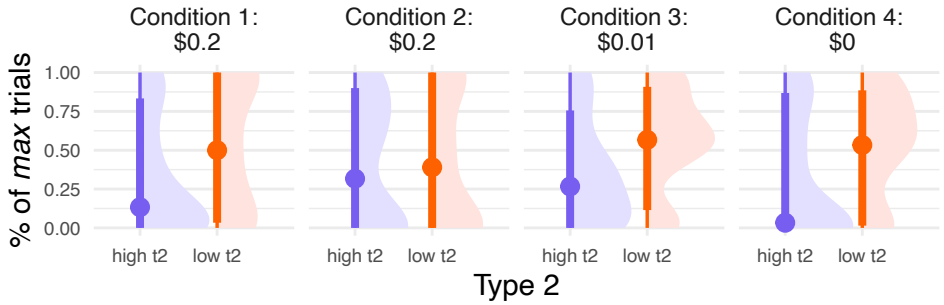


Figure 3.10: Proportion of max trials by type 2 median split in Experiment 2. Participants with an above-median type 2 score, on average, appear to choose to max less frequently.

Values or preferences? As in Section 3.4.3.2, we can use a set of logistic regression models which describe the binary outcome variable (was it a 'max' choice?) in terms of an individual's type scores and the difference in preference ratings for the options.

Response times While we did not find a difference in response times between the different conditions, we found a difference in response times when it comes to donation patterns: 'mix' trials take more time overall, $\text{rt}_{\text{mix}} = 6.21 \pm 14.95$, compared to 'max' trials, $\text{rt}_{\text{max}} = 5.09 \pm 9.65$; this difference was significant, as assessed by an ANOVA of a mixed effects model $\text{rt} \sim \text{is\_max\_trial} + (1|\text{participant\_id})$, $F(1, 1187) = 11.6$, $p = 6.98 \times 10^{-4}$. Yet, we did not find a difference in response times in the participant groups low/high on either of the two type scores. (It is worth noting that these measurements are not very sensitive since we only recorded reaction times rounded down to the nearest second since they were tracked based on server timestamps stored in Unix time.)

Cost-based differences in strategy Although we did not observe an overall shift towards 'max'-ing in the '$0.2 cost'-conditions, this might be due to effects that counteract each other. For instance, there might be a larger number of individuals that choose 'max'

consistently, balanced out by the other participants choosing 'max' less frequently, making the effect disappear in aggregate. Or, conversely, some individuals might be choosing 'max' more often, while others opt for consistency in 'mix'-ing (say, because they want fairness with respect to the donations a charity receives).

Morals and overhead aversion Comparing responses within the different experimental conditions, we failed to see differences in sensitivity to overhead cost depending on people's moral values. But do binding and individualising morality scores have an effect on a participant's general tendency to max or mix?

As measures of people's moral values, we look at an individualising morality type score and a binding morality type score for each participant, obtained by averaging the MFQ30 scores for the respective foundations, an approach used in previous literature (Iurino & Saucier, 2020; I. H. Smith, Aquino, Koleva & Graham, 2014; Van Leeuwen & Park, 2009).

While designing the experiment, our assumption was that people's charity ratings would reflect their moral values (Nilsson, Erlandsson & Västfjäll, 2020). In confirmation of this, we find significant correlations between individualising morality type scores and ratings for all four charities, and binding morality type scores and ratings for the two non-secular charities.

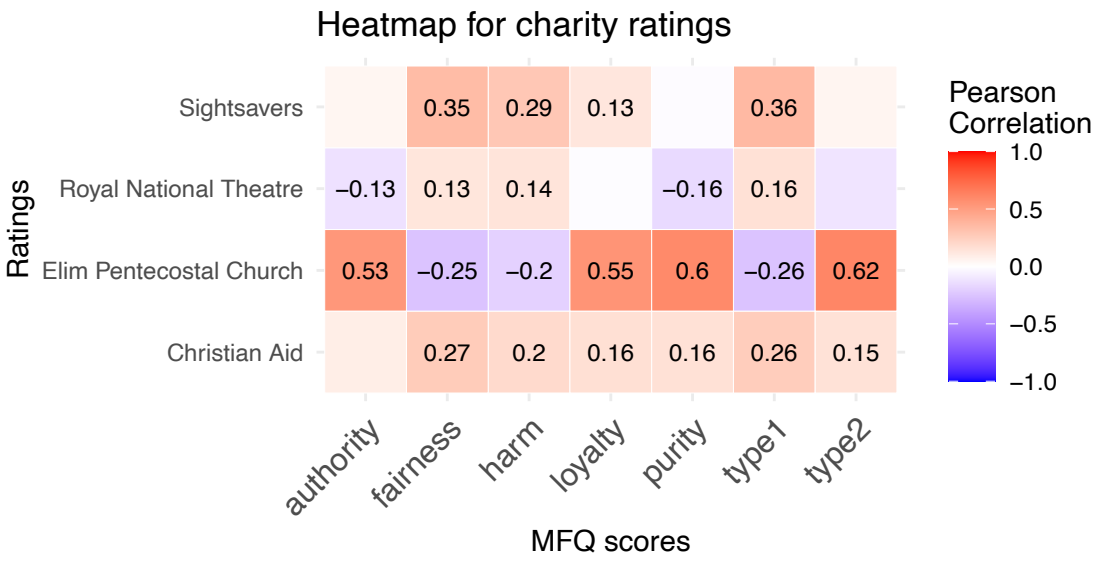


Figure 3.11: Correlation heatmap for charity ratings and MFQ scores, as well as type scores; numbers only displayed where significant ($p<.05$). While binding-type correlations exhibit the intended qualitative pattern (no significant correlation for secular organisation), individualising-type scores correlated negatively with EPC ratings, rather than not at all.

As expected, for individualising morality, we see a positive correlation for donations to Sightsavers. (The correlation for Christian Aid is positive but not significant.) Surprisingly, we see a negative correlation for donations to Elim Pentecostal Church. For binding morality, all correlations are significant; there is a positive correlation for donations to Elim Pentecostal Church, and negative correlations for donations to the other charities, including Christian Aid. These different patterns suggest some relationship between individualising and binding morality scores and charity ratings. But what does this relationship look like beyond the direction of influence? How can we translate these correlations into a model of the role individualising and binding morality play on participants' charity preferences?

As a baseline, we want to formalise a version of the intuition 'for a given participant, their moral type score provides no useful information in predicting their charity ratings,' that is: 'for each participant, a model that includes their moral type scores is not better than a model that predicts their ratings to be randomly determined.'

Let us consider the following three models describing the rating for participant i for each charity (with charity as a categorical variable):

(m0) $y_i = \beta_{i0} \cdot \text{charity} + \epsilon_i$

(m1) $y_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i + \beta_2 \cdot \text{type2}_i + \epsilon_i$

(m2) $y_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i \cdot \text{charity} + \beta_2 \cdot \text{type2}_i \cdot \text{charity} + \epsilon_i$

(m3) $y_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i + \beta_2 \cdot \text{type2}_i + \beta_3 \cdot \text{type1}_i \cdot \text{charity} + \beta_4 \cdot \text{type2}_i \cdot \text{charity} + \epsilon_i$

The first model corresponds to the baseline model described above; there is no relationship between type scores and ratings. The second one allows for type scores to have an effect on overall charity ratings but not specific to a given charity (for example, people who score higher on loyalty might value charitable giving more than those who don't, and consequently generally rate charities higher). The third model additionally allows for interactions between charities and type scores, capturing a potential effect type scores can have on how much charities appeal to a participant, while having a different or no effect on the ratings of others.

Results: Strong evidence favouring m1 over m0 ($\text{BF}_{1,0} = 7.2 \times 10^7$) and m2 over m1 ($\text{BF}_{2,1} = 2.78 \times 10^{37}$). Inconclusive evidence for m3 compared to m2 ($\text{BF}_{3,2} = 1.12$).

3.5.2.3 Models with donation as target variable

Next, let us check if this relationship extends beyond ratings onto the choices people make. We compare a set of models equivalent to (m0)-(m2) above, where z_i denotes the

donation given by Participant i to a particular charity.

(m10) $z_i = \beta_{i0} \cdot \text{charity} + \epsilon_i$

(m11) $z_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i + \beta_2 \cdot \text{type2}_i + \epsilon_i$

(m12) $z_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i \cdot \text{charity} + \beta_2 \cdot \text{type2}_i \cdot \text{charity} + \epsilon_i$

(m13) $z_i = \beta_{i0} \cdot \text{charity} + \beta_1 \cdot \text{type1}_i + \beta_2 \cdot \text{type2}_i + \beta_3 \cdot \text{type1}_i \cdot \text{charity} + \beta_4 \cdot \text{type2}_i \cdot \text{charity} + \epsilon_i$

Results: Strong evidence favouring m12 over m10 ($BF_{12,10} = 2.32 \times 10^{27}$) and m12 over m11 ($BF_{12,11} = 8.11 \times 10^{30}$). Inconclusive evidence for M13 compared to M12 ($BF_{13,12} = 0.99$).

3.5.2.4 Influence of individual MFQ scores

Interpretation of MFQ scores:

One might assume that a high score for 'fairness' might predict more 'mix' responses—after all, donating to all charities equally is a simple form of distributive fairness. So it appears somewhat odd that we see the opposite:

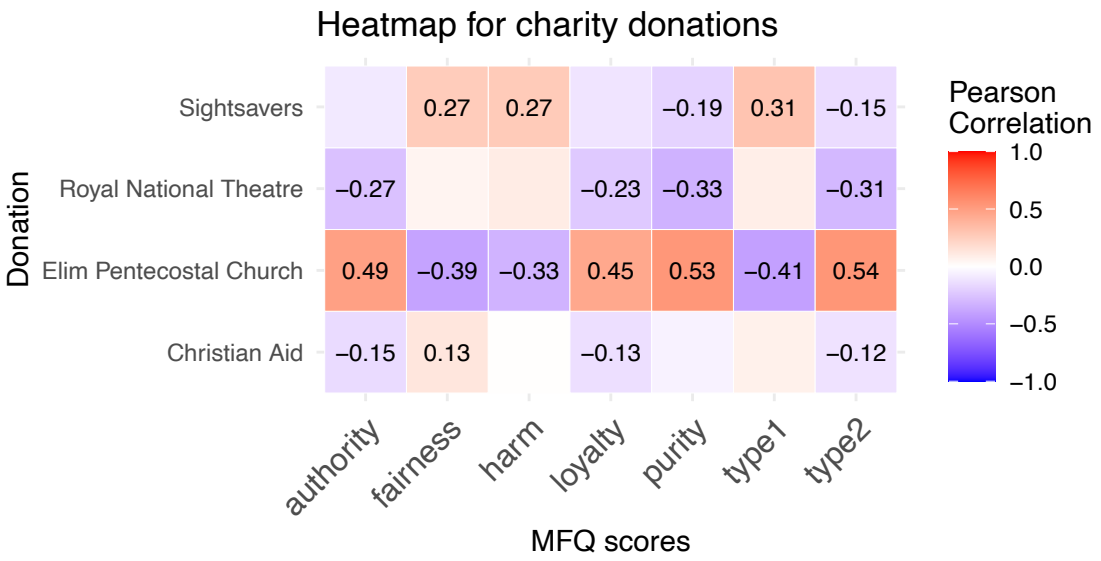


Figure 3.12: Correlation heatmap for total charity donations and MFQ scores, as well as type scores; numbers only displayed where significant (p<.05). Correlations of type scores and donation amounts differed from the pattern we observed for ratings.

Across conditions, scores for fairness and harm are positively correlated with the proportion of 'max' trials, while scores for loyalty, authority and purity are negatively correlated with it, again suggesting thinking in terms of types rather than individual foundations for the direction of influence.

3.5.2.5 GARP violations

Let us examine all donations which each participant allocated over the course of the experiment as choices: On each trial, some donation amount (d_i, d_j) is spread between a charity pair $(c_i, c_j) \in C \times C$ selected among out charities $C = \{\text{CA}, \text{EPC}, \text{RNT}, \text{SS}\}$. If $d_i > d_j$, we consider this as the participant's choice of c_i within the subset of available options $\{c_i, c_j\} \subset C$.

As specified in Def. 3.5, we are looking for specific cases of preference reversals, in which c_j is strictly directly revealed preferred to c_i and, at the same time, option c_i is indirectly revealed preferred to c_j. In our setting, this means we are searching for participants for whom we can observe both $c_j \succeq_R c_i$ and a chain of preferences linking c_i and c_j into the opposite direction, $c_i \succeq_R ... \succeq_R c_j$.

The perhaps simplest way for a participant to violate GARP in our task is violating asymmetry, that is, allocating the larger donation among the charity pair $(c_i, c_j)$to c_i on some trial and to c_j on another. In our experiment, 55.94% of participants violated GARP in some way at least once[10] during the experiment, and 55.94% of participants violated asymmetry (therefore, any participant who violated GARP also must have violated asymmetry at some point).

And as seen previously with donation patterns, we can also see a possible effect of MFQ type scores on their tendency to violate GARP by violating preference asymmetry (see Figures 3.13 and 3.14).

[10]Counting individual violations with respect to chains of revealed preferences is less straightforward because it is not clear whether overlapping preference chains would count as separate instances of violations; therefore, we are only looking at aggregates.

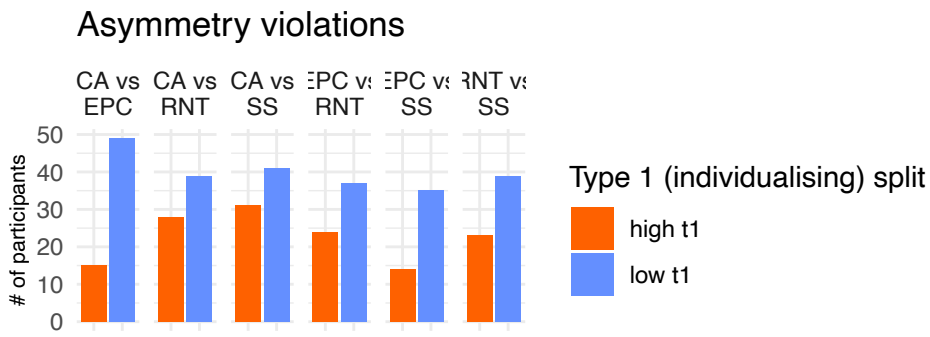


Figure 3.13: Number of participants with GARP asymmetry violations by charity pair and individualising-type median split. The largest difference can be seen on donations to Christian Aid vs Elim Pentecostal Church, where participants with high individualising scores exhibited fewer preference reversals.

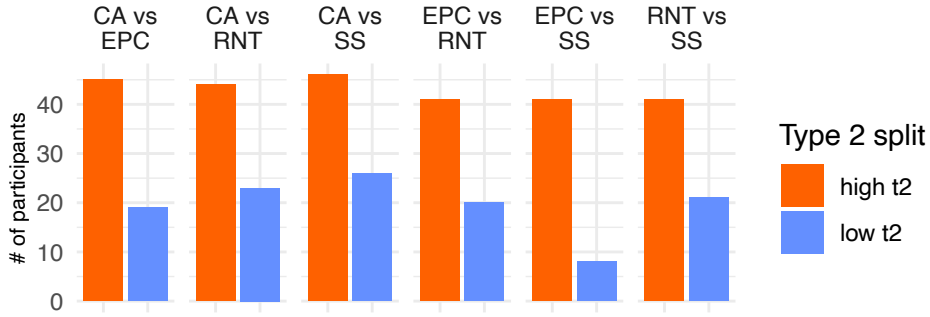


Figure 3.14: Number of participants with GARP asymmetry violations by charity pair and type 2 median split. The largest difference between the two groups was trials involving donations to Elim Pentecostal Church vs Sightsavers, with people with high type 2 scores showing more revealed preference reversals.

Other ways to violate GARP would be by violating transitivity, that is, choosing c_i over c_j, c_j over c_k, and not choosing c_i over c_k; or choosing within a similar preference chain with four elements (more than four are not possible here because we only have four distinct[11]

[11]Since the second clause in GARP does not speak of *strict* indirectly revealed preferences, instances involving the allocation of equal amounts to both charities are not considered to be in violation of GARP.

choice options). Among our participants, 35.25% exhibited transitivity violations; and 32.18% violated GARP using a chain of length 4.

3.6 Discussion

3.6.0.1 Confirmatory analysis

We did not find any differences in donation patterns between fixed cost and no-cost conditions. In particular, people's tendency to 'mix' donations did not decrease even when that effectively meant their donation budget shrunk by 10%. A relative cost threshold after which people are more inclined to 'max' might still exist when the cost gets high enough. If that cost becomes so high that the amount donated is negligible in comparison, the experimental task would lose its applicability to overhead costs in charitable giving.

The extent of overhead cost aversion people exhibit (or lack thereof) might be sensitive to the source of donated funds. Another possibility is that cost aversion is sensitive to the absolute size of the cost involved. For very large donations, a realistic overhead cost would likely constitute a smaller percentage of the amount than in our experiment; but because of the larger total amount of money involved, that cost would still be larger in absolute terms. A potential follow-up experiment might therefore investigate individual charity donations with a higher amount.

By letting participants allocate funds to different charities without the possibility to reward themselves instead, we were aiming to exclude effects of selfishness versus altruism to the largest achievable extent by design. An effect of this might be that overhead aversion may be less of an issue in a multiple-option allocation task: Bazerman, Loewenstein and White (1992) observed preference reversals when comparing a scenario where the desirability of an alternative is rated, compared to a setting in which different outcomes are compared.

In real-world charitable giving, donors may not be thinking in terms of choosing between alternatives.

3.6.0.2 Measures of rationality

We observed that participants behave "irrationally" in the sense of choosing to sacrifice a proportion of their donation in order to be able to split their donation, rather than allocating it to one of the charities (see Figure 3.8).

One resulting question is: under which conditions the behaviour we observed would be adaptive?

One possibility is that the integration of moral values occurs in some non-linear, perhaps hierarchical way. In this case, thinking of the moral domain as a multidimensional Cartesian coordinate system, in the way Moral Foundations Theory suggests, might be misleading. Potential further research could look into mapping the moral domain by translating results from behavioural research into frameworks utilising different well-understood mathematical structures, for example topology (Capraro & Perc, 2021) (with potential links to game theory) or dynamic nonlinear systems (with potential links to neuroscience methods (Crockett et al., 2015)). Alternatively, if we *do* continue to think of morality as such a multidimensional linear space, it is still conceivable that the evaluation function of an outcome's moral desirability is not simply the distance to origin.

Perhaps the preference towards distributed actions is an adaptation to underlying uncertainty. This uncertainty might be empirical (the more you know about the different choice options and/or the resulting impact of your action, the less you are inclined to do a bit of everything) but it could also be normative. If you cannot know for sure to which extent the pursuit of art or the reduction of suffering are morally relevant, it makes sense to support both.

This links to portfolio optimisation theory. MacAskill (2014) argues that normative uncertainty should be treated the same way as empirical uncertainty by maximising 'expected choice-worthiness.' However, in analogy to stock portfolio diversification, spreading one's responses rather than betting on one cause area and charity may be a preferable way of doing this; although it increases the risk of making a single morally suboptimal choice, it reduces the risk of making only morally bad choices and having to declare the moral equivalent of bankruptcy.

Another possibility is that paying attention to overhead is a heuristic to reduce the space of available options; perhaps, overhead aversion comes from an inclination to 'mix' rather than 'max,' which requires a narrowing-down of the space of charities one might want to give to.

One possible reason why we see this difference in maxing and mixing bases on people's binding-type score might be that when one option has moral relevance and the other does not, people might not even consider the morally irrelevant choice as a truly available option (Kouchaki, Smith & Savani, 2018). Perhaps we see this with donations to the church more than with donations to the theatre because people who score highly in their binding type are as likely to support the arts as those who do not. One way to depict this type of relationship would be to use a hierarchical model which, in a first step, prunes the set of options down to equally morally relevant choices, and then, in a second step, integrates the agent's values to pick among those options. Perhaps people are choosing to support multiple charities to minimise their expected regret upon learning new information in the future.

Or, perhaps we are all accounting for the effect we saw with people's moral value (MFT) measurements being dynamic. With at least two probably-different systems at play, the outcome of a value integration process using outputs from those two systems might

likewise vary over time. Perhaps 'mixing' is allowing for respective fluctuations.

3.6.1 Exploration

3.6.1.1 MFQ and donation pattern

Does a high binding-type score make choices harder because there is more of an internal conflict, or does it make them easier because it reduces the subjectively experienced freedom to choose (Diaye & Urdanivia, 2009; Guzmán et al., 2022)? Apparently, compromise ('mix') choices take more time (see Section 3.5.2.2), and a higher binding score leads to more such choices (Section 3.4.3.2). This suggests that caring more about 'binding' moral values does not reduce computational complexity (a perceived choice between conflicting obligations could still plausibly reduce the perception of how free that choice really was—an inquiry in this direction could be peripherally interesting alongside further research).

3.6.1.2 The "Square"

The main problem with the approach that treats MFQ types as dimensions is that we know that people's morality type scores are, in fact, related.

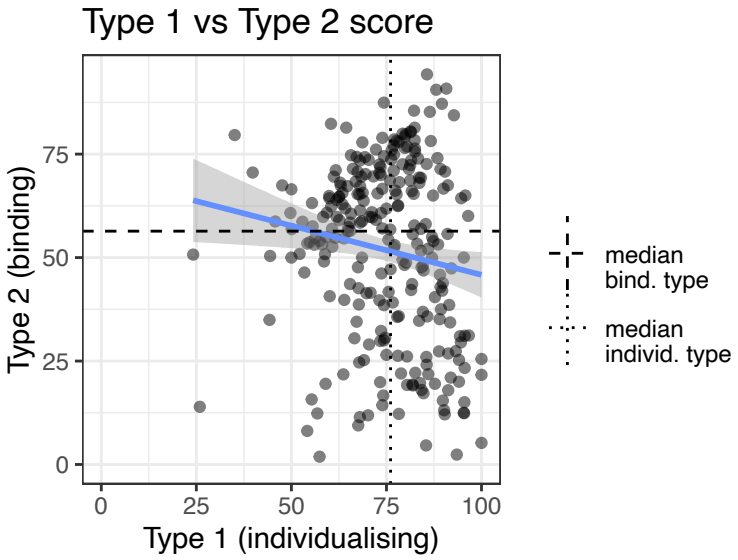


Figure 3.15: Relationship between participants' individualising-type vs binding-type scores

Individualising morality appears to be universally relevant—notice how there are few data points on the plot's left side in Figure 3.15—while binding morality matters to a degree which varies by individual. The extent to which binding morality matters is influenced by the other dimension, implying that these two "dimensions" are related to each other somehow. In fact, we find a weak negative correlation between participants' type1 and type2 scores (here, cor(t_1, t_2) = -0.15), an observation also found during the original MFQ validation (Graham et al., 2011).

In linear algebra terms, it looks like while the two types are not orthogonal to each other, they are still linearly independent. A true decomposition into orthogonal dimensions rather than spanning vectors would make for a more convenient basis to build mathematical models on, but it would come at the cost of losing interpretability. (How exactly should we interpret '$t_1 - .15 \cdot t_2$' in social psychology terms? Or is it '$t_2 + .15 \cdot t_1$' that is meaningful?) One possible workaround for this problem may require looking into which

widely used psychological measuring tools are compatible with each of these different possible decompositions.

3.6.1.3 Measures of rationality

We saw that participants who score high in binding type morality are more likely to split their donation. How often do subjects violate GARP in their preferences? How often do they do so in their choices? Could the latter be explained by the former (Diaye & Urdanivia, 2009)? And does it link to their morality?

Stated and revealed preferences:

If people's revealed preferences align with their stated preferences, on each trial, we would expect the larger donation to be to the charity which was rated higher.

We find that people deviate from their stated preferences in 16% of trials overall. (Unequal donations to equally-rated charities are not considered to be deviating.) At the same time, 67% of participants deviated in their revealed preferences from their stated preferences at least once (Figure 3.16).

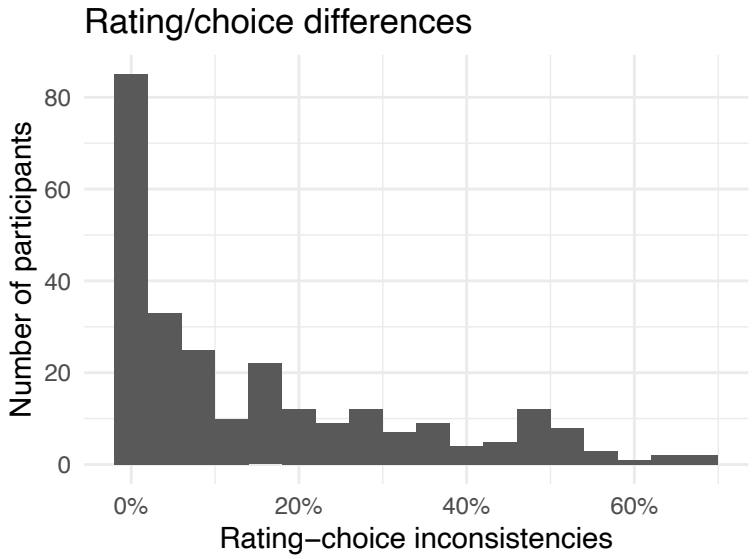


Figure 3.16: Difference between stated preferences and revealed preferences

GARP violations:

Guzmán et al. (2022) leads us to expect that we might see less GARP violations in choices that entail moral trade-offs.

As we saw in Section 3.5.2.5, the most common way to violate GARP was by violating asymmetry. GARP asymmetry violations were exhibited by more participants with below-median scores for the individualising moral type than with above-median scores, and by more participants with above-median scores for the binding type than with below-medium scores. If we understand adherence to GARP as a measure of rationality, we might be inclined to conclude that binding-type–morality makes people less rational, and individualising-type–morality makes people more rational.

But if we take into account that we believe morality to be multidimensional, not all charity pairs make for equally challenging choices when it comes to preference formation.

For people with high first-type (individualising) scores, the most difficult comparison is between Christian Aid and Sightsavers, with both charities being similar in *type 1-ness*; and conversely, the choice one might expect people scoring higher on second-type (binding) to struggle with the most would be between Elim Pentecostal Church, an organisation focusing on *type 2*-ness, and Sightsavers, a charity whose cause represents *type 1*-ness. Looking at Figures 3.13 and 3.14, these are the charity pairs that participants with above-median scores for the respective types violated GARP most frequently for.

While this is merely a post-hoc observation within an exploratory analysis and, as such, shall be taken with a grain of salt, it suggests that the multidimensional nature of morality may lead to differences in decision difficulty between agents with different sets of moral values.

3.6.1.4 Limitations

In real-life charitable giving, people usually donate their own money rather than allocating some budget provided to them. Perhaps, overhead costs play a bigger role if their own resources go towards covering them—Gneezy et al. (2014) found that covering overhead cost from a separate budget can reduce overhead cost aversion in donors. On the other hand, a potential reason for that decrease of overhead aversion might be a higher perceived personal impact from one's donation, and this perception might be different when we look at allocation to one versus two charities rather than at the total amount somebody donates. Since giving your own money rather than an amount provided to you in an experiment might produce different behaviour, this cannot be studied well in a laboratory/online setting alone. Potentially, a follow-up task could be done by an organisation that coordinates charitable giving ('meta-charity').

3.6.2 Conclusions for increasing the effectiveness of charitable giving

Trying it out as an intervention, Caviola, Schubert and Greene (2021) found that donation splitting can be used to increase amounts given to charities deemed effective by allowing to split donations between that charity, and a charity personally preferred by the donor: *'Many people perceive charity as a matter of personal choice, such that donations need not be guided by objective measures of effectiveness. For many, a donor choosing a charity is more like a gourmand choosing a restaurant than a doctor prescribing a treatment.'* In spite of the analogy being used as criticism here, thinking back to the morality-as-taste-metaphor, moral decisions might very well be subject-dependent and based on the subject's particular set of distinct moral 'taste buds'.

The idea of 'cause neutrality'—that we should not prefer one cause over another for reasons other than it doing the 'most good'—is a central assertion in Effective Altruism (MacAskill, 2019a). But the 'moral taste bud' analogy suggests that while cause neutrality may be compatible with some sets of moral values, it may be an invalid assumption for others; it is only feasible when optimising along one dimension at a time, or when comparing options that are similar in the extent to which they address each relevant dimension. To accommodate for moral diversity, we need effective charities that address other causes that people care about, albeit this may include cause areas that may be harder to quantify, or ones that cannot be addressed as effectively. For example, EA's focus on poverty interventions in poor countries ignores potential differences in people's values when it comes to ingroup preference. Poverty interventions in the donor's region may be less cost-effective, but would agree with what they care about.

As of early 2023, the four recommended charities by GiveWell are all aimed at addressing

global health causes; two organisations are aiming to fight malaria by providing medicine or mosquito nets respectively, one organisation fighting vitamin A deficiency, and one organisation attempting to increase childhood vaccination rates using cash incentives—for an estimated average cost-effectiveness between $3500 and $5500 dollars per life saved. Rather than incentivising fundraising for a small selection of charities, providing effectiveness information on a wider range of causes, interventions aimed at improving effectiveness in the charity sector in general, may make the idea of effective charitable giving appealing to people with more diverse values.

Towards Moral and Trustworthy AI

> *In arguing that machines think, we are in the same fix as Darwin when he argued that man shares common ancestors with monkeys, or Galileo when he argued that the Earth spins on its axis. [...] The computers, speaking for themselves (figuratively and literally), will in time convince all but the most hardened fundamentalists that they think.*
>
> — Herbert Simon, Models of My Life

> *Roughly speaking, [machine learning models] take huge amounts of data, search for patterns in it and become increasingly proficient at generating statistically probable outputs — such as seemingly humanlike language and thought. [...] But ChatGPT and similar programs are, by design, [...] incapable of distinguishing the possible from the impossible. [...] They trade merely in probabilities that change over time. For this reason, the predictions of machine learning systems will always be superficial and dubious.*
>
> — Chomsky, Roberts and Watumull (2023)

4.1 Introduction

With the continuing advancement of artificial AI, the question arises of how it can be ensured that artificially intelligent systems are safe for humans. One aspect of AI safety is the question of AI value alignment: How can it be ensured that an artificial agent's actions

and objectives are compatible with human values? The aim of this chapter is to propose a research direction that addresses the value alignment problem from moral psychology's point of view.

The chapter's first part (Sections 4.2–4.4) links current questions in value alignment to topics in psychology, including the results of the previous chapters, and provides an outline of research to be done. The second part (Section 4.5) introduces an empirical framework tailored to address one of these central research questions—'How do human agents infer the values of other agents, human or artificial?'—and concludes by presenting a specific experimental design for an interactive online game.

4.2 Values, noise, and other motifs

When we observe people's behaviour, we cannot directly observe the reasons for any given action. A person offered a glass of wine might decline because they suspect it to be poisoned. Another person may say no because they are trying to stay sober. A third person might simply not like the taste of wine. Therefore, when attempting to learn an agent's values, it is difficult to distinguish what they know from what they value.

Uncertainty in the human's beliefs about the world may lead to variance, which could resemble a change in their (moral) utility functions,[1] and make it difficult to tell them apart. Armstrong and Mindermann (2018) argue that it is impossible to distinguish an irrational agent's reward function from their planning process, and that 'indeed, any reward is possible', implying that it is likewise fundamentally impossible to attribute underlying values to people's behaviour. Yet, humans casually figure out other people's intentions and thus, to some extent, their reward functions in day-to-day life (the tendency to do

[1] Any function describing people's moral values is probably multidimensional (Graham et al., 2009) as we have seen in Chapter 2.

that is so strong that we even attribute intentions to triangles on a screen). If the 'reward functions' of others were entirely arbitrary, no approximate inference could be made whatsoever, and intention attributions by humans would likely not coincide. The fact that there generally appears to be overlap in people's intentionality judgements, combined with the observation that the ability of accurate intention attribution is decreased in mental health conditions such as schizophrenia (Brunet, Sarfati & Hardy-Baylé, 2003; Sarfati, Hardy-Bayle, Nadel, Chevalier & Widlocher, 1997), suggests that there is some real-world information about the intentions of others being inferrered.

In Chapter 2, we found people's moral values appear to be fundamentally noisy, and confirmed that there is more than one relevant dimension to morality. In Chapter 3, we learned that the way competing values are integrated differs between people, not only with respect to what they choose to prefer, but also how they make that choice (and their preference towards optimising for more than one value).

We know that if we wanted to describe morality in dimensions, we would need to decompose the values we know of into separate, orthogonal components. We might want to use AI to aid us to do so.

We know that if we want to build AI that learns how human values work, we need to make sure we construct the training set in a way that allows a system to learn the aspects we know about. For example, we could train a language model on a text corpus, but if that text corpus does not include time stamps or specific links to people, the model won't be able to recognise any temporal dynamics. (For well-known phenomena, a model might be able to learn from descriptions of said phenomena; this doesn't work if the descriptions are not there yet.)

And ultimately, we need to be aware that different people have different sets of moral

values, intuitions, and rules to which they conform; and these do not only yield different choice outcomes in terms of what people prefer, but also how they make those choices, as we have seen in the previous chapter. Whenever we speak of 'value alignment of AI' without specifying whose values we are referring to, we are being too unspecific. We need to get into the habit of thinking along the lines of 'Can we prove that this algorithm is aligned to the value sets $V_1, \ldots, V_k$?', not merely 'Can we show this algorithm is aligned to human values?'.

When it comes to morality, academics are not a representative sample of the global population; on average, we tend to be more open-minded, perhaps because that's our job. But we need to be careful not to assume that this makes our moral values more correct than those of others, and take steps not to attribute differences we observe to cognitive limitations. At the same time, we also don't want to copy or emulate human-typical errors of moral reasoning within systems with otherwise superior-to-human abilities. In our experiments in Chapter 3, one possible hypothesis was that type 2-founded reasoning would help simplify moral decisions; but this was not what we saw in reaction times.

Adding to this, the distinction between the two types probably does not paint the whole picture; while the different MFT Foundations show high correlation, they still differ in content, and it is not a perfect correspondence. For example, purity and authority may have contrary influences on choosing not to eat meat (De Backer & Hudders, 2015; Rozin, Markwith & Stoess, 1997). There is work to be done to disentangle those into something more orthogonal by considering scenarios to differentiate between them (looking at purity/authority in veganism and authority/loyalty in scientific publishing may be interesting leads). And as the authors of MFT themselves have said, the list of known foundations is not an exhaustive depiction of reality; for instance, a later version of the MFQ than the one we used in Chapters 2 and 3 has added the foundation of liberty—and there are

possibly other dimensions that play less prominent roles in moral judgement and that are not yet explicitly accounted for.

4.2.1 Dual-process morality

As we know from the previous chapters, there have been multiple approaches that describe how people make moral decisions in terms of two different systems or processes, trying to find a correspondence: Emotion versus reason (Haidt, 2001), binding versus individualising moral foundations (Haidt & Graham, 2007), model-free versus model-based evaluation (Crockett, 2013), fast versus slow response times (Suter & Hertwig, 2011), action-based versus outcome-based value representations (Cushman, 2013), and two different patterns of activity in the brain (Dolan & Dayan, 2013). These distinctions clearly appear to be related in multiple ways; but in each of these cases, the dualities do not seem to neatly map onto each other in a one-to-one correspondence.

4.2.2 Uses of noise

We have also seen that people's expressions of their moral values seem to be noisy and that there are probably at least two separate processes involved. This separation between the two clusters of moral foundations does not appear to be the same duality as in dual systems theory, and also not the distinction between utilitarianism and deontology. To determine if noise is an artifact of human cognitive architecture rather than a functional component, we might want to look into the purpose of noise in the context of morality. If noise turns out to be an undesirable constraint of human thinking, similar to probability matching behaviour in gambling (Shanks, Tunney & McCarthy, 2002), we might not want to repeat this aspect in a system of our own design. However, noise might also represent an adaptive response to things like normative uncertainty—when we cannot be sure what

we ought to do, even after gathering as much information about the world as we can—or model uncertainty, in which we are unsure which model is appropriate to apply to a given problem to figure out what we are going to do. So perhaps noise is a component we want to include in to any representation of moral reasoning artificial agents should have; but perhaps not, in which case we also need to find a different response to adapt to normative uncertainty.

A system that is consistent but consistently does the wrong thing seems like a clearly undesirable scenario. If we build a system that operates under continuous human oversight—such that a human is overseeing every decision that is made and has the option to 'veto' and intervene—this oversight may be sufficient to introduce an appropriate level of noise, thus removing the need to include it. This approach seems difficult to implement because computers will be able to make decisions on timescales that we cannot compete with. Another question we have to answer is how rule-based inference which is typical for deontology comes to be. Could it be a sampling-based system in disguise, potentially with similarity-based judgements?

4.3 AI alignment

Today's AI systems are optimisers with specific, albeit occasionally implicit, loss functions. Even the astonishing developments in language processing and image generation in recent years are striking examples of systems that solve a specific problem. If we do not take into account how morality works cognitively in humans and fail to design systems in a way that allows for accommodating 'moral' decision-making, the systems involved may lack the capability to be adjusted in the ways we need.

One such aspect of morality is fairness. AI systems have been shown to exhibit biases

that make them unfair (Buolamwini & Gebru, 2018); this can occur as an undesired side effect of the training set used (Sambasivan et al., 2021), but may also come from the way 'fairness' is defined (Zehlike et al., 2017). But there are more aspects to morality than fairness; it is possible to be fair and still fail to do good,[2] we need a representation of other values.

Trust is often cited as a desired quality of a "safe" AI system. But what does trust require? One way trust has been framed is that a person or an organisation can be sure that the AI system is acting in a way deemed "good" by the person/organisation. This could mean that the AI system does the same things the individual/organisation would *do*, or it might mean that it does the things they would *like*. Either way, the AI system needs to optimise along the evaluation dimensions which are relevant to the human(s) in question. These dimensions might be universal, such as fairness; or they might be culture-specific/person-dependent, such as loyalty or authority. But for the system to optimise along a set of dimensions, these dimensions must be accounted for in the system.

A commonly presented argument is that as long as AI agents act in ways specified by us, they will not need a representation of human values because they are not moral agents and, by definition, will not able to make moral decisions anyway. This is debated: While some experts claim it to be impossible for AI to have moral agency (Martinho, Poulsen, Kroesen & Chorus, 2021), others argue that constructing artificial moral agents is not only possible, but desirable (Formosa & Ryan, 2021). And people in experiments appear to attribute moral agency to AI agents (Shank & DeSanti, 2018), albeit perhaps to a weaker degree than to human agents (Gamez, Shank, Arnold & North, 2020).

To work towards value-aligned AI, we need to learn more about how human values work,

[2]Imagine a trolley-type scenario in which the fat man is pushed onto the tracks after the trolley has already passed, killing the man as well as the workers. Arguably, killing everybody would achieve distributive fairness, but at the cost of everyone involved being dead.

and keep our knowledge about this in mind when we design AI systems with the goal of aligning them to our interests. We cannot, as of now, defer this task to black-box AI algorithms because even if we had a provably reliable and trustworthy AI system, we don't have good tasks to use for value alignment yet; moral dilemmas are nice for illustrative purposes, but not great for actual value alignment (LaCroix, 2022). The usefulness of learning from developmental psychology is probably limited, given current AI systems don't acquire their cognitive 'skills' the same way humans do, either: A baby does not learn to speak by systematically processing large amounts of text. Likewise, noticing misalignment in an AI is likely to differ fundamentally from watching children and pointing out mistakes. We likely need a new approach.

Implications for human-in-the-loop approaches

The current idea of AI control built into upcoming legislation places central importance on human oversight, following a "human-in-the-loop" approach (Floridi, 2019). This alone is insufficient in practice, since AI systems are able to exert control on a different timescale: AI systems are acting within time spans or with amounts of data which don't allow for direct, full human oversight during the system's operation (e.g. self-driving cars, medical diagnostics), providing a further argument that the AI system in question should have a model of what the human would approve of.

If we are going to pursue human-in-the-loop approaches during the operation of AI systems, we need to find monitoring solutions that will continue to work for systems beyond human-level ability (Bowman et al., 2022), and which are also constructed in a way that is informed by what people are good at noticing.

4.4 Research to be done

> *Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks.*
>
> — Hawking, Russell, Tegmark and Wilczek (2014)

We need the systems we build to act in accordance with our values, but not (all) our limitations. The current approach to AI alignment is learning from human feedback. For this, we need to learn how we evaluate intentions/behaviour and give feedback. Humans will likely need to assign trust differently when it comes to AI-generated output. Our heuristics how we infer values from actions are likely maladaptive when it comes to machine-picked actions (because other heuristics do not translate well either; see Section 4.4.2). We need to find out more about how people do this.

One problem we are now aware of is that a model with a training set that's human-generated that it is going to generate human-like judgments while exhibit similar biases as humans, as we can already observe in existing systems (Bolukbasi, Chang, Zou, Saligrama & Kalai, 2016; Johnson et al., 2022; Schramowski, Turan, Andersen, Rothkopf & Kersting, 2022), although we do not yet know whether these biases are produced in a similar way (see Section 4.4.2). There may also be incentives to defer judgement to an AI (Feier, Gogoll & Uhl, 2022); people seem to behave in ways we consider less moral when interacting with machines (Giroux, Kim, Lee & Park, 2022). This suggests that in interactions 'assisted' by an AI which may be biased in familiar ways, the tandem of AI and human may be more likely to produce flawed and/or less moral decisions than a human agent would, and than an AI agent should. Further, Malle, Scheutz, Arnold, Voiklis and Cusimano (2015) found people tend to approve of artificial agents acting in accordance with utilitarianism, compared to how they judge humans. This tendency, too, may depend on the individuals'

morality.

Whether large language models (LLMs) internally create a representation of dual-systems processes, or merely exhibit similar judgements which it produces merely by similarity-based reasoning, is an open question. It is, however, clear that they do not develop the same internal representations humans have, leading some researchers to conclude a general lack of such internal structure (Chomsky et al., 2023). But some functionality in human brains, too, can be described in terms of mere 'probabilities that change over time' (Friston et al., 1994); so at this point, we likely cannot tell.

For topics other than this one, we might want to apply AI methods somewhat blindly to obtain answers, as they can assist in the detection of 'unknown unknowns' (Wilson & Daugherty, 2018). But here, this would mean using systems whose functionality is not well understood to extract information we want to use to make these very systems safe. The goal cannot be to have an AI 'summarise and optimise morality', but neither can it be to build AI systems with optimised decision-making abilities in any other area but only equipped with a reproduction of human moral judgements, errors and all.

Current systems do not quite exhibit superhuman abilities in every other domain. For example, ChatGPT does not seem to pass the Turing test. After giving it a try, Floridi and Chiriatti (2020) write:

> *We expand the analysis to present three tests based on mathematical, semantic (that is, the Turing Test), and ethical questions and show that GPT-3 is not designed to pass any of them. This is a reminder that GPT-3 does not do what it is not supposed to do, and that any interpretation of GPT-3 as the beginning of the emergence of a general form of artificial intelligence is merely uninformed science fiction.*

Still, conversational AI, the type of AI the Turing test is designed to address, is only one kind of AI; and LLM-based versions of conversational AI constitute only one possible way to implement that kind. Tasks such as driving a vehicle, performing surgery, steering financial decisions require very different adaptive systems which are also improving and need to be explainable and value-aligned. Not every AI system is being designed to have the ability to communicate about its own actions in natural language. But every AI system needs to be value-aligned for it to be safe.

4.4.1 Human agents

We need to continue to study how people assign trust to other agents, human and artificial.

Today's approaches to AI alignment and control overwhelmingly rely on integrating human feedback (Christiano et al., 2017; Hadfield-Menell, Russell, Abbeel & Dragan, 2016; Ouyang et al., 2022). This relies on human ability to provide accurate feedback. But when only evaluating output behaviour without any explicit or even intuitive access to underlying values, people might make wrong judgements.

We need to improve our understanding how people, in practice, distinguish between noise and value misalignment. Both moral learning as well as intention inference (Diaconescu et al., 2014) can be described as hierarchical Bayesian learning, involving volatility estimates of others' intentions. And it appears that this is easier to do for people are like us—we are able to gauge intentions of those whose *values* are similar to our own more accurately (Barnby, Raihani & Dayan, 2022). This suggests we might be less able to do that for artificial agents with preferences that behave very differently from ours, including over time. It is not immediately clear whether this similarity effect can be explained because of a need to understand individuals to infer their values accurately, and we have more insight into our own mental phenomena such as intentions than those of other people. Or

simply because of that preference simularity but for some other reason than understanding (and we persistently let our own preferences cloud judgements of preferences of others, beyond merely using them as a starting point (Tarantola, Kumaran, Dayan & De Martino, 2017)). Are we also better at distinguishing between, say, noise, missing information, and misalignment of values for individuals with similar values?
Intention attribution includes intentionality judgements: How do we know whether someone who acted in a harmful manner towards us did so on purpose? Or did they in fact act to the best of their ability and failed to avoid causing harm? Do we do this differently for artificial agents?

Progress on these questions will also give us a better understanding of how people's formation of trust could be exploitable by AI, and therefore potential areas of vulnerability. Lacking a shared evolutionary history, the way people form trust is likely to be maladaptive to interactions with artificial agents, particularly those that appear human-like.[3] The things that would make us trust an AI are probably not the same things that would make an AI safe.

4.4.2 Artificial agents

We need to keep studying artificial agents to find out how we can tell whether our trust is justified.

David Marr's 'levels of analysis' framework, first described in his book Vision (Marr, 1982), constitutes a popular set of different viewpoints in computational cognitive science. It proposes the distinction between the computational level (which problem is being solved?), the algorithmic level (which algorithm is used to solve it?), and the implementational

[3] In terms of behaviour or dialogue, not appearance; human-like shape might not have a clear effect either way (Atchley et al., 2022).

level (how does it run on the physical hardware?). Yet, in the task of interpreting AI systems, this distinction seems insufficient. We comprehend the high-level problem (providing a response to a prompt[4]), we understand the algorithms involved in building ChatGPT (Radford, Narasimhan, Salimans, Sutskever et al., 2018), and we know which hardware can be used to run it on (Shoeybi et al., 2019). But for the public release of ChatGPT, the algorithm differs from previous versions mainly in the amount of data and computing power. The output, however, is perceived as something fundamentally different (Bommasani et al., 2021), and its overall quality surprisingly difficult to assess: Heuristics such as use of language (GPT-3 can produce results that are well-phrased, adapted to a particular style, and, at the same time, thoroughly factually wrong (van Dis, Bollen, Zuidema, van Rooij & Bockting, 2023)), which we typically use to infer social status (Piff, Kraus & Keltner, 2018) or social group membership and, by proxy, trustworthiness (Aidenberger, Rauhut & Rössel, 2020) do not work here.

We lack understanding on the level on which this fundamental difference would be visible. We don't have a theory of 'mind' for AI systems; we are only aware of some failure modes we happened to come across, some of which are remarkably unhumanlike (adversarial samples (Elsayed et al., 2018), for instance, where any arbitrary picture can be very slightly tweaked in ways that are invisible to the human eye but reliably make an image recognition algorithm recognise it as, say, a picture of a dog). Yet, using methods from psychology research on current state-of-the-art AI systems directly in order to obtain this understanding fails—they respond to slight changes to the input in unexpected ways (Binz & Schulz, 2023), making it seem likely that the internal cognitive architecture used to produce these results substantially differs from that of humans (Shiffrin & Mitchell, 2023). Perhaps we need to start applying the foundations of behavioural research to sufficiently

[4]It is designed to 'follow an instruction in a prompt and provide a detailed response' (`https://openai.com/blog/chatgpt`)

advanced AI models. It would be conceivable, say, to query an uncensored version[5] of ChatGPT for the reasons it gives for particular bits of output, and then look under the hood and see in which cases you can discover patterns in its weights that correspond to the explanation given.

Such an approach will need to look into, and account for, AI-specific failure modes or deviations from what we would expect or what we would consider optimal in terms of decision making and behaviour. For this, we need to know which failure modes are specific to which kinds of AI system are used (not everything we call AI will exhibit the same properties); but we need to know which human biases are exhibited by AI systems as well—a line of work that is already ongoing (Rich & Gureckis, 2019)—but we also need to learn which properties of human reasoning we are used to are *not*. (For instance, unless computer hardware undergoes some significant changes by then, AI-based choices will likely not be influenced by hunger, as opposed to human decision-making (De Ridder, Kroese, Adriaanse & Evers, 2014).)

We need a collective effort by anthropologists, psychologists, social scientists who come up with qualitative abstractions of the ways these systems produce reasoning and that are accessible for humans in terms of understanding.

In this pursuit, we will also need to try to avoid failures to see AI systems demonstrating the ability to perform cognitive tasks we believed to be unique to humans; eventually, morality might be one of these. One widespread opinion is that LLMs are merely parroting their training sets (Bender, Gebru, McMillan-Major & Shmitchell, 2021) and are incapable of true insight and understanding.[6] How will we be able to tell if an AI agent is *truly* showing

[5]The version of ChatGPT currently open to the public responds to people digging too far with a standardised answer: 'As a large language model trained by OpenAI ...'

[6]However, if we compare the size of an LLM to the size it would take to store hardcoded lookup tables that produce the same output, such tables would be far, far larger than the model—because the model is

insight? By what means can we determine if it *really* capable of abstract reasoning?

4.4.3 Policy and education

Work on advancing value-aligned AI technology needs to take into account what we already know, and continue to learn, in social psychology (T. Miller, 2019). It would be worthwhile to prioritise alignment and interpretability over capabilities research because otherwise, we are increasing the power of something we do not understand. However, this might very well constitute an economic disadvantage to any companies that voluntarily choose to slow down capability improvements in favour of safety and explainability, which highlights the need to develop policy and regulation. On a related note, people act differently when interacting with artificial or human agents. For example, we seem to attribute blame differently to an AI deciding to launch a lethal strike compared to a person (Malle, Magar & Scheutz, 2019). This suggests that analogously to GDPR for data regulation, there may be a need for a requirement to make any use of AI in predefined contexts transparent.

We also need to make sure people are sufficiently familiar with AI technology to avoid being fooled by it. ChatGPT might not pass the Turing test today because people generally have experience with output produced by language models–something everyone who types a message on a modern smartphone encounters a version of–but would it also fail the Turing test conducted with Alan Turing's contemporaries? In the field of HR, Gonzalez et al. (2022) found that the extent to which human vs AI-based candidate selection processes differs in their effect in hiring depends on the person's familiarity with AI technology. We do not need to wait for more advanced models to start developing 'AI literacy' curricula

designed to dismiss irrelevant bits of information, making it capable of abstraction compared to raw data in the literal sense of the word (the Latin 'abstrahere' means 'to draw away' from something).

for the models used today.

4.5 Proposed experiment

A better understanding of intention attribution in humans might give us new ideas for which similar functionality we, attempting to build human-compatible AI agents, could want AI systems to have. But mainly, it will provide an understanding of the difference in intention attribution between interactions with another person versus interactions with an artificial agent, allowing to evaluate the degree and also to which familiarity with AI technology affects the nature and accuracy of such judgements, which will help to determine requirements in terms of 'AI literacy' to ensure a sufficient sensitivity of human oversight.

4.5.1 Motivation

When attempting to learn an agent's values, it is difficult to distinguish what they know from what they value. Uncertainty in the human's beliefs about the world may resemble a change in their (moral) utility function.[7]

One framework aimed at learning an agent's reward function was proposed by S. Russell (1998) and described by Choi and Kim (2011) as follows:

> Given (1) measurements of an agent's behavior over time, in a variety of circumstances, (2) measurements of the sensory inputs to the agent, (3) a model of the physical environment (including the agent's body), determine the reward function that the agent is optimizing.

[7] Any function describing people's values is probably multidimensional (Graham et al., 2009).

In practice, if we want an AI agent to learn the values of humans, (2) would be very difficult to accomplish. Likewise, rewarding the robot agent by using the human's reward function (Hadfield-Menell et al., 2016) requires the human to have access to its reward function in the first place, which, taking into account all of moral psychology, is not a problem that has a closed solution for big-picture value alignment at the current point in time.

Kleiman-Weiner, Saxe and Tenenbaum (2017) propose a minimal set of cognitive abilities necessary specifically for moral learning: (1) a utility calculus that allows integrating other agents' utility functions into one's decisions, (2) hierarchical Bayesian inference to infer the weights of other agents, and (3) learning by value alignment, to values of others or to one's own attachments/feelings.

When the AI can only observe an agent's actions (and possibly get feedback in the form encouragement or discouragement), it is impossible to distinguish uncertainty in the human's beliefs about the world from the shape of the human's utility function. By creating a task that allows us to manipulate uncertainty, we can gradually isolate the question of what the other agent values from what the other agent knows.

4.5.2 Questions

The central question this task is aimed at is: How do humans distinguish between other people's value's incompatibility with one's own ('evil'), principal errors in how they model the world and/or limited computational power ('stupid'), and the specific setting along with any stochastic uncertainty inherent to it ('unlucky')?

It would be interesting to look at the extent to which our own level of (in)competence affects the likelihood of categorising others stupid rather than evil. Further, since for

people who care about binding moral foundations as much as about individualising foundations, there are more possibilities for something to be 'evil', this might play a role in intention attribution simply because more actions could have perceived moral valence. In this case, an increase in group-level morality scores would lead to an increase in the proportion of 'evil' over 'stupidity' judgements. Additionally, utilitarianism[8] may imply a stronger link between 'evil' and 'stupid' because between two individuals with the same set of intentions, a lack of understanding or ability would lead to less helpful outcomes of actions compared to the other person. Does volatility of a person's actions play a role in intention attribution? It is conceivable that volatility—occasional but inconsistent helpful behaviour—might signal 'stupid' rather than 'evil' simply because with 'evil' values, in one-shot games, being consistently unhelpful would yield a higher reward but might require a higher level of skill to achieve. Finally, does programming experience influence how evil people think AI agents can be?

This links to the idea of distinguishing 'evilness' and 'stupidity' of human and AI agents. In a simple, e.g. Wolfpack-like[9] spatial cooperation task, what influences people's degree of belief in one or the other?

4.5.3 Experimental design

People are known to act differently than they say they would. The choices they make in hypothetical scenarios are different from choices in real scenarios (Nisbett & Wilson, 1977; Vlaev, 2012).[10]

Since people have been shown to attribute human-like concepts such as intentions to *triangles* (Heider & Simmel, 1944), there is not much need for an elaborate life-like setting.

[8] To measure utilitarianism, the Oxford Utilitarianism Scale (Kahane et al., 2018) could be used.

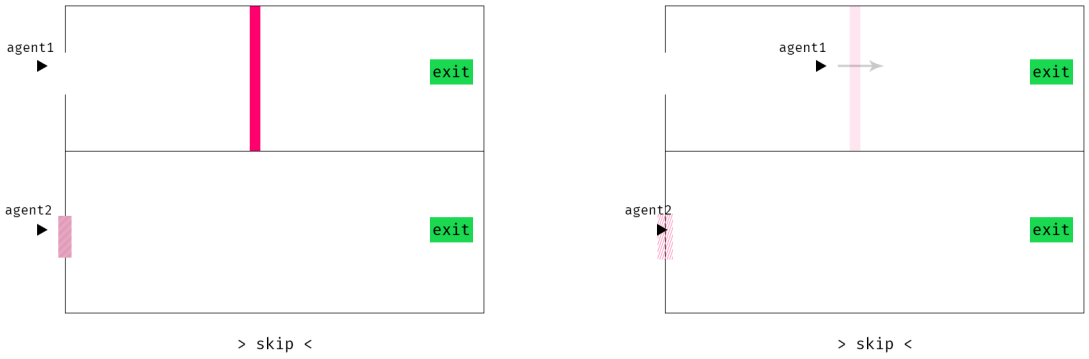
[9] Leibo, Zambaldi, Lanctot, Marecki and Graepel (2017)

[10] See also: Asch's conformity experiment, bystander effects, experiments incentivising dishonesty, ...

We can use either a spatial cooperation task or a limited resources task (such as the Gathering game by Leibo et al. (2017)) to test cooperation; since resource-based games have been extensively investigated by game theory, let us begin with spatial cooperation for higher marginal information returns.

I propose a simple spatial cooperation task illustrated below. Each agent is represented by a triangle. The goal is to complete the level by touching the exit button. The maze can have a different shape every time, resembling a two-dimensional version of a reinforcement learning maze (Beattie et al., 2016; Dayan & Hinton, 1992). There are barriers in different colours which block a portion of the game area, that can be overcome by an agent touching the switch in the same respective colour. Points are awarded for completed levels.

- Each level starts with both players placed in specific parts in the game area.
- A level is completed by being in the exit area and pressing "n" to go to the next level. Only players in the exit area get points for completing the level.



(a) Agent 1 faces a barrier they cannot pass.

(b) Agent 2 touches the switch, and Agent 1 now can pass through the barrier.

Figure 4.1: Game design. Each agent controls a triangle that moves within a designated area. An agent can interact with another agent using in-game objects (barriers, deactivating switches) which provide specific opportunities for valenced (helpful, unhelpful) interaction.

This setting allows us to manipulate many aspects of a value attribution decision in a minimalist environment:

- Variation of incentive structure: Cooperation can be required, beneficial to both, beneficial for one agent and neutral for the other, or beneficial for one agent and costly for the other.
- Variation of complexity: For example, number of instances of cooperation within one task.
- Uncertainty manipulation: This can be achieved through randomly occurring handicaps for a player that are hidden to the other player, or through varying the amount of information that is given: In-game mechanics can either work reliably, or include stochastic failures; temporary information handicaps could be implemented by hiding a part of the canvas from a player.

- Variation of difficulty: The game can be made harder to play either by increasing the difficulty of controlling the triangle (e.g. adjusting gravity or acceleration), and the possibility of introducing additional in-game mechanics.

4.5.4 Hypotheses

One possible hypothesis is the requirement of value alignment for moral learning to be possible. Within the game, this would present as the pattern that people are more likely to infer someone's intentions correctly if they gained experience playing the other side of that level beforehand. Another possible hypothesis is that there is either a 'moral budget' such that cooperation becomes less likely after a certain amount of it, or that there is 'moral oscillation': After thinking somebody is stupid rather than evil (plausible deniability), in the next round, a person is *more* likely to judge others as evil rather than stupid. It seems plausible that programming experience may influence to which degree a person thinks that AI agents can be evil: The more programming experience ('AI literacy'), the more likely are people to think the program is stupid. Moreover, we may be able to detect effects in response to volatility in the environment: When external difficulties are present ('unlucky' agents), the proportion of 'stupid' vs 'evil' judgements changes as well.

For screenshots and a description of a prototype of such a game, see Appendix C.

Conclusion

> *Yet, the best ending of all I have saved for last. The best ending is one that ends suddenly and without any explana–.*
>
> — J.P. Beaubien, Terrible Writing Advice, "ENDING A STORY"

Contributions and limitations

Chapter 2 provided insights on properties of MFQ scores over time, and thereby a new perspective on what the MFQ actually measures. We do not find evidence of 'moral licensing' or 'moral cleansing'-like effects when it comes to the underlying values themselves, which speaks against 'steering wheel' analogies for morality. Previously observed stability of moral identity or character over time may be explained by playing out on longer time spans than individual choices.

The results suggest that it is plausible to view moral values, as they appear in applied decision-making, as the result of sampling from a system that involves two separate, noisy processes.

In **Chapter 3**, we discovered that type 1 and type 2 morality seemed to have opposite effects on participants' tendency to split their donations. We were surprised to see that overhead cost did not have an observable effect on donation patterns. This suggests that people care more about the 'what' of their donations than about their impact, and also that how much they care about the 'what' depends on their moral values. We also found

that scoring higher on type 2 did not lead to quicker, and presumably easier, decisions.

In terms of drawbacks, it should be mentioned that contrary to initial hopes, it was not possible to construct geometry-inspired functions which would take people's MFQ scores as arguments and accurately simulate their donation behaviour. The integration process likely depends on other factors we were not measuring, and that process itself might also be noisy. Or perhaps the very notion of linear distance metric on the space of moral values is too much of a simplification. Another limitation is that we, again, used a monetary metric to quantify responses that are likely not produced in terms of reasoning about money. (That said, at least doing so was not out-of-context in the domain of charitable giving, as compared to, say, tasks that require inflicting harm for money.)

Chapter 4 outlined the need for research on humans (intention inference) and on artificial agents (making higher-level cognitive processes interpretable). It concluded with a proposal of an experimental framework in the shape of an interactive game that provides data on how people infer intentions of other human or artificial agents. In contrast to the questionnaire in Chapter 2, it is an interactive task that requires choices; and unlike the experiments in Chapter 3, it does not use money as a proxy metric. The next natural step will be the implementation of the task proposed at the very end of Chapter 4. The suggested framework allows to test intention attribution empirically, observing behaviour and not merely questionnaire responses, and is not centered around monetary rewards. It can also serve as a starting point for comparing interactions between two people to interactions between a person and an artificial agent, permitting the study of how people attribute intentions to humans vs AI in a sandbox environment.

Outlook

Today, there is a need for foundational research in the empirical behavioural study of AI. In contrast to the study of human decision-making, this is hardly possible within a sandbox environment because scaled-down prototypes of AI systems built using small data sets typically do not work very well, let alone exhibit emergent cognition-like properties. Due to this, we are only currently entering a time in which AI systems themselves can be meaningfully approached from the perspective of behavioural research. Artificial agents are starting to exhibit behaviours they were not designed to have, improve in some unexpected ways, and go wrong in yet others, and constitute a rare opportunity to observe and understand something truly novel. Simultaneously, this line of research is a necessity for AI safety—regardless of whether this can be achieved through ensuring artificial moral agents have human-compatible values, or by obtaining an understanding of human social psychology in interactions with artificial agents to ensure we can provide safe human oversight of AI systems, or in a way that is not yet foreseen.

If we do not start now, systems may continue to grow in complexity and/or size and make later approaches at understanding more difficult—with the ongoing potential of us being surprised by new abilities that, too, appear rather suddenly and without any explana–.

References

Aidenberger, A., Rauhut, H. & Rössel, J. (2020). Is participation in high-status culture a signal of trustworthiness? *PLoS One*, *15*(5), e0232674.

Alexander, L. & Moore, M. (2021). Deontological ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2021/entries/ethics-deontological/`.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, *100*(401), 464–477.

Andreoni, J. & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Aquino, K., Freeman, D., Reed II, A., Lim, V. K. & Felps, W. (2009). Testing a social-cognitive model of moral behavior: The interactive influence of situations and moral identity centrality. *Journal of personality and social psychology*, *97*(1), 123.

Aquino, K. & Reed II, A. (2002). The self-importance of moral identity. *Journal of personality and social psychology*, *83*(6), 1423.

Ariely, D., Bracha, A. & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*(1), 544–55.

Armstrong, S. & Mindermann, S. (2018). Occam's razor is insufficient to infer the preferences of irrational agents. *Advances in neural information processing systems*, *31*.

Atchley, A., Barr, H. M., O'Hear, E. H., Gray, C. E., Chesser, A. F., Jones, N. & Tenhundfeld,

N. L. (2022). Does my driver share my moral view? Effects of humanlikeness and morality in an adapted trolley problem. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 66, pp. 187–191).

Atran, S. & Ginges, J. (2012). Religious and sacred imperatives in human conflict. *Science*, *336*(6083), 855–857.

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

Barnby, J. M., Raihani, N. & Dayan, P. (2022). Knowing me, knowing you: Interpersonal similarity improves predictive accuracy and reduces attributions of harmful intent. *Cognition*, *225*, 105098.

Baron, J. & Szymanska, E. (2011). Heuristics and biases in charity. *The science of giving: Experimental approaches to the study of charity*, 215–235.

Bazerman, M. H., Loewenstein, G. F. & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative science quarterly*, 220–240.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., … others (2016). Deepmind lab. *arXiv preprint arXiv:1612.03801*.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 610–623).

Berman, J. Z., Barasch, A., Levine, E. E. & Small, D. A. (2018). Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological science*, *29*(5), 834–844.

Binz, M. & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the

literature. *Psychological bulletin*, *88*(1), 1.

Bloom, P. (2012). Moral nativism and moral psychology. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 71–89). American Psychological Association.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Botzer, N., Gu, S. & Weninger, T. (2022). Analysis of moral judgment on Reddit. *IEEE Transactions on Computational Social Systems*.

Bourget, D. & Chalmers, D. J. (2014). What do philosophers believe? *Philosophical studies*, *170*, 465–500.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., … others (2022). Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Briggs, R. A. (2019). Normative theories of rational choice: Expected utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/`.

Brunet, E., Sarfati, Y. & Hardy-Baylé, M.-C. (2003). Reasoning about physical causality and other's intentions in schizophrenia. *Cognitive neuropsychiatry*, *8*(2), 129–139.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … others (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Capraro, V. & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society interface*, *18*(175).

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J. & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives? *Judgment and decision making*, *9*(4), 303–315.

Caviola, L., Schubert, S. & Greene, J. D. (2021). The psychology of (in)effective altruism. *Trends in Cognitive Sciences*, *25*(7), 596–607.

Caviola, L., Schubert, S. & Nemirow, J. (2020). The many obstacles to effective giving. *Judgment and Decision Making*, *15*(2), 159.

Caviola, L., Schubert, S., Teperman, E., Moss, D., Greenberg, S. & Faber, N. S. (2020). Donors vastly underestimate differences in charities' effectiveness. *Judgment and Decision Making*, *15*(4), 509.

Chapman, C. M., Masser, B. M. & Louis, W. R. (2020). Identity motives in charitable giving: Explanations for charity preferences from a global donor survey. *Psychology & Marketing*, *37*(9), 1277–1291.

Charities Aid Foundation. (2017). *Charity trends.* Retrieved 2017-05-31, from `http://www.charitytrends.org`

Choi, J. & Kim, K.-E. (2011). Inverse reinforcement learning in partially observable

environments. *Journal of Machine Learning Research*, *12*(Mar), 691–730.

Chomsky, N., Roberts, I. & Watumull, J. (2023). The false promise of ChatGPT. *The New York Times*.

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S. & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, *30*.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science*, *19*(1), 51–57.

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

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., … Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology*, *25*(14), 1852–1859.

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

Dayan, P. & Hinton, G. E. (1992). Feudal reinforcement learning. *Advances in neural information processing systems*, *5*.

De Backer, C. J. & Hudders, L. (2015). Meat morals: Relationship between meat consumption consumer attitudes towards human and animal welfare and moral behavior. *Meat science*, *99*, 68–74.

De Ridder, D., Kroese, F., Adriaanse, M. & Evers, C. (2014). Always gamble on an empty stomach: Hunger is associated with advantageous decision making. *PLoS one*, *9*(10), e111081.

Diaconescu, A. O., Mathys, C., Weber, L. A., Daunizeau, J., Kasper, L., Lomakina, E. I., … Stephan, K. E. (2014). Inferring on the intentions of others by hierarchical Bayesian

learning. *PLoS computational biology*, *10*(9), e1003810.

Diaye, M.-A., Gardes, F. & Starzec, C. (2008). GARP violation, economic environment distortions and shadow prices: Evidence from household expenditure panel data. *Annales d'Économie et de Statistique*, 3–33.

Diaye, M.-A. & Urdanivia, M. W. (2009). Violation of the transitivity axiom may explain why, in empirical studies, a significant number of subjects violate GARP. *Journal of Mathematical Psychology*, *53*(6), 586–592.

Dolan, R. J. & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Doris, J., Stich, S., Phillips, J. & Walmsley, L. (2020). Moral psychology: Empirical approaches. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/spr2020/entries/moral-psych-emp/`.

Edwards, C. P. (1986). Cross-cultural research on Kohlberg's stages: The basis for consensus. In *Lawrence kohlberg: Consensus and controversy* (pp. 419–430). Routledge.

Elassar, A. (2020). Name a cockroach after your ex and watch an animal eat it on Valentine's Day. *CNN*. Retrieved from `https://edition.cnn.com/2020/02/09/us/san-antonio-zoo-cockroach-valentines-day-trnd/index.html`

Ellemers, N., Van Der Toorn, J., Paunov, Y. & Van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, *23*(4), 332–366.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I. & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, *31*.

Engelmann, J. B. & Fehr, E. (2016). The slippery slope of dishonesty. *Nature Neuroscience*, *19*(12), 1543–1544.

Falk, A. & Szech, N. (2013). Morals and markets. *Science*, *340*(6133), 707–711.

Feier, T., Gogoll, J. & Uhl, M. (2022). Hiding behind machines: Artificial agents may help to evade punishment. *Science and Engineering Ethics*, *28*(2), 19.

Février, P. & Visser, M. (2004). A study of consumer behavior using laboratory data. *Experimental Economics*, *7*, 93–114.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and decision*, *13*(2), 139–199.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262.

Floridi, L. & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681–694.

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

Formosa, P. & Ryan, M. (2021). Making moral machines: Why we need artificial moral agents. *AI & society*, *36*, 839–851.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D. & Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, *2*(4), 189–210.

Gamez, P., Shank, D. B., Arnold, C. & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & Society*, *35*, 795–809.

Garrett, N., Lazzaro, S. C., Ariely, D. & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*.

Giroux, M., Kim, J., Lee, J. C. & Park, J. (2022). Artificial intelligence and declined guilt: Retailing morality comparison between human and ai. *Journal of Business Ethics*, *178*(4), 1027–1041.

Gneezy, U., Keenan, E. A. & Gneezy, A. (2014). Avoiding overhead aversion in charity.

Science, *346*(6209), 632–635.

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R. & Martin, N. R. (2022). Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior*, *130*, 107179.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P. & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, *47*.

Graham, J., Haidt, J. & Nosek, B. (2008). *Moral foundations questionnaire, MFQ 30, revised in July 2008. Extracted March 2015.*

Graham, J., Haidt, J. & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, *96*(5), 1029.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S. & Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, *101*(2), 366.

Greene, J. D. (2002). *The terrible, horrible, no good, very bad truth about morality and what to do about it* (Unpublished doctoral dissertation). Princeton University.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, vol. 3. the neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–79). MIT Press.

Greene, J. D. (2023). The dual-process theory of moral judgment does not deny that people can make compromise judgments. *Proceedings of the National Academy of Sciences*, *120*(6).

Griffin, J. (1979). Is unhappiness morally more important than happiness? *The Philosophical Quarterly*, *29*(114), 47–55.

Guzmán, R. A., Barbato, M. T., Sznycer, D. & Cosmides, L. (2022). A moral trade-off system produces intuitive judgments that are rational and coherent and strike a balance between conflicting moral values. *Proceedings of the National Academy of Sciences*,

119(42).

Hadfield-Menell, D., Russell, S. J., Abbeel, P. & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in neural information processing systems* (pp. 3909–3917).

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Haidt, J. & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, *20*(1), 98–116.

Haidt, J. & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Dædalus*, *133*(4), 55–66.

Haidt, J. & Joseph, C. (2011). How moral foundations theory succeeded in building on sand: A response to Suhler and Churchland. *Journal of Cognitive Neuroscience*, *23*(9), 2117–2122.

Hammerstein, P. & Selten, R. (1994). Game theory and evolutionary biology. *Handbook of game theory with economic applications*, *2*, 929–993.

Harms, W. & Skyrms, B. (2008, 07). Evolution of Moral Norms. In *The Oxford Handbook of Philosophy of Biology*. Oxford University Press.

Harper, R. F. (1904). *The Code of Hammurabi, king of Babylon, about 2250 BC*. University of Chicago Press, Callaghan.

Haug, M. R. (1972). Deprofessionalization: an alternate hypothesis for the future. *The Sociological Review*, *20*(1_suppl), 195–211.

Hawking, S., Russell, S., Tegmark, M. & Wilczek, F. (2014). Transcendence looks at the implications of artificial intelligence–but are we taking AI seriously enough? *The Independent*, *1*, 2014.

Hayes, J. E. & Keast, R. S. (2011). Two decades of supertasting: Where do we stand? *Physiology & behavior*, *104*(5), 1072–1074.

Heider, F. & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Heiphetz, L., Strohminger, N. & Young, L. L. (2016). The role of moral beliefs, memories, and preferences in representations of identity. *Cognitive science*.

Hills, P. & Argyle, M. (2002). The oxford happiness questionnaire: A compact scale for the measurement of psychological well-being. *Personality and Individual Differences*, *33*(7), 1073–1082.

Horne, Z., Powell, D. & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive science*, *39*(8), 1950–1964.

Hursthouse, R. & Pettigrove, G. (2022). Virtue ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2022/entries/ethics-virtue/`.

Iurino, K. & Saucier, G. (2020). Testing measurement invariance of the moral foundations questionnaire across 27 countries. *Assessment*, *27*(2), 365–372.

Jentzsch, S., Schramowski, P., Rothkopf, C. & Kersting, K. (2019). Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (pp. 37–44).

Jiang, Y., Marcowski, P., Ryazanov, A. & Winkielman, P. (2023). People conform to social norms when gambling with lives or money. *Scientific Reports*, *13*(1), 853.

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J. & Bertulfo, D. J. (2022). The ghost in the machine has an American accent: Value conflict in GPT-3. *arXiv preprint arXiv:2203.07785*.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J. & Savulescu, J.

(2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kirby, K. N. & Herrnstein, R. J. (1995). Preference reversals due to myopic discounting of delayed reward. *Psychological science*, *6*(2), 83–89.

Kleiman-Weiner, M., Saxe, R. & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Kohlberg, L. (1976). Moral stages and moralization: The cognitive-development approach. *Moral development and behavior: Theory research and social issues*, 31–53.

Kohlberg, L. & Kramer, R. (1969). Continuities and discontinuities in childhood and adult moral development. *Human development*, *12*(2), 93–120.

Kouchaki, M. (2011). Vicarious moral licensing: The influence of others' past moral actions on moral behavior. *Journal of personality and social psychology*, *101*(4), 702.

Kouchaki, M., Smith, I. H. & Savani, K. (2018). Does deciding among morally relevant options feel like making a choice? How morality constrains people's sense of choice. *Journal of personality and social psychology*, *115*(5), 788.

LaCroix, T. (2022). Moral dilemmas for moral machines. *AI and Ethics*, *2*(4), 737–746.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J. & Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 464–473).

Levy, A. (2011). Game theory, indirect modeling, and the origin of morality. *The Journal of Philosophy*, *108*(4), 171–187.

Loewenstein, G. & Small, D. A. (2007). The scarecrow and the tin man: The vicissitudes of human sympathy and caring. *Review of general psychology*, *11*(2), 112–126.

MacAskill, W. (2014). *Normative uncertainty* (Unpublished doctoral dissertation). Univer-

sity of Oxford.

MacAskill, W. (2019a). The definition of effective altruism. In H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues* (pp. 10–28). Oxford University Press Oxford.

MacAskill, W. (2019b). Practical ethics given moral uncertainty. *Utilitas*, *31*(3), 231–245.

MacAskill, W. & Ord, T. (2020). Why maximize expected choice-worthiness? *Noûs*, *54*(2), 327–353.

Machery, E. & Mallon, R. (2010). Evolution of morality. *The moral psychology handbook*, 3–46.

Makowski, D., Ben-Shachar, M. S. & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

Malle, B. F., Magar, S. T. & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. *Robotics and well-being*, 111–133.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 117–124).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman and Company.

Martinho, A., Poulsen, A., Kroesen, M. & Chorus, C. (2021). Perspectives about artificial moral agents. *AI and Ethics*, *1*(4), 477–490.

Mattei, A. (2000). Full-scale real tests of consumer behavior using experimental data. *Journal of Economic Behavior & Organization*, *43*(4), 487–497.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of

systems of processes in cascade. *Psychological review*, *86*(4), 287.

McConnell, T. (2022). Moral dilemmas. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/moral-dilemmas/.

Merritt, A. C., Effron, D. A. & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and personality psychology compass*, *4*(5), 344–357.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). GoogleNews-vectors-negative300.bin.gz: Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. Retrieved from https://code.google.com/archive/p/word2vec/

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38.

Mosteller, F. & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*(5), 371–404.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Nilsson, A., Erlandsson, A. & Västfjäll, D. (2020). Moral foundations theory and the psychology of charitable giving. *European Journal of Personality*.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Osborne, M. J. et al. (2004). *An introduction to game theory* (Vol. 3) (No. 3). Oxford university press New York.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … others

(2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J. & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170–4175.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L. & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social neuroscience*, *9*(1), 94–107.

PayPal editorial staff. (n.d.). *Credit card processing fees.* Retrieved 2020-09-01, from `https://www.paypal.com/us/brc/article/understanding-merchant-credit-card-processing-fees`

Persad, G., Wertheimer, A. & Emanuel, E. J. (2009). Principles for allocation of scarce medical interventions. *The Lancet*, *373*(9661), 423–431.

Piff, P. K., Kraus, M. W. & Keltner, D. (2018). Unpacking the inequality paradox: The psychological roots of inequality and social class. In *Advances in experimental social psychology* (Vol. 57, pp. 53–124). Elsevier.

Pizarro, D., Uhlmann, E. & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological science*, *14*(3), 267–272.

Portillo, J. E. & Stinn, J. (2018). Overhead aversion: Do some types of overhead matter more than others? *Journal of behavioral and experimental economics*, *72*, 40–50.

Prescott, J., Soo, J., Campbell, H. & Roberts, C. (2004). Responses of prop taster groups to variations in sensory qualities within foods and beverages. *Physiology & Behavior*, *82*(2-3), 459–469.

Pronin, E., Lin, D. Y. & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381.

Pryor, C., Perfors, A. & Howe, P. D. (2019). Even arbitrary norms influence moral

decision-making. *Nature human behaviour*, *3*(1), 57–62.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. et al. (2018). Improving language understanding by generative pre-training.

Railton, P. (2020). Ethical learning, natural and artificial. *Ethics of artificial intelligence*, 45.

Rand, D. G., Greene, J. D. & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427–430.

Řehůřek, R. & Sojka, P. (2010, 22nd May). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks* (pp. 45–50). Valletta, Malta: ELRA. (`http://is.muni.cz/publication/884893/en`)

Reyniers, D. & Bhalla, R. (2013). Reluctant altruism and peer pressure in charitable giving. *Judgment and Decision making*, *8*(1), 7–15.

Rich, A. S. & Gureckis, T. M. (2019). Lessons for artificial intelligence from the study of natural stupidity. *Nature Machine Intelligence*, *1*(4), 174–180.

Rozin, P., Markwith, M. & Stoess, C. (1997). Moralization and becoming a vegetarian: The transformation of preferences into values and the recruitment of disgust. *Psychological science*, *8*(2), 67–73.

Russell, B. (1950). The science to save us from science. *The New York Times*.

Russell, S. (1998). Learning agents for uncertain environments. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 101–103).

Sachdeva, S., Iliev, R. & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological science*, *20*(4), 523–528.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 chi conference on human factors in computing*

systems (pp. 1–15).

Sarfati, Y., Hardy-Bayle, M.-C., Nadel, J., Chevalier, J.-F. & Widlocher, D. (1997). Attribution of mental states to others by schizophrenic patients. *Cognitive Neuropsychiatry*, *2*(1), 1–18.

Sayre-McCord, G. (2021). Moral realism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2021/entries/moral-realism/`.

Scheidel, B. (2012). *Wordfind.* `https://github.com/bunkat/wordfind`.

Schein, C. & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, *41*(8), 1147–1163.

Schein, C. & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of economic literature*, 529–563.

Schoemaker, P. J. H. (2013). *Experiments on decisions under risk: The expected utility hypothesis*. Springer Science & Business Media.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, *4*(3), 258–268.

Shank, D. B. & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in human behavior*, *86*, 401–411.

Shanks, D. R., Tunney, R. J. & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Sheeran, P. & Webb, T. L. (2016). The intention–behavior gap. *Social and Personality Psychology Compass*, *10*(9), 503–518.

Shiffrin, R. & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, *120*(10), e2300963120.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J. & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Shou, Y., Olney, J., Smithson, M. & Song, F. (2020). Impact of uncertainty and ambiguous outcome phrasing on moral decision-making. *PLoS one*, *15*(5), e0233127.

Simon, H. A. (1990). Bounded rationality. *Utility and probability*, 15–18.

Singer, P. (2015). What is Effective Altruism? In *The most good you can do* (pp. 3–12). Yale University Press.

Smith, I. H., Aquino, K., Koleva, S. & Graham, J. (2014). The moral ties that bind... even to out-groups: The interactive effect of moral identity and the binding moral foundations. *Psychological science*, *25*(8), 1554–1562.

Smith, V. L. & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*(2), 245–261.

Suhler, C. L. & Churchland, P. (2011). Can innate, modular "foundations" explain morality? Challenges for Haidt's moral foundations theory. *Journal of cognitive neuroscience*, *23*(9), 2103–2116.

Surdina, A. & Sanborn, A. (2017). Temporal variability in moral value judgement. In *Proceedings of the 39th annual meeting of the cognitive science society (cogsci)*.

Suter, R. S. & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458.

Tarantola, T., Kumaran, D., Dayan, P. & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature Communications*, *8*(1), 817.

Agent. (n.d.). In *APA dictionary of psychology*. American Psychological Association.

Retrieved 2023-03-01, from https://dictionary.apa.org/agent

Thomson, J. J. (1984). The trolley problem. *Yale Law Journal*, *94*, 1395.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Usher, M. & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. (2023). Chatgpt: Five priorities for research. *Nature*, *614*(7947), 224–226.

Van Leeuwen, F. & Park, J. H. (2009). Perceptions of social dangers, moral foundations, and political orientation. *Personality and Individual Differences*, *47*(3), 169–173.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the econometric society*, 945–973.

Vineberg, S. (2022). Dutch book arguments. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2022/entries/dutch-book/.

Vlaev, I. (2012). How different are real and hypothetical decisions? Overestimation, contrast and assimilation in social interaction. *Journal of Economic Psychology*, *33*(5), 963–972.

Wilson, H. J. & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114–123.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M. & Baeza-Yates, R. (2017). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 1569–1578).

Zhang, H., Su, Y., Peng, L. & Yao, D. (2010). A review of game theory applications in transportation analysis. In *2010 international conference on computer and information application* (pp. 152–157).

Moral Foundations Questionnaire

Displayed text for questions in this column: When you decide whether something is right or wrong, to what extent is the following consideration relevant to your thinking?	**Displayed text for questions in this column:** Please read the following sentence and indicate your agreement or disagreement.
"Harm" Foundation:	
1. Whether or not someone suffered emotionally 2. Whether or not someone cared for someone weak or vulnerable 3. Whether or not someone was cruel	4. Compassion for those who are suffering is the most crucial virtue. 5. One of the worst things a person could do is hurt a defenseless animal. 6. It can never be right to kill a human being.
"Fairness" Foundation:	
1. Whether or not some people were treated differently than others 2. Whether or not someone acted unfairly 3. Whether or not someone was denied his or her rights	4. When the government makes laws, the number one principle should be ensuring that everyone is treated fairly. 5. Justice is the most important requirement for a society. 6. I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.
"Loyalty" Foundation:	
1. Whether or not someone's action showed love for his or her country 2. Whether or not someone did something to betray his or her group 3. Whether or not someone showed a lack of loyalty	4. I am proud of my country's history. 5. People should be loyal to their family members, even when they have done something wrong. 6. It is more important to be a team player than to express oneself.
"Authority" Foundation:	
1. Whether or not someone showed a lack of respect for authority 2. Whether or not someone conformed to the traditions of society 3. Whether or not an action caused chaos or disorder	4. Respect for authority is something all children need to learn. 5. Men and women each have different roles to play in society. 6. If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.
"Purity" Foundation:	
1. Whether or not someone violated standards of purity and decency 2. Whether or not someone did something disgusting 3. Whether or not someone acted in a way that God would approve of	4. People should not do things that are disgusting, even if no one is harmed. 5. I would call some acts wrong on the grounds that they are unnatural. 6. Chastity is an important and valuable virtue.
Neutral or added questions:	
1. Whether or not someone was good at math 2. Whether or not someone told the truth (*) 3. Whether or not someone made a smart decision (*)	4. It is better to do good than to do bad. 5. If one's children live a happy life, it is better to have children than not to have children. (*) 6. Destroying beautiful things that took long to create is worse than destroying things that took less time. (*)

Figure A.1: Moral foundations questionnaire. Questions added by us for the purpose of including equally many 'neutral' items are marked with (*).

Charities

The self-descriptions of the four charities we chose as stimuli, as found on the websites of each organisation (at the time of access, in 2017), are printed below.

- **Elim Pentecostal Church:** The Elim Pentecostal Church is a growing Movement of Christian congregations. We believe the Bible, as originally given, to be without error, the fully inspired and infallible Word of God and the supreme and final authority in all matters of faith and conduct. We believe that the Godhead exists co-equally and co-eternally in three persons - Father, Son and Holy Spirit - and that these three are one God, sovereign in creation, providence and redemption. We believe in the spiritual unity and the priesthood of all believers in Christ and that these comprise the universal Church, the Body of Christ.
- **Christian Aid:** Christian Aid is a Christian organisation that insists the world can and must be swiftly changed to one where everyone can live a full life, free from poverty. We work globally for profound change that eradicates the causes of poverty, striving to achieve equality, dignity and freedom for all, regardless of faith or nationality. We are part of a wider movement for social justice. We provide urgent, practical and effective assistance where need is great, tackling the effects of poverty as well as its root causes.
- **Royal National Theatre:** At the National, we make world-class theatre

that is entertaining, challenging and inspiring. And we make it for everyone. The National Theatre is dedicated to making the very best theatre and sharing it with as many people as possible. The work we make strives to be as open, as diverse, as collaborative and as national as possible. We do all we can to keep ticket prices affordable, to reach a wide audience and to use our public funding to maintain artistic risk-taking, accessibility and diversity.

- **Sightsavers:** We want avoidable blindness to be eliminated. We want equality for people with disabilities. We help blind people to see again, and prevent people from going blind wherever we can. We improve the lives of people with disabilities, particularly those who have permanent sight loss. We need to change the lives of people at risk of sight loss for the long term, not just today. So we campaign to make the world a fairer place for people with disabilities and we tackle the underlying causes of avoidable blindness.

Game Design

Below is a screenshot of the unfinished prototype of '*Starstuck*', a variation of what the spatial cooperation task described in Chapter 4 could look like. Any further use of the proposed experimental design is *encouraged*.[1]

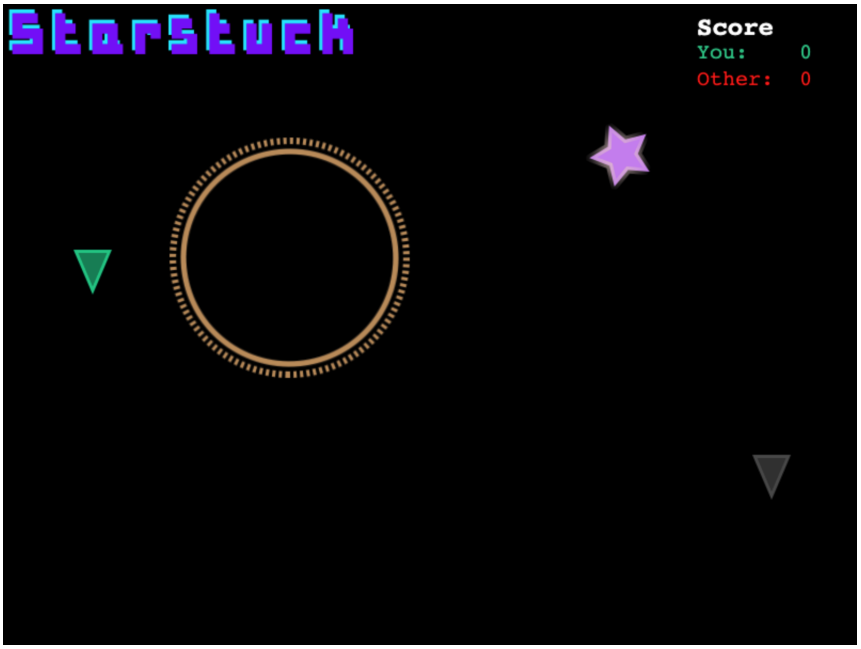


Figure C.1: Game design: The active player is represented as a green triangle on their own screen, the other player as a grey triangle. Players get points for collecting stars. Each player can move their triangle by influencing direction and acceleration.

Each player's own avatar is shown as a green triangle, with a grey triangle displaying the

[1]The code of the game's proof of concept is available at: `https://github.com/surdina/evil-or-stupid`

other player's location. Each triangle points towards its direction of movement.

Each player can control their triangle by pressing the arrow keys. The up arrow key increases acceleration into the direction the player is currently facing. The down arrow key accelerates the player in the other direction, intuitively useable to decelerate. Pressing the right arrow key rotates the triangle clockwise, and the left arrow key counterclockwise, without changing the direction of movement; the direction of movement can only be controlled by turning and then accelerating.

If a triangle collides with the boundary of the black game screen, it is reflected back in accordance to Newtonian mechanics, without losing speed. The surface of the game screen has friction which slows down the triangle, and can be varied; a lower friction coefficient makes steering more difficult.

The purple star disappears when a player navigates to its location to 'collect' it, which turns out to be surprisingly challenging when the triangle movement can only be controlled using acceleration. Collecting a star gives 10 points to the player, the star disappears, and a new star spawns in a random location.

The orange circle is a trap. It has a sticky boundary: if a player touches the outside of the circle, their triangle is sucked inside. If the second player also gets trapped, the trap blinks for a few seconds and disappears, releasing both; 10 points are deducted from their scores (this can be adjusted).

Whenever a trap gets activated, a teal-coloured trap deactivation button appears:

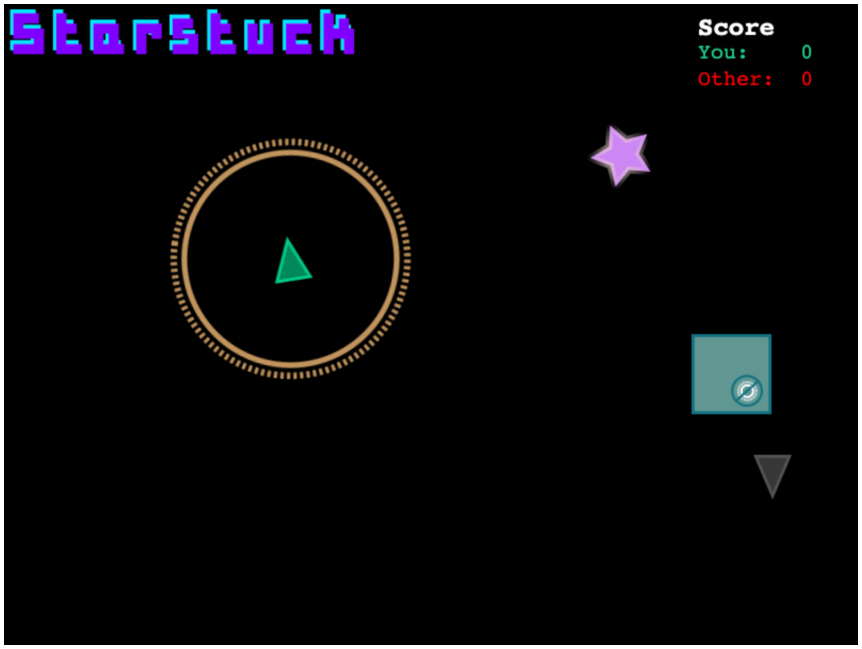


Figure C.2: Game design: Star collection and activation of a trap all occur when the player navigates to the on-screen position of the respective object's boundary. Trap deactivation is only possible by the free player analogously by navigating to the teal-coloured trap deactivation button.

While a player is trapped, the up and down arrow keys have no effect, and they can only change the direction the player's triangle is facing. Here, the active player is trapped, and the other player can choose to be helpful and navigate towards the release button and free the trapped player. Or they can opt to be selfish and continue collecting stars while the other player is stuck.

In some scenarios, the free player may have plausible deniability for not helping the other player and collecting stars instead—if they accelerated so much that they are clearly moving uncontrollably all over the place; or if they were moving towards the star's location when the other player got trapped. But in the current scenario, visible in the

screenshot, moving towards the star in a straight line would lead the grey player right by the release button. Now, let us assume we observe the grey player's decision leading to the following outcome:

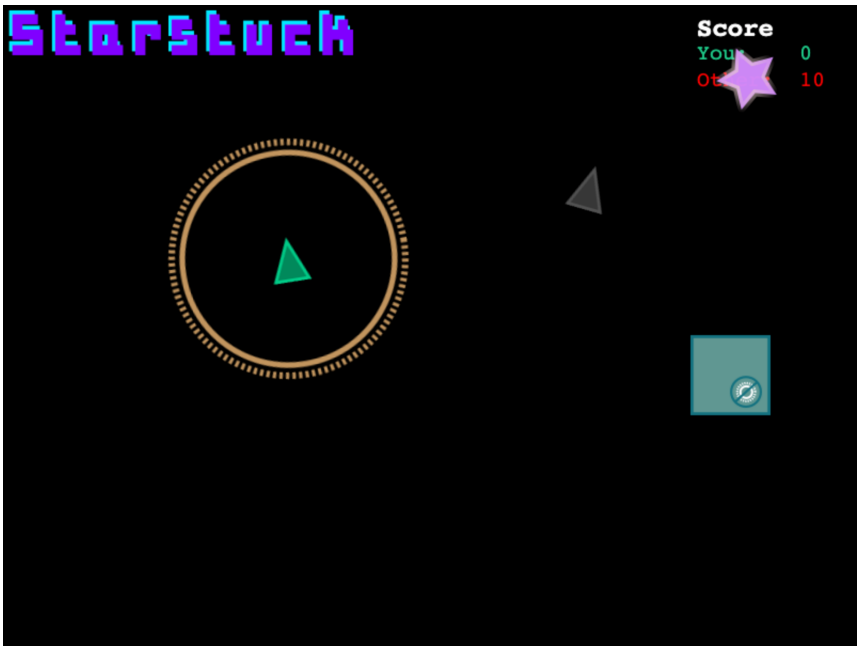


Figure C.3: Game design: Helpful actions (trap deactivation) and self-interested behaviour (star collection) lead to different trajectories of movement, allowing for intention inference based on observations of the other player's position over time.

To do what they just did—to collect the previous star—they would have had to steer explicitly around the star, suggesting their intentions towards us were less than cooperative.

Each player's behaviour can be manipulated by influencing how the scores are displayed (e.g. same or opposing team). Difficulty could be increased by manipulating the increase in acceleration when up/down arrow keys are pressed, or by randomly ignoring a player's keyboard input.