Identifying moral antecedents of decision-making in discrete choice models

Szép, T.

DOI
10.4233/uuid:1ca8f755-74ee-440a-8a78-550268c9ef54

Publication date
2022

Document Version
Final published version

Citation (APA)
Szép, T. (2022). *Identifying moral antecedents of decision-making in discrete choice models*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:1ca8f755-74ee-440a-8a78-550268c9ef54

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

IDENTIFYING MORAL ANTECEDENTS OF DECISION-MAKING IN DISCRETE CHOICE MODELS

IDENTIFYING MORAL ANTECEDENTS OF DECISION-MAKING IN DISCRETE CHOICE MODELS

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of Rector Magnificus Prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended in public on
Wednesday, 19 October 2022 at 10:00.

by

Teodóra SZÉP
Master of Science in Economics,
Vrije Universiteit Amsterdam
born in Budapest, Hungary

This dissertation has been approved by the promotors:

promotor: prof. dr. ir. C.G. Chorus
copromotor: dr. ir. S. van Cranenburgh

Composition of the doctoral committee:

Rector Magnificus,	chair
Prof. dr. ir. C.G. Chorus,	Delft University of Technology, promotor
Dr. ir. S. van Cranenburgh,	Delft University of Technology, copromotor
Independent members:	
Prof. dr. M. Bierlaire	École polytechnique fédérale de Lausanne, Switzerland
Prof. dr. C.F. Choudhury	University of Leeds, United Kingdom
Prof. dr. mr. ir. N. Doorn	Delft University of Technology
Prof. dr. S. Hess	Delft University of Technology
Prof. dr. ir. L.A. Tavasszy	Delft University of Technology, reserve member

This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 724431).

Keywords: Discrete Choice Modelling, Moral Decision Making, Identifiability, Methodological and Empirical Research
Printed by: Ridderprint, www.ridderprint.nl
Cover layout: Ridderprint, www.ridderprint.nl

ISBN 978-94-6384-375-1

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

CONTENTS

SUMMARY

Discrete Choice Models are valuable tools for quantitative decision-making analysis: they allow analysts to draw behavioural conclusions from data, better understand and predict choices, and evaluate policies. However, up until recently, they had a blind spot for morality. Moral values often play an essential role in decision-making; fairness or loyalty can deter people from following self-interest. Moral motivations can also prompt decision-makers to change their minds when contemplating a dilemma or hide their preferences when they want to avoid judgement. These notions are not aligned with crucial behavioural assumptions traditional Discrete Choice Models are based on, such as stable preferences echoing through choices or decision-makers maximizing their utility.

This thesis aims to develop and test new Discrete Choice Models that help identify morality in a mathematically rigorous framework, thus increasing the behavioural realism of Discrete Choice Models in moral decision-making. To do this, it uses two approaches.
First, in Part I, it tests two recently developed models, Decision Field Theory and the obfuscation model. These models relax the assumptions of utility maximization, stable, and revealed preferences and thus are promising tools for morality analysis. The thesis tests whether parameters can be uniquely identified or recovered without bias when only choice data is available.
Second, in Part II, it uses additional morality data to enrich mainstream Discrete Choice Models by using moral incentives, standard morality surveys, Likert-type contextual questionnaires, and extracting moral values from text with Natural Language Processing.

In Part I, Chapter 2 examines the identifiability of the recently adapted model of Decision Field Theory (DFT). DFT is a process model that aims to capture the contemplation process in a decision-maker's mind. I use analytical derivations to find DFT specifications equivalent to probit models and show that two special cases of DFT have identifiability issues; the process parameters cannot be identified. Examining the generic model using Monte Carlo simulations, I also find that psychological parameters cannot be recovered without bias. These results suggest that a deliberation process cannot be accurately retrieved from merely choice data.
Chapter 3 proposes an extension to the recently developed Obfuscation model, namely sequential obfuscation. Sequential obfuscation postulates that the decision-maker considers that not only their current choice but the previous ones were also observed. Chapter 3 examines the identifiability of preferences under both the original model's and the sequential extension's assumptions. I use Monte Carlo simulations and find that parameters can be recovered without bias, but an obfuscating intention reduces the analyst's confidence about parameter estimates.

In Part II, Chapter 4, morality data is collected through moral incentives, a standard morality survey and Likert-type contextual questions in a social routing context. I estimate standard linear additive utility mixed logit models to identify which moral motivations play a role when different incentives, namely a collective good based and a sacrifice-based scheme, are used. The results can be interpreted in light of moral psychology, in order to help identify which moral personality traits indicate that one is drawn to a collective good scheme or is willing to sacrifice their own free time for others' benefit.
Chapter 5 proposes a method to use Natural Language Processing output as input in Discrete Choice Models. To illustrate and test the method, I collect voting data from the European Parliament and text data from decision-makers. I use Natural Language Processing to extract moral rhetoric from the text and then estimate standard Discrete Choice Models enriched with these moral features. Estimates show that moral rhetoric has significant explanatory power when modelling voting behaviour and sheds light on which moral foundations can be connected to strategic language use.

In conclusion, this thesis contributes to the literature in three ways. First, it rigorously examines models that aim to model moral motivations in Discrete Choice Models by various structural innovations, relying on only choices and behavioural theory. Second, it operationalizes a theory from moral psychology in Discrete Choice Models in order to identify moral motivations and behavioural constructs, which would not be detectable relying on choice data alone. Third, this thesis contributes to the growing literature which connects data-driven solutions to theory-driven Discrete Choice Models. It does so by using, among the first, Natural Language Processing in Discrete Choice Models.

SAMENVATTING

Discrete keuzemodellen zijn waardevolle hulpmiddelen voor kwantitatieve besluitvormings analyse: ze stellen analisten in staat gedragsconclusies te trekken uit datagegevens, keuzes beter te begrijpen en te voorspellen en beleid te evalueren. Tot voor kort hadden ze echter een blinde vlek voor moraliteit. Morele waarden spelen vaak een essentiële rol bij besluitvorming; eerlijkheid of loyaliteit kan mensen ervan weerhouden hun eigenbelang te volgen. Morele motivaties kunnen besluitvormers van gedachten laten veranderen wanneer ze over een dilemma nadenken of ze motiveren om hun voorkeuren te verbergen wanneer ze een oordeel liever willen vermijden. Deze noties zijn niet in lijn met een aantal cruciale gedragsaannames waarop traditionele discrete keuzemodellen zijn gebaseerd, zoals het idee van stabiele voorkeuren die terugkomen in keuzes of dat besluitvormers hun nut maximaliseren.

Dit proefschrift heeft tot doel om nieuwe discrete keuzemodellen te ontwikkelen en te testen die, ingebed in een grondig wiskundig raamwerk, helpen om moraliteit te identificeren, en zo het gedragsrealisme van discrete keuzemodellen in morele besluitvorming vergroten. Om dit te doen, wordt er gebruik gemaakt van twee benaderingen.
Ten eerste, worden in Deel I twee recent ontwikkelde modellen getest: de Beslissingsveldtheorie en het obfuscatiemodel. Deze modellen versoepelen de aannames van nutsmaximalisatie, stabiele voorkeuren en onthulde voorkeuren en zijn dus veelbelovende hulpmiddelen voor moraliteitsanalyse. Het proefschrift test of parameters uniek kunnen worden geïdentificeerd of kunnen worden teruggevonden zonder dat dit leidt tot bias , wanneer alleen keuzedata beschikbaar zijn.
Ten tweede, in deel II, worden aanvullende moraliteitsgegevens gebruikt om de reguliere discrete keuzemodellen te verrijken. Hiervoor maak ik gebruik van morele stimuli, standaard moraliteitsenquêtes, Likert-achtige contextuele vragenlijsten en het extraheren van morele waarden uit tekst, door middel van natuurlijke taalverwerking.

In Deel I onderzoekt Hoofdstuk 2 de identificeerbaarheid van het recent aangepaste Beslissingsveldtheorie-model (Decision Field Theory, DFT). DFT is een procesmodel dat het overwegingsproces in het hoofd van de beslisser probeert vast te leggen. Ik gebruik analytische afleidingen om DFT-specificaties te vinden die gelijkwaardig zijn aan probitmodellen en laat zien dat twee speciale gevallen van DFT identificeerbaarheidsproblemen hebben; de procesparameters kunnen niet worden geïdentificeerd. Bij het onderzoeken van het generieke model, met behulp van Monte Carlo-simulaties, vind ik ook dat psychologische parameters niet kunnen worden teruggevonden zonder dat dit tot bias leidt. Deze resultaten suggereren dat een overwegingsproces niet nauwkeurig kan worden opgehaald uit louter keuzegegevens.
Hoofdstuk 3 stelt een uitbreiding voor op het recent ontwikkelde Obfuscatiemodel, namelijk sequentiële obfuscatie . Sequentiële obfuscatie stelt dat de beslisser van mening is dat niet alleen zijn huidige keuze, maar ook zijn eerdere keuzes zijn waargenomen.

Hoofdstuk 3 onderzoekt de identificeerbaarheid van voorkeuren onder zowel de aannames van het originele model als die van de sequentiële extensie. Ik gebruik Monte Carlo-simulaties en vind dat parameters kunnen worden hersteld zonder dat dit leidt tot bias, maar een verduisterende intentie vermindert het vertrouwen van de analist in parameterschattingen.

In Deel II, Hoofdstuk 4, worden moraliteitsgegevens verzameld door middel van morele stimuli, een standaard moraliteitsvragenlijst en Likert-achtige contextuele vragen in een social routingcontext. Ik schat standaard mixed logit modellen met lineaire nutsfunctie om te identificeren welke morele motivaties een rol spelen wanneer verschillende stimuli, namelijk een collectief goed-gebaseerd en een opoffering-gebaseerd schema, worden gebruikt. De resultaten worden geïnterpreteerd in het licht van de morele psychologie: ze identificeren welke morele persoonlijkheidskenmerken aangeven dat iemand zich aangetrokken voelt tot een collectief goed plan of bereid is zijn eigen vrije tijd op te offeren voor het welzijn van anderen.
Hoofdstuk 5 stelt een methode voor om Natural Language Processing output te gebruiken als input voor discrete keuzemodellen. Om de methode te illustreren en te testen, verzamel ik stemgegevens van het Europees Parlement en tekstgegevens van besluitvormers. Ik gebruik Natural Language Processing om morele retoriek uit de tekst te extraheren en schat vervolgens standaard discrete keuzemodellen die zijn verrijkt met deze morele kenmerken. Schattingen laten zien dat morele retoriek een grote verklarende kracht heeft bij het modelleren van stemgedrag en werpt licht op welke morele kenmerken kunnen worden verbonden met strategisch taalgebruik.

Concluderend draagt dit proefschrift op drie manieren bij aan de literatuur. Ten eerste onderzoekt het op een grondige manier modellen die gericht zijn op het modelleren van morele motivaties in discrete keuzemodellen door verschillende structurele innovaties, waarbij het alleen gebruik maakt van keuzegegevens en van gedragstheorie. Ten tweede operationaliseert het een theorie uit de morele psychologie in discrete keuzemodellen met het doel om morele motivaties en gedragsconstructies te identificeren, die niet beschikbaar zouden zijn op basis van keuzegegevens alleen. Ten derde draagt dit proefschrift bij aan de groeiende literatuur die data gestuurde oplossingen verbindt met theorie gestuurde discrete keuzemodellen. Het doet dit door, als een van de eersten, gebruik te maken van Natural Language Processing in discrete keuzemodellen.

1

INTRODUCTION

1.1. RESEARCH BACKGROUND

Decision-making is an everyday task for each individual: deciding what to wear, when to leave for work, what travel mode to take, or what groceries to buy. In many situations, choices have a high impact on one's life: which career path to take, how to raise children, or in which country to live. Moreover, in some cases, decisions affect not only a few but also many people's lives: how to develop an urban area or which politician to elect. Hence, it is no surprise that decision-making is a subject of several studies in a wide range of domains, from psychology to political science. Observing and analyzing individual choices can lead to insights into *how* a decision is made, what trade-offs are relevant, and infer the importance weights one attaches to different aspects of a decision. For instance, when one decides which route to take when going to work, decision-making can be travel time, congestion, weather, or one's morning schedule at work. Observing route choice behaviour several times for one or more individuals, analysts can estimate the importance weight of such aspects. The estimation outcomes can often be used, for instance, to predict travel flow. Such predictions are often used as input for cost-benefit analysis of product development, urban development or policy implementations. Thus, in order to better understand, predict, or evaluate human decision-making, quantitative analysis of choices is crucial.

The standard practice to quantitatively analyze preferential choice between discrete alternatives is the Discrete Choice Model (DCM) family (McFadden, 1973). DCMs have been used for several decades by now and have proved to be highly useful for scholars in many fields, including but not limited to transportation, marketing and health care. DCMs are formally connected to the widely used economic theory of utility maximization (McFadden, 1973), which won a Nobel Prize to Daniel McFadden. DCMs allow for several behavioural inferences, such as one's willingness to pay for an additional feature or preference order between the attributes, thus proving instrumental in the economic

appraisal of policies or products.

In order to be tractable and allow for behavioural and economic inferences, DCMs impose three major psychological assumptions on decision-makers.

First, that decision-makers are maximizing their own utility. The model postulates that decision-makers are willing to make trade-offs between the attributes of alternatives, such as choosing a worse quality product to pay less for it. DCMs estimate the relative importance of the attributes, the weights that represent the decision-maker's preferences. According to the model, the utility of each alternative is calculated based on their attributes and corresponding preference weights. The alternative with the highest utility is going to be chosen.

Second, that preferences are complete and stable. Complete preferences mean that for every pair of alternatives in the choice set, the decision-makers can decide whether they prefer one or the other or they are indifferent. Stability of preferences means that decision-makers do not change their minds during or after deliberation.

Third, that the preferences echo through the choices decision-makers make. This means that the observed choices directly result from the observed attributes. If someone chooses a blue car over an identical red one, they prefer blue to red.

Despite their long history and a broad range of use cases, up until recently, DCMs had a blind spot for moral aspects of decision-making. Morality can be defined as an individual's normative judgement about what is right and wrong. In a decision-making context, it can be the case that one has to choose between two wrong alternatives (classic moral dilemmas, such as the trolley problem by Foot (1967)) or between one morally right and one morally wrong, where the wrong is usually motivated by self-interest or care (e.g., working for an environment polluting corporation for a high salary, or stealing food for a starving loved one). Whether a choice task has moral dimensions depends on how the decision-maker perceives it. However, people generally have a shared common sense of morality (e.g., "do not cause pain", "do not deprive of freedom", "help others" in, for instance, Gert (2004)), which helps an outside observer or analyst to judge whether the choice task has moral dimensions or not. For example, if we observe someone choosing a travel mode, it is not evident whether they consider moral dimensions to it; often, people only evaluate the travel time and cost. However, if their attention is drawn to how their action might affect others adversely (i.e., congested roads cause pollution and travel time loss for everyone), then the task is one with a moral dimension.
Morality is often a part of small everyday choices and high impact decisions as well. Looking out for the benefit of others or complying with the rules of society at large is crucial in human interactions, and is often studied in research fields such as evolutionary theory, genetics, psychology or anthropology (e.g., Curry, 2016; Israel et al., 2015; Tomasello et al., 2012; Tomasello and Vaish, 2013). These kinds of moral behaviours can contribute to the formation of social norms, which then create other incentives, such as avoidance of shame. In many everyday situations, one has to navigate between several moral incentives (e.g., lying not to hurt someone's feelings), which can lead to insecurities about preferences in terms of what is right or wrong. This can result in long contemplation processes, one changing their mind several times or making a quick emotional

decision. High impact decisions, such as voting in an election or implementing a policy, often have clear moral components. Parties in the political arena often separate themselves from each other based on moral values; for instance, whether a party supports gender equality or traditional gender roles has a significant effect on partisanship.

Although morality has been widely studied in philosophy and psychology, DCMs traditionally cannot capture the moral dimensions. The main reason for it is that the nature of underlying preferences is quite different when it comes to consumer (non-moral) and moral attributes in a decision-making task. As described above, DCMs have crucial behavioural assumptions about the decision-maker's decision rule (utility maximization) and the nature of their preferences; stable, complete and echo through the choices (e.g., Deaton and Muellbauer, 1980). However, when it comes to moral decisions, scholars have often theorized and empirically observed that these behavioural and psychological assumptions on which DCMs are based do not hold. As these assumptions are the fundamental building blocks of decision-making modelling, it is crucial to understand the different theories and observations related to these concepts.

Utility maximization is the most prevalent decision rule of people making choices, according to traditional DCMs. Several studies find that in many situations, this assumption does not hold (see for an overview, e.g. Leong and Hensher, 2012). There is a set of arbitrary decision rules that reflect that cognitive capacities are limited. This set of rules is called bounded rationality (Simon, 1957). Bounded rationality means that individuals' rational behaviour is bounded because of their limits in information cognition, processing, and time dedicated to a task. Therefore they use heuristics or shortcuts when confronted with a decision. Gigerenzer (2010) presents a few of such moral heuristics such as default bias, which can steer people towards being organ donors, or peer imitation, which can foster prosocial behaviour such as contributing to charity, but undesirable behaviour such as discrimination too. Another line of the literature refuting the utility maximization principle emphasizes the role of emotions and suggests that moral choices are predominantly driven by intuition, and reasoning is built up in order to justify the judgement, not the other way around (Haidt, 2001). For example, wrongdoing against us from someone we trust is judged worse than the same action of a stranger: the emotion of anger might trigger a strong response that is later justified on the ground of betrayal. Haidt and Joseph (2004) argue that moral intuition is triggered by so-called moral foundations, which are essential building blocks of human morality; everyone is born with it, and culture and environment affect to what extent they become important to different individuals.

Preference stability is often not a realistic assumption (e.g., Braga and Starmer, 2005; Tversky and Thaler, 1990), especially in situations involving morality. There can be different reasons behind this, for example, reinforcement from society or emotional decisions followed by rationalization. A next decision based on emotions might result in a completely different rationalization of preferences. These preference changes can mean that two choices are not based on the same preferences; thus, estimating a model where stability is assumed would result in misguided conclusions.

When someone acts altruistically or cooperates with others or contributes to something deemed good by society, others often praise them. On the contrary, causing harm, be-

trayal or taking advantage of others is frowned upon. This mechanism can foster prosocial behaviour more and more as social norms become stronger. When someone acts on instinct or based on a 'gut feeling', arguably, they have no initial set of preferences based on which trade-offs are evaluated. However, once their decision is made, the brain creates an ex-post narrative, an explanation that potentially involves moral values, and it has been argued that people often internalize these narratives; thus in a later decision, they can serve as an initial preference (Haidt, 2001). Theoretical explanations for these phenomena are provided by the discovered preference hypothesis or coherent arbitrariness. According to discovered preference hypothesis, preferences do exist; however, the individual in some new or complex situations might not know them (Plott, 1993). The decision-maker discovers his preferences through the repetition of the task. Coherent arbitrariness suggests that the stable, well-defined preferences do not exist prior to an unfamiliar choice task, but they are constructed with an inner drive for consistency as the individual gains familiarity (Ariely et al., 2003).

Incomplete preferences mean a person is not able to decide[1]. This can be the case in very abstract moral dilemmas such as the trolley problem (Foot, 1967) but in more realistic scenarios such as referendums (Søberg and Tangerås, 2007). Not being able to decide can manifest itself in the decision-maker removing themselves from the situation; for example, by not answering the dilemma or not voting on the referendum. When making a choice in such situations is unavoidable, people might contemplate for a while and choose when one alternative seems to be good enough, they might choose randomly or try to obfuscate their incomplete preferences.

Preferences often do not echo through the choices one makes. For instance, when someone makes random choices (which can be a sensible choice when one has incomplete preferences) or tries to deceive others (which is against the moral imperative of 'do not lie'), their true underlying preferences cannot be inferred. In some choice situations, agents may wish to obfuscate the preferences underlying their choices from the observers while also fulfilling their true preferences to some extent (Chorus et al., 2021). The reason for such behaviours could be incomplete preferences, avoiding judgement, embarrassment or shame.

Identifying antecedents of moral decision-making, such as perceptions, emotions, and preference evolution, are not traditionally considered in discrete choice models. Traditional choice models ignore why preferences form a certain way or, in other words, why decision-makers want what they want, and for a long time, the examined variables were observable, such as the attributes of alternatives or socioeconomic variables (Vij and Walker, 2016a). In the past two decades, however, there has been an increased interest in a higher behavioural realism in choice modelling to reach more accurate estimations or predictions or to gain additional insights, for example, on the effect of environmental concern (being a latent feature of a person, estimated using psychometric data) on vehicle choice (Bolduc et al., 2008). Besides the traditional cost-benefit trade-offs humans make, additional insights into how different latent phenomena, such as environmental concern, or more generally, morality or perceptions, affect decision-making can

[1] Not being able to decide is not the same as being indifferent. Indifference rather manifests itself in random choice, as the decision-maker is not (necessarily) uncomfortable with either of the alternatives.

be instrumental in policy making or corporate arenas. Targeting intrinsic motivations, for instance, with informational campaigns or sharing mobile applications, are more and more widely used and researched, for instance, in the field of mobility to decrease congestion or environmental pollution (e.g., Klein and Ben-Elia, 2018; Mariotte, Leclercq, Gonzalez Ramirez, et al., 2021; Van Essen et al., 2016). Latent variable (or hybrid) choice models are specifically designed to identify different perceptions, attitudes, emotions and latent motivations in discrete choice analysis (Ashok et al., 2002; Ben-Akiva et al., 2002). Other advanced models such as mixed logit models or latent class models were also used to capture behavioural phenomena such as interaction with other decision-makers (Lovreglio et al., 2016) or preference endogeneity (Vij and Walker, 2016a). The theoretical and empirical identifiability of advanced discrete choice models[2] is crucial in order to draw behavioural conclusions from the estimates, but it is often a challenging task. Theoretical identifiability issues may arise when the model is specified in a way that allows for several parameter combinations to give the same result. Empirical identifiability issues arise when the data on which the models are estimated do not allow for unique recoverability of the true parameters (e.g., Raveau et al., 2012). Both forms of identifiability were addressed in several studies, which provided important guidelines for identification, precise interpretation, and how far inferences can be taken with particular types of data. For mixed logit, for instance, Hensher and Greene (2003) examines the effect of the number of draws, and Walker (2002) and Walker et al. (2007) give theoretical guidelines regarding parameter identifiability. For latent class models Gonzalez-Valdes et al. (2022) establish necessary conditions on the classes to be identifiable jointly; sufficient behavioural difference among the classes and a sufficient number of cases that expose the difference. Latent variable models' estimates were found to be often wrongfully interpreted as latent attitudes directly affecting choice behaviour (which would allow a policy maker to reach the desired choice behaviour by influencing the attitudinal factors); however, endogeneity of the latent variable and cross-sectional data preclude such causal inferences (Chorus and Kroesen, 2014). Vij and Walker (2016b) derives guidance on the interpretation and practical usefulness of latent variable models under various circumstances and research goals. Identifying latent behavioural constructs and precisely interpreting the outcomes of discrete choice models remains a non-trivial problem for each model and dataset.

Moral choice situations trigger emotions, intuitions, and alternative decision-making rules; thus, moral attributes are more challenging to understand for decision-makers themselves as well as for the analyst than consumer preferences in a DCM framework. However, representing morality in rigorous quantitative analysis of choice behaviour is crucial to increase the behavioural realism and predictions of such analyses. Section 1.1.1 introduces DCMs which aim, or have the potential, to model moral antecedents of decision-making, and section 1.1.2 presents theories of moral psychology that aim to categorize moral values and thus have the potential to serve as a base for moral attributes.

[2] By *advanced*, I refer to models with higher complexity than the standard multinomial logit model, which is identifiable and readily interpretable, but it comes at the cost of several limitations, such as not being able to capture random (i.e., not connected to observed variables) taste variation.

1

1.1.1. RECENT MODEL DEVELOPMENTS IN MORAL DISCRETE CHOICE MODELLING

Few recently developed or adapted (from mathematical psychology, for instance) models in the field of choice modelling have several advantageous qualities that make them potentially valuable tools for modelling moral dimensions.
The Taboo-Trade-off Aversion model (TTOA, Chorus et al., 2018) uses utility maximization but also constructs a moral attribute: a penalty term when the decision-maker is making a trade-off that is taboo. A trade-off is considered taboo when the subjects belong to different 'spheres'—for instance, paying less tax resulting in a less safe road that leads to more fatal accidents. Paying less tax belongs to a 'market sphere', while lost human life due to the accidents belongs to a 'moral sphere'.
The recently adapted version of Decision Field Theory (DFT, Hancock et al., 2018) is a dynamic model that aims to capture the deliberation process in a decision-maker's mind. Thus, it has the potential to accommodate insecurity, contemplation or changing minds, which are often the case in rather complex tasks, such as moral decision-making.
Quantum models [3] of moral decision-making (Hancock et al., 2020) use the mathematical formalization of quantum probability rules, which allow the analyst to model and explain, for instance, changing perspectives in morally salient situations, such as considering another person's interest when making a decision.
The Obfuscation model (Chorus et al., 2021) relaxes the assumption that preferences echo through the choices and aims to capture obfuscating behaviour (i.e. hiding one's true underlying preferences) through information entropy in choice tasks.

Table 1.2 shows these recent models, which have advantageous features for modelling morality. The table summarizes what behavioural phenomena are captured that in traditional DCMs are overlooked ("Behavioural phenomena"), how does the model capture it ("Structure") and an example where it is relevant in morality ("Moral example").

[3] For non-morality related quantum choice models see Lipovetsky (2018) and Yu and Jayakrishnan (2018).

Table 1.1: *Discrete choice models with potential in modelling morality.* Behavioural phenomena *shows what the models can capture, that in traditional DCMs are overlooked.* Structure *shows how the models capture it, and* Moral example *describes an example where it is relevant in morality.*

	TTOA	**DFT**	**Quantum**	**Obfuscation**
Behavioural phenomena	The trade-off between two attributes is prohibited in one direction	Decision-making with shifting attention, evolving preference states	Choice behaviour violates classical probability rules, such as law of total probability	People want to hide their preferences rather than reveal them
Structure	Logit model with an additional taboo penalty term	Dynamic process model with random attention-shifting, transition matrix and multivariate normal distribution of choice probabilities	Preferences are in a "superposition" which collapse to a choice based on rules of quantum probability theory	Logit model with additional obfuscation term calculated with information entropy and Bayesian updates
Moral example	Paying money for improving health is acceptable, while profiting from causing injuries is a taboo	Someone decides to lie to a loved one to avoid conflict, but as they face the person, they change their mind	Prisoner dilemma[4]: the probability of choosing to defect without prior information on the other person's choice is significantly smaller than that of calculated from the probabilities of defecting while knowing the other cooperates and defecting while knowing the other defects.	Someone may hide their preferences to avoid social judgement; politicians cut back on welfare policies where it is less transparent who the beneficiaries are

1.1.2. MORAL VALUE TAXONOMIES IN MORAL PSYCHOLOGY

Theories of morality developed in the interdisciplinary field of moral psychology have also potential to represent moral values in DCMs. Moral values can be categorized several ways; using Moral Foundations Theory (MFT, Graham et al., 2009), Morality-as-Cooperation (MAC, Curry et al., 2019), or Schwartz Value Theory (SVT, Schwartz, 1992) to name a few. These taxonomies aim to capture universal aspects of morality, meaning the categories described by them are valued by all human beings, only their extent or manifestation differs across people or cultures. They can be operationalized through questionnaires (e.g. Curry et al., 2019; Graham et al., 2011; Schwartz and Cieciuch, 2021), dictionaries (e.g. Frimer et al., 2019; Graham et al., 2009) or labelling (e.g. Hoover et al., 2020). These methods allow researchers to quantify the relative importance of these values and draw conclusions on political differences (e.g. Haidt and Graham, 2007), cross-cultural morality (e.g. Struch et al., 2002) or context-dependent morality (e.g. Chowdhury, 2021).

Table 1.2: *Major moral value taxonomies from moral psychology; their covered values and ways to operationalize them in empirical research.*

	MFT	**MAC**	**SVT**
Values	(1) care/harm, (2) fairness/cheating, (3) loyalty/betrayal, (4) authority/subversion (5) sanctity/degradation and the later added (6) freedom/oppression	(1) family values, (2) group loyalty, (3) reciprocity, (4) bravery, (5) respect, (6) fairness, and (7) property rights	(1) self-direction, (2) stimulation, (3) hedonism, (4) achievement, (5) power, (6) security, (7) conformity, (8) tradition, (9) benevolence, (10) universalism
Operationalization	Moral Foundations Dictionary and its updates (Araque et al., 2020; Frimer et al., 2019; Graham et al., 2009; Hopp et al., 2021), Moral Foundations Questionnaire (MFQ, Graham et al., 2008), labelled data (Araque et al., 2020; Hoover et al., 2020)	MAC questionnaire (Curry et al., 2019)	Schwartz Value Survey (Schwartz and Cieciuch, 2021), Portrait Values Questionnaire (PVQ, Schwartz et al., 2001), Personal Values Dictionary (Jones et al., 2018)

There are several overlaps among these values; each postulate that, for instance, 'helping your group' or 'respecting superiors' are widely considered to be morally good (Curry et al., 2019). There are a few papers that aim to synthesize these moral theories, SVT and MFT in particular, in order to create a consistent and comparable operationalization of the underlying constructs in such theories (e.g., McNeace and Sinn, 2018; Vaisey and Miles, 2014; Zapko-Willmes et al., 2021). MFT and SVT, being developed decades ago, have been used, validated and criticized several times over the years. Empirical applications operationalize them in several contexts, such as examining political orientation (e.g., Caprara et al., 2009; Federico et al., 2013), ethical consumer behaviour (e.g., Culiberg et al., 2022; Shaw et al., 2005), or cross-cultural comparisons (e.g., Doğruyol et al., 2019; Struch et al., 2002). MAC, a relatively new theory, is put forward as a potentially superior alternative to MFT and SVT; however, it has significantly fewer empirical applications to support or refute this claim.

1.2. RESEARCH GOALS

Quantitative analysis of decision-making proved to be valuable in many fields. Morality is a significant factor in many situations; however, as moral preferences substantially differ from non-moral ones, DCMs currently have a blind spot, with a few notable exceptions detailed in 1.1.1, for moral decision-making analysis.

In the light of this, the research aim of this thesis can be formulated as follows.

To develop and evaluate the potential of new discrete choice modelling methods to identify latent morality, thus increasing the behavioural realism of DCMs in moral decision-making.

1.3. RESEARCH FOCUS AND METHODS

To achieve the research goal stated in section 1.2, I use two approaches.

First, in Part I, I test two recently developed or adopted Discrete Choice Models that relax the assumptions of utility maximization, stable and revealed preferences, thus are promising tools for moral discrete choice analysis (see section 1.1.1). I test whether the behavioural constructs they aim to capture are indeed identifiable from observing choices only.

Second, in Part II, I collect additional data on moral values using standard morality surveys, Likert-scale type contextual questionnaires and natural text, which are then inserted into mainstream discrete choice models. I rely on moral psychology to create different types of attributes, which can quantify moral values one endorses or moral frames one uses in their language. Section 1.1.2 introduces different theories of morality. In this thesis, without claiming that the other theories are incorrect or less useful, I rely on MFT, as it has several ways of operationalization and several empirical studies that can serve as a reference point in the analyses of Part II.

Figure 1.1 shows the outline of the thesis. Chapter 2 and Chapter 3 correspond to the first approach; they examine two recently developed or adapted models that have the potential to give novel behavioural insights into moral decision-making based on only observed choices. Chapter 4 and Chapter 5 correspond to the second approach, as they use additional data to choices on different moral dimensions. Chapter 6 concludes the thesis with scientific reflection, practical implications and future research directions.

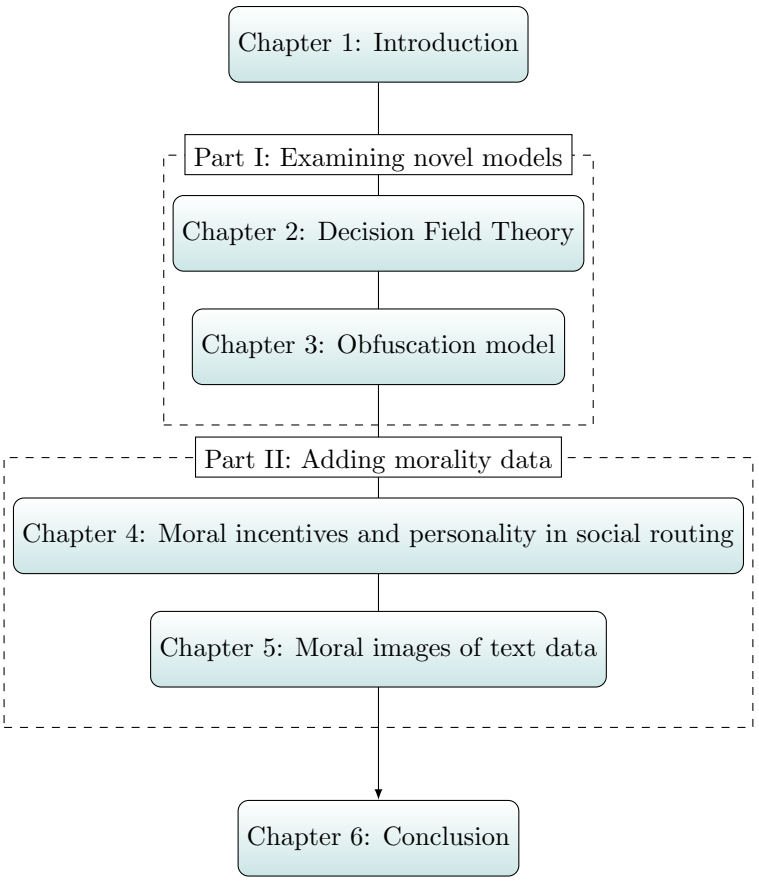


Figure 1.1: *The outline of the thesis*

1.3.1. STUDY 1: DECISION FIELD THEORY: IDENTIFIABILITY, DISTINGUISHABILITY AND EQUIVALENCE WITH PROBIT MODELS

Cognitive processes, such as contemplation or decision-makers changing their minds, serve as a base for Decision Field Theory (DFT). The model of DFT is specifically designed to capture cognitive processes (Busemeyer and Townsend, 1993). It is a dynamic model, and instead of having one utility for each alternative, there is a 'preference value' attached to the alternatives that change over time as the attention of the decision-maker wanders during deliberation. Although the current applications do not involve morality, DFT's features can be connected to cognitive processes or emotions that characterize morality. For instance, a child might decide they are going to lie about something to avoid being told off, but once they are facing their mother, honesty might become more important than avoiding being lectured. According to DFT, a decision is made when the

preference value of an alternative reaches the level of 'good enough' or when the time runs out; in this case, the alternative with the highest preference value at the time will be chosen. This is in line with the satisficing behaviour (Simon, 1957) that often characterizes decisions with moral dimensions.

Research sub-goal:
To test the parameter identifiability in the recent adaptation of Decision Field Theory, which aims to capture the contemplation process in a decision-maker's mind during deliberation.

This study tests the recent adaptation of DFT (Hancock et al., 2018) that allows the estimation of several parameters that characterize a cognitive process in the DCM framework. The study uses mathematical derivations to find that in some cases, these parameters are not jointly identifiable when the analyst has only choice data[5]. Monte Carlo simulations also show that the psychological parameters of the DFT model are biased and that converged models often come with large and infinite standard errors for parameter estimates. It concludes with practices to prevent identification issues.

1.3.2. STUDY 2: IDENTIFICATION OF PREFERENCES UNDER OBFUSCATING BEHAVIOUR

The Obfuscation model postulates that agents may try to hide the preferences underlying their decisions (Chorus et al., 2021). It captures the behaviour of an agent who believes an onlooker is observing them. The assumed onlooker observes the actions made by the agent and updates his beliefs on the agent's underlying motivations based on Bayesian inference. The agents (who wish to hide their preferences consciously or subconsciously) try to obfuscate by maximizing the information entropy resulting from their decision. That means they generate the highest level of uncertainty about the preferences from an observer's perspective. This can be the case when decision-makers are insecure about their preferences or do not want to be judged because of them. This relates to the concept of moral wiggle room, which means that decision-makers might intentionally avoid relevant information, which allows for a wiggle room when an explanation is needed due to the imperfect incentives created by social norms under conditions of uncertainty (Spiekermann and Weiss, 2016). For instance, in the 2015 refugee crisis in Sweden, it has been observed that the more refugees are allocated in a municipality, the more people tend to avoid clicking on the news that may encourage them to welcome the refugees (Freddi, 2015). This can allow for an explanation of "I did not know that help is needed" instead of "It would have been inconvenient" or "I did not really want to help". Due to the decision-maker trying to hide their preferences, the question arises: can the analyst identify their preferences and obfuscating intention? Take an example where a politician has to cast public votes on different policy packages. If the politician uses an obfuscation strategy, the accurate estimates of their preference weights can be used to infer the future behaviour of the politician in public or even non-public voting scenarios. In case the parameters cannot be recovered due to obfuscation, such infer-

[5] This means there is no time measurement or dynamic data such as eye-tracking. This assumption reflects the most common type of data in the field of choice modelling.

ences cannot be made.

Research sub-goal:
To test whether obfuscation behaviour can be identified from observing choices in the recently proposed Obfuscation model.

To achieve this goal, this study uses Monte Carlo simulations and examines whether the true data generating process can be recovered under different levels of obfuscation intention. Furthermore, it proposes an extension to the original model called sequential obfuscation, which allows the agent to obfuscate throughout several choice tasks. Thus, if one decision is 'too revealing', they have a chance to offset it in a subsequent choice task. Again, the study uses Monte Carlo simulations to test whether the true data generating process can be recovered under varying obfuscation intention and a varying number of choice tasks under sequential obfuscation.

1.3.3. STUDY 3: MORAL ASPECTS OF DECISION-MAKERS' INTENTIONS TO PARTICIPATE IN SOCIAL ROUTING SCHEMES

Moral psychology research regarding moral values argued that some basic units of morality transcend contexts (e.g. Haidt and Joseph, 2004). This means, for instance, that if fairness is important to someone when they have to allocate benefits to their employees, it is also important to them when they vote on tax reform. However, recent investigations find the opposite: general moral value measurements are more stable but less predictive of actual behaviour, than contextual moral motivations (Kroesen and Chorus, 2018). Several empirical investigations also find that subtle differences in the presentation of a task in the very same situation can lead to different choices.

Research sub-goal:
To investigate the relationship between generic moral values, contextual moral values and moral decision-making under different incentives to participate in social routing schemes.

To achieve this goal, this study uses a stated intention experiment related to an everyday choice task, namely route choice. The motivation for this is a recent line of literature arguing that moral incentives can help to significantly reduce traffic congestion. The idea behind the so-called *social routing* schemes is that car users voluntarily agree, every once in a while, to choose a different route with higher travel time than their regular route for the benefit of the system at large (Klein et al., 2018; Mariotte, Leclercq, Ramirez, et al., 2021; van Essen et al., 2020). This study addresses an aspect of social routing schemes that have not received much attention: the role of morality. The study adds three dimensions of morality to the discrete choice analysis of the collected data. First is a widely established morality scale (Moral Foundations Questionnaire or MFQ; Graham et al., 2009) to measure the general moral inclinations of travellers. The second is a context-specific set of questions on moral motivations. And the third moral dimension is the nature of the presented scheme, which can be either sacrifice-based or fairness-based. The study estimates Discrete Choice Models with these moral dimensions and finds that general moral values, as well as contextual ones, have similar ex-

planatory power in the outcomes and that, albeit more difficult to implement, fairness-based social routing schemes are more viable in the long run than sacrifice-based ones.

1.3.4. STUDY 4: MORAL IMAGES IN DISCRETE CHOICE MODELS: A NATURAL LANGUAGE PROCESSING APPROACH

Recent work regarding a broad range of latent variables, including latent moral motivations, shows that the joint identification of underlying preferences and other latent determinants of decision-making is a very challenging task (e.g. Vij and Walker, 2016b). Although progress is being made to advance the identification of such models based on choice data and additional closed-ended questionnaires, one obvious potential solution has not received the attention it deserves: the use of additional text data to help identify latent behavioural constructs. One central argument for using text data in the choice analysis is that the nuances that are present in free text often cannot be grasped with standard, closed-ended responses (Baburajan et al., 2020). This is even more relevant when the subjects are abstract and complex phenomena, such as moral values (Boyd et al., 2015).

Research sub-goal:
To investigate the relationship between moral decision-making and moral images projected by decision-makers' natural language use.

This study proposes a method to combine choice- and text data to infer moral motivations in a decision-making situation. Moral features are extracted from text using Natural Language Processing. These features are called moral images and are used as input in Discrete Choice Models. The study presents how this novel approach can lead to new, subtle insights regarding latent motivations of moral choice, which would be very difficult – if not impossible – to obtain using traditional choice models based on observed choices only. To test and illustrate the proposed approach, a case study investigates the voting behaviour of Members of the European Parliament. Results indicate that moral images have significant explanatory power in modelling voting behaviour. Behavioural insights are presented in the light of political science literature, and potential routes for future investigations conclude.

REFERENCES

Araque, O., Gatti, L., & Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, *191*, 105184.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106.

Ashok, K., Dillon, W. R., & Yuan, S. (2002). Extending discrete choice models to incorporate attitudinal and other latent variables. *Journal of marketing research*, *39*(1), 31–46.

Baburajan, V., e Silva, J. d. A., & Pereira, F. C. (2020). Open-ended versus closed-ended responses: A comparison study using topic modeling and factor analysis. *IEEE Transactions on Intelligent Transportation Systems*, *22*(4), 2123–2132.

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges*, *2002*, 431–470.

Bolduc, D., Boucher, N., & Alvarez-Daziano, R. (2008). Hybrid choice modeling of new technologies for car choice in canada. *Transportation Research Record*, *2082*(1), 63–71.

Boyd, R., Wilson, S., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1).

Braga, J., & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics*, *32*(1), 55–89.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432.

Caprara, G., Vecchione, M., & Schwartz, S. H. (2009). Mediational role of values in linking personality traits to political orientation. *Asian Journal of Social Psychology*, *12*(2), 82–94.

Chorus, C. G., & Kroesen, M. (2014). On the (im-) possibility of deriving transport policy implications from hybrid choice models. *Transport Policy*, *36*, 217–222.

Chorus, C. G., Pudāne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: A discrete choice model and empirical analysis. *Journal of choice modelling*, *27*, 37–49.

Chorus, C. G., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. *Mathematical Social Sciences*, *109*, 28–44.

Chowdhury, R. M. (2021). The ethics of nudging: Using moral foundations theory to understand consumers' approval of nudges. *Journal of Consumer Affairs*.

Culiberg, B., Cho, H., Kos Koklic, M., & Zabkar, V. (2022). The role of moral foundations, anticipated guilt and personal responsibility in predicting anti-consumption for environmental reasons. *Journal of Business Ethics*, 1–17.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In T. K. Shackelford & R. D. Hansen (Eds.), *The evolution of morality* (pp. 27–51). Springer International Publishing.

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality, 78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Deaton, A., & Muellbauer, J. (1980). *Economics and consumer behavior*. Cambridge university press.

Doğruyol, B., Alper, S., & Yilmaz, O. (2019). The five-factor model of the moral foundations theory is stable across weird and non-weird cultures. *Personality and Individual Differences, 151*, 109547.

Federico, C. M., Weber, C. R., Ergun, D., & Hunt, C. (2013). Mapping the connections between politics and morality: The multiple sociopolitical orientations involved in moral intuition. *Political Psychology, 34*(4), 589–610.

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

Freddi, E. (2015). Do people avoid morally relevant information? evidence from the refugee crisis. *Exploiting the unexpected in inflow of refugees to Sweden during*.

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*.

Gert, B. (2004). *Common morality: Deciding what to do*. Oxford University Press.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science, 2*(3), 528–554.

Gonzalez-Valdes, F., Heydecker, B. G., & Ortúzar, J. D. D. (2022). Quantifying behavioural difference in latent class models to assess empirical identifiability: Analytical development and application to multiple heuristics. *Journal of Choice Modelling*, 100356.

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.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Spassena, K., & Ditto, P. H. (2008). Moral foundations questionnaire. *Journal of Personality and Social Psychology*.

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

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. *Daedalus, 133*(4), 55–66.

Hancock, T. O., Broekaert, J., Hess, S., & Choudhury, C. F. (2020). Quantum choice models: A flexible new approach for understanding moral decision-making. *Journal of choice modelling*, *37*, 100235.

Hancock, T. O., Hess, S., & Choudhury, C. F. (2018). Decision field theory: Improvements to current methodology and comparisons with standard choice modelling techniques. *Transportation Research Part B: Methodological*, *107*, 18–40.

Hensher, D. A., & Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, *30*(2), 133–176.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, *53*(1), 232–246.

Israel, S., Hasenfratz, L., & Knafo-Noam, A. (2015). The genetics of morality and prosociality. *Current Opinion in Psychology*, *6*, 55–59.

Jones, K. L., Noorbaloochi, S., Jost, J. T., Bonneau, R., Nagler, J., & Tucker, J. A. (2018). Liberal and conservative values: What we can learn from congressional tweets. *Political Psychology*, *39*(2), 423–443.

Klein, I., & Ben-Elia, E. (2018). Emergence of cooperative route-choice: A model and experiment of compliance with system-optimal ATIS. *Transp. Res. Part F Traffic Psychol. Behav.*, *59*, 348–364.

Klein, I., Levy, N., & Ben-Elia, E. (2018). An agent-based model of the emergence of cooperation and a fair and stable system optimum using atis on a simple road network. *Transportation research part C: emerging technologies*, *86*, 183–201.

Kroesen, M., & Chorus, C. G. (2018). The role of general and specific attitudes in predicting travel behavior–a fatal dilemma? *Travel behaviour and society*, *10*, 33–41.

Leong, W., & Hensher, D. A. (2012). Embedding decision heuristics in discrete choice models: A review. *Transport Reviews*, *32*(3), 313–331.

Lipovetsky, S. (2018). Quantum paradigm of probability amplitude and complex utility in entangled discrete choice modeling. *Journal of choice modelling*, *27*, 62–73.

Lovreglio, R., Fonzone, A., & Dell'Olio, L. (2016). A mixed logit model for predicting exit choice during building evacuations. *Transportation Research Part A: Policy and Practice*, *92*, 59–75.

Mariotte, G., Leclercq, L., Gonzalez Ramirez, H., Krug, J., & Bécarie, C. (2021). Assessing traveler compliance with the social optimum: A stated preference study. *Travel Behaviour and Society*, *23*, 177–191.

Mariotte, G., Leclercq, L., Ramirez, H. G., Krug, J., & Becarie, C. (2021). Assessing traveler compliance with the social optimum: A stated preference study. *Travel behaviour and society*, *23*, 177–191.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

McNeace, M., & Sinn, J. (2018). Moral foundations theory vs. schwartz value theory: Which theory best explains ideological differences? *The Winthrop McNair Research Bulletin*, *4*(1), 6.

Plott, C. R. (1993). Rational individual behavior in markets and social choice processes.

Raveau, S., Yáñez, M. F., & de Dios Ortúzar, J. (2012). Practical and empirical identifiability of hybrid discrete choice models. *Transportation Research Part B: Methodological*, *46*(10), 1374–1383.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology* (pp. 1–65). Elsevier.

Schwartz, S. H., & Cieciuch, J. (2021). Measuring the refined theory of individual values in 49 cultural groups: Psychometrics of the revised portrait value questionnaire. *Assessment*, 1073191121998760.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of cross-cultural psychology*, *32*(5), 519–542.

Shaw, D., Grehan, E., Shiu, E., Hassan, L., & Thomson, J. (2005). An exploration of values in ethical consumer decision making. *Journal of Consumer Behaviour: An International Research Review*, *4*(3), 185–200.

Simon, H. A. (1957). Models of man; social and rational.

Søberg, M., & Tangerås, T. P. (2007). Voter turnout in small referendums. *Electoral Studies*, *26*(2), 445–459.

Spiekermann, K., & Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of 'moral wiggle room'. *Games and Economic Behavior*, *96*, 170–183.

Struch, N., Schwartz, S. H., & Van Der Kloot, W. A. (2002). Meanings of basic values for women and men: A cross-cultural analysis. *Personality and social psychology bulletin*, *28*(1), 16–28.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., Herrmann, E., Gilby, I. C., Hawkes, K., Sterelny, K., Wyman, E., Tomasello, M., et al. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current anthropology*, *53*(6), 000–000.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual review of psychology*, *64*, 231–255.

Tversky, A., & Thaler, R. H. (1990). Anomalies: Preference reversals. *Journal of Economic Perspectives*, *4*(2), 201–211.

Vaisey, S., & Miles, A. (2014). Tools from moral psychology for measuring personal moral culture. *Theory and society*, *43*(3), 311–332.

van Essen, M., Thomas, T., van Berkum, E., & Chorus, C. G. (2020). Travelers' compliance with social routing advice: Evidence from SP and RP experiments. *Transportation*, *47*(3), 1047–1070. https://doi.org/10.1007/s11116-018-9934-z

Van Essen, M., Thomas, T., van Berkum, E., & Chorus, C. G. (2016). From user equilibrium to system optimum: A literature review on the role of travel information, bounded rationality and non-selfish behaviour at the network and individual levels. *Transport reviews*, *36*(4), 527–548.

Vij, A., & Walker, J. (2016a). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, *90*, 192–217.

Vij, A., & Walker, J. (2016b). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, *90*, 192–217.

Walker, J. (2002). Mixed logit (or logit kernel) model: Dispelling misconceptions of identification. *Transportation Research Record*, *1805*(1), 86–98.

Walker, J., Ben-Akiva, M., & Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics*, *22*(6), 1095–1125.

Yu, J. G., & Jayakrishnan, R. (2018). A quantum cognition model for bridging stated and revealed preference. *Transportation Research Part B: Methodological*, *118*, 263–280.

Zapko-Willmes, A., Schwartz, S. H., Richter, J., & Kandler, C. (2021). Basic value orientations and moral foundations: Convergent or discriminant constructs? *Journal of Research in Personality*, *92*, 104099.

I

IDENTIFICATION OF PARAMETERS USING CHOICE DATA ALONE

2

DECISION FIELD THEORY: IDENTIFIABILITY, DISTINGUISHABILITY AND EQUIVALENCE WITH PROBIT MODELS

Part I of this thesis examines parameter identifiability and recoverability in novel Discrete Choice Models. This chapter concerns a model that has the potential to capture contemplation in a decision-maker's mind, which is highly relevant to moral dilemmas. Decision Field Theory is a dynamic cognitive process model recently adapted to Discrete Choice Modelling. In Discrete Choice Modelling, the most common data is only choice; it is not common to obtain data on intermediate steps of deliberation or measure deliberation time.

This chapter examines whether the process parameters that capture preference updating time steps, memory, and sensitivity are identifiable in DFT relying on only choice data. It uses analytical derivations to first find in what cases are DFT models equivalent to probit models. Then, established methods designed for probit models are applied to examine parameter identifiability in these cases. Section 2.1 introduces the motivation and related literature, section 2.2 gives a detailed, formal introduction on the recently adapted version of DFT. Section 2.3 introduces probit models formally and describes their relevant identifiability steps. Section 2.4. presents the results: first the equivalence between DFT and probit models, then the parameter identifiability and recoverability findings. Section 2.5 provides an overview of DFT model specifications that avoid these issues and avenues for further research.

A shortened version of this chapter is accepted for publication as research note in the Journal of Choice Modelling entitled 'Decision Field Theory: equivalence with probit models and guidance for identifiabilty' by Teodóra Szép, Sander van Cranenburgh, and Caspar Chorus (Szép et al., 2022). The research note contains (in full length or in a shortened version) sections 2.1, 2.2, 2.3, 2.4.1,2.4.2, 2.4.3, 2.5 and appendices 2.A, 2.B, 2.C,2.D, 2.F.

2.1. INTRODUCTION

Decision Field Theory (DFT) models are dynamic cognitive *process* models that have been used in mathematical psychology for over two decades (Busemeyer and Townsend, 1993) and have been widely popular. They have been used to analyse monetary gambles (Scheibehenne et al., 2009; Hey et al., 2010), risk-taking in sports (Raab and Johnson, 2004), and consumer decisions (Noguchi and Stewart, 2014; Berkowitsch et al., 2014), for instance. Recently, DFT has caught the travel behaviour research community's attention and several contributions have been made to adapt it to the field of discrete choice analysis, which is the most widely used method to study choice behaviour in Transportation (Hancock, Hess, and Choudhury, 2018; Hancock, Hess, and Choudhury, 2018). DFT models are put forward as a more behaviourally rich alternative to conventional Discrete Choice Models (DCMs) (Busemeyer, Rieskamp, et al., 2014). The DFT model consists of three main ingredients. Firstly, weight parameters that are associated with attributes are similar to the taste parameters in conventional DCMs. Secondly, psychological parameters represent deliberation processes in the decision maker's mind. Specifically, a memory parameter captures how the previous state of preference for an alternative affects the current one, and a sensitivity parameter captures how the presence and performance of an alternative affects the decision maker's preference for another one. The third ingredient is the timestep parameter that stands for the number of times the decision maker updates their state of mind during deliberation.

Despite the fact that the DFT model has been around for over two decades and has been widely cited and used in an abundance of studies into choice behaviour, its inner workings and econometric properties are not yet fully understood. In particular, it is unclear whether the model's parameters are identifiable. Since the model is put forward as being able to capture a psychological (decision-making) *process*, based only on data concerning observed final decision *outcomes*[1], it is a crucial question whether the parameters that are representing this process are in fact identifiable. The identifiability of a model means that there are no two different sets of parameters capable of being estimated that give the same probability distribution function on any data; this notion has also been called observational equivalence (e.g. Rothenberg, 1971). In the context of choice behaviour, this implies that there are no two sets of parameters that generate the same choice probabilities for choice alternatives in the data set. It is widely accepted that in a case where a model tries to reconstruct meaningful state variables that cannot be measured directly (e.g. memory or preference), identifiability and distinguishability (a closely related concept, see Section 2.3.2 for more details) are crucial for drawing

[1]There are exceptions that also take things like eye-tracking data into account (Noguchi and Stewart, 2014).

meaningful behavioural conclusions from the data (Walter and Pronzato, 1996). In existing DCM literature, conditions for identifiability have been addressed extensively and thoroughly (e.g. Bunch, 1991, Walker, 2002), highlighting that identification issues can lead to biased estimates in choice models and a loss of model fit (Walker, 2002). Therefore, for DFT models to become a viable addition to the travel behaviour researcher's toolbox, in-depth understanding on their identifiability is compulsory.

This paper investigates the identifiability and distinguishability of DFT models using two methods: analytical derivations and Monte Carlo experimentation. In order to obtain analytical results, first we show that four DFT specifications can be recast as special cases of a probit model, one of the classic DCMs. More generally, we derive the conditions under which the theoretical equivalence of DFT and probit models holds. This enables us to build on the existing, well-developed literature on identifiability in Discrete Choice Theory in order to obtain robust results concerning the identifiability of the DFT model. This method (i.e. establishing conditions for equivalence between a process model and a classical DCM, and subsequently using identifiability results from discrete choice theory to obtain corresponding results for the process model) has not been applied previously in existing literature for process models. We consider it to be a promising avenue that may help pave the way towards incorporating alternative models from mathematical psychology to the transportation modelling domain. Applying our method to DFT, we have found four cases where the probit equivalence holds: (1) when there is just one timestep, (2) when the sensitivity parameter is relatively high, (3) when the memory decay parameter is zero and (4) when there are only two alternatives. If one of these conditions is met, then the DFT model can be considered a probit model with a particular structure of its covariance matrix. For special cases (1)-(3), we were able to draw conclusions concerning the identifiability and distinguishability of the DFT model, using analytic derivations which capitalize on the probit equivalence. In case (4), and in the cases of other, more generic specifications of DFT models, we applied Monte Carlo simulation to explore identification problems. In our analytical derivations, we found that the high sensitivity case is unidentifiable, and is indistinguishable from the zero memory decay case. Using the Monte Carlo experiments we found that unrestricted, DFT models applied in binary and multi-nominal choice situations also exhibit identification problems (specifically, we show that the estimated psychological parameters are biased in such contexts). This identification issue is likely to result in misguided behavioural inferences when interpreting estimated DFT process parameters.

The remaining part of the paper is organized as follows. Section 2.2 introduces the DFT model in full detail: we focus on the building blocks of the model, following the specification and notation most often used in transportation literature. Section 2.3 briefly introduces conventional discrete choice theory and probit models, with particular attention to the literature on structured covariance matrices and the notions of identifiability and distinguishability. We present our results in Section 2.4. Here, we establish the theoretical equivalence between special cases of DFT and probit models, and study the DFT model's identifiability and distinguishability issues using analytical derivations for the special cases, and using Monte Carlo experiments for non-restricted DFT models. Sec-

tion 2.5 discusses the implications of our results in terms of the estimation of DFT models and, more generally, the making of behavioural inferences based on process models which are estimated on outcome data.

2.2. DECISION FIELD THEORY

DFT was developed by Busemeyer and Townsend (1992) and over the last few decades it has had several variations applied in different fields for different problems. In this section we introduce the variation that has appeared in the transportation and choice modelling domain in recent years, following the developments and notation of Hancock, Hess, and Choudhury (2018). DFT is a process-oriented approach to describing human decision making. It assumes that from the moment decision makers face a choice task, their attention wanders from one attribute to the other. That means, when the decision maker has to choose between travel modes when planning a trip, they might first think about the travel time, then the price or convenience or other qualities of the mode, and then their attention may wander back to the travel time, and so on. In DFT this translates to an attention vector, that has a size of $m \times 1$ where m is the number of attributes. The vector is a zero-vector with one element set to 1 at each timestep[2] t, that corresponds to the attribute being focused on. The attention of each attribute i follows a Bernoulli-distribution, being 1 with probability w_i. This is called the weight of the attribute and also forms a vector of size $m \times 1$.

Therefore, at each timestep, one attribute (ith attribute) is being focused on, which creates a momentary valence V in the decision maker's mind. The valence of an alternative j is the difference between x_{ij} and the mean of $x_{i,k\neq j} \ \forall k$. x_{ij} stands for the value of the ith attribute of alternative j. It is also assumed that the individual's momentary preference (valence) at each timestep has a random component, which is represented by an added error term in the model. Mathematically, this is formulated as

$$V_t = CMW_t + \varepsilon_t \tag{2.1}$$

where C is a contrast matrix of size $n \times n$ with 1 on the diagonal and $-\frac{1}{n-1}$ on the off-diagonal, with n being the number of alternatives. M is a matrix of the attributes of alternatives (x_{ij}), where each row corresponds to an alternative and each column to an attribute. W_t is the $m \times 1$ attention vector at timestep t, all zeros with one element (i.e. the one corresponding to the attribute being focused on in that timestep) being one. ε is the error term vector of size $n \times 1$. ε is distributed identically, independently ,across alternatives, individuals and timesteps, following a normal distribution with zero mean and s variance. The valence therefore corresponds to a momentary preference that results from a comparison between one alternative and all the others on a single attribute.

[2] Note that although the timestep parameter can be a function of real time (e.g. Hancock, Hess, and Choudhury, 2018) in this paper parameter t always refers to the number of updating timesteps in a decision maker's mind.

When receiving the choice task, the decision maker is assumed to have an initial preference value, represented by P_0. At the first timestep, $t = 1$, focus is on one of the attributes, and a momentary valence with an error term is calculated. Then this valence is added to the previous preference state P_0 multiplied by a so-called feedback matrix (equation 2.2) to get P_1. Then again, at $t = 2$, one of the attributes is being focused on, a new valence is calculated, which is then added to P_1, multiplied by the feedback matrix. Figure 2.1 illustrates an example of the evolution of P_t over time, based on a DFT process.

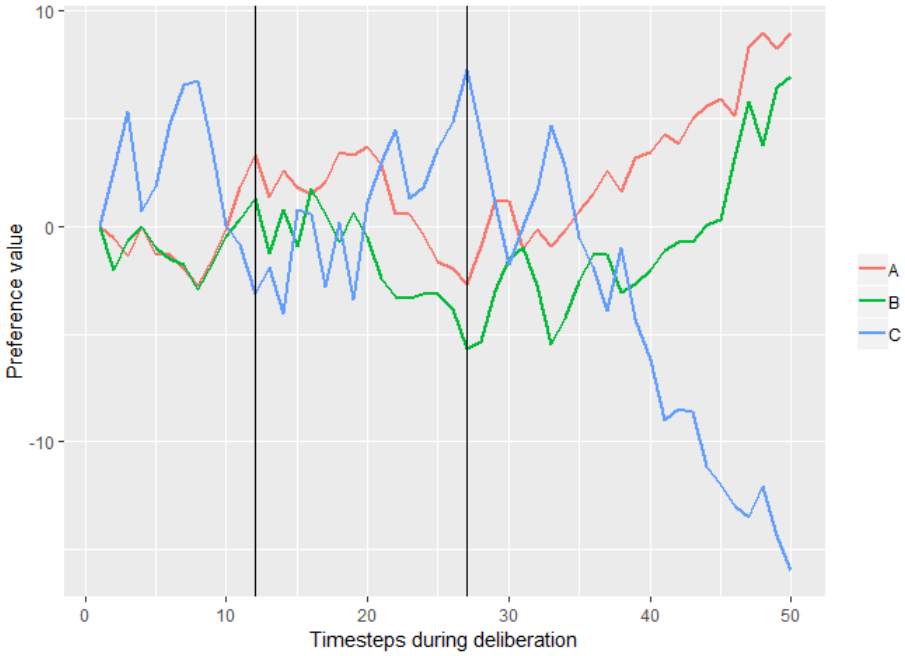


Figure 2.1: *The horizontal axis shows the timesteps of deliberation, the vertical axis shows the preference value. Three competing alternatives are plotted. The vertical black lines illustrate that at different timesteps different alternatives have the highest preference value.*

The feedback matrix can be understood as the effect of the previous state of mind on the current one. It contains the memory (ϕ_2) and sensitivity (ϕ_1, the effect of one alternative on the other) parameters. Its form following Hotaling et al. (2010)[3] using element-wise notation is:

$$S = I - \phi_2 \cdot \exp(-\phi_1 \cdot D^2) \tag{2.2}$$

where the D matrix contains the Eucledian distance between the alternatives in the multi-attribute space and I is the the identity matrix of size D ($n \times n$). The feedback

[3] It is possible to use other specifications for the feedback matrix, however this is the most common one used in choice modelling literature, so we have therefore used it in this paper. Our results are generalised for other specifications as well, see Appendix 2.A for the conditions on a general feedback matrix.

matrix therefore is a symmetrical matrix, which reduces to a diagonal matrix if $\exp(-\phi_1 \cdot D^2)$ is very close to zero, and to an identity matrix, if ϕ_2 is zero.

The preference value at any timestep can be written in recursive form as:

$$P_t = S \cdot P_{t-1} + V_t \tag{2.3}$$

The DFT model postulates that at any timestep the alternative with the highest preference value will be chosen. This can be reached two ways: either one of the alternatives crosses an internal threshold (meaning the alternative will be "good enough", and therefore deliberation stops) or the deliberation stops first (due to internal or external pressure) and the alternative that is leading at that moment will be chosen. Following on from existing transportation literature (Hancock, Hess, and Choudhury, 2018; Hancock, Hess, and Choudhury, 2018), we examine the second kind of model where the deliberation stops at some point in time. In other words, the decision maker is not taking any more timesteps to update their preference. In order to compute choice probabilities for alternative i at timestep t, the following formula applies:

$$Pr[P_{ti} - P_{tj} > 0 \; \forall \, j \neq i] = \int_{X>0} \frac{exp\left[-\frac{1}{2}(X-\Gamma_i)'\Lambda_i^{-1}(X-\Gamma_i)\right]}{\sqrt{(2\pi)^{n-1}|\Lambda_i|}} \, dX \tag{2.4}$$

X follows the multivariate normal distribution with Γ_i mean and Λ_i covariance matrix. Γ_i is the vector of expected differences between P_{ti} and all the other alternatives, Λ_i is the corresponding covariance matrix. Both Γ_i and Λ_i are constructed with the help of an L_i matrix, which is constructed in the following way: we insert a column vector of 1s as the ith column of a diagonal matrix (of size $n-1 \times n-1$) with -1s on the diagonal. This gives us an $n-1 \times n$ matrix. For example, for 3 alternatives when i is 1:

$$L_1 = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

The mean of equation 2.4 therefore is

$$\Gamma_i = L_i \xi_t \tag{2.5}$$

and the covariance matrix is

$$\Lambda_i = L_i \Omega_t L_i' \tag{2.6}$$

The expected preference value at timestep t is denoted as ξ_t and calculated as:

$$\begin{aligned} \xi_t &= \sum_{k=0}^{t-1} S^k \mu + S^t \cdot P_0 \\ &= (I-S)^{-1}(I-S^t)\mu + S^t \cdot P_0 \end{aligned} \tag{2.7}$$

where μ is the expected valence

$$\mu = CMw \tag{2.8}$$

with w being the weight-vector, containing the probabilities that each attribute is being focused on at any moment in time.

The corresponding covariance matrix is[4]

$$\begin{aligned}\Omega_t &= \sum_{k=0}^{t-1} S^k \Phi S^{k'} \\ &= (I - Z)^{-1}(I - Z^t)\bar{\Phi}\end{aligned} \tag{2.9}$$

where $\bar{\Phi}$ is the vectorized form of Φ, the covariance of the valence, which is

$$\Phi = CM\Psi M'C' + s \tag{2.10}$$

where Ψ is the covariance of the attention vector W.

Z is the Kronecker-product[5] of S with itself:

$$Z = S \otimes S \tag{2.11}$$

2.3. PROBIT MODELS, IDENTIFIABILITY AND DISTINGUISHABILITY

2.3.1. PROBIT MODELS

The standard practice to analyse preferential choice between discrete alternatives is the Discrete Choice Model family (McFadden, 1973), predominantly used with a linear in parameters utility maximizing assumption. One of the most widely used DCMs is the probit model. Probits estimate the relative importance of the attributes, the weights that represent the preferences of the decision maker. According to probit the utility of each alternative is calculated based on the attributes of the alternatives and the attached weights that stand for underlying preferences. The systematic part of the utility is noted as V, and for alternative i can be written as:

$$V_i = \sum_{m}^{M} \beta_m x_{mi} \tag{2.12}$$

where x_{mi} denotes the mth attribute of alternative i, and β_m the taste parameter for attribute m. DCMs postulate that the alternative with the highest utility is going to be chosen, formally:

$$Pr[i] = Pr[V_i + \varepsilon_i > V_j + \varepsilon_j \; \forall j \neq i] \tag{2.13}$$

The probability formula of a choice depends on the error term (ε) that is added to the systematic utility. In a probit model, this error term has a multivariate normal distribution, therefore the probability is

$$Pr[i] = \int I(\widetilde{V_i} + \widetilde{\varepsilon}_i > 0)\phi(\widetilde{\varepsilon}_i)d\widetilde{\varepsilon}_i \tag{2.14}$$

[4] Details of the derivation of the covariance matrix for any number of timesteps can be found in Hancock, Hess, and Choudhury (2018).

[5] Given an $m \times n$ matrix A and an $p \times q$ matrix B, their Kronecker product $C = A \otimes B$, which is a $(mp) \times (nq)$ matrix with elements defined by $c_{\alpha\beta} = a_{ij}b_{kl}$, where $\alpha = p(i-1)+k$ and $\beta = q(j-1)+l$.

where I is a function that evaluates to 1 if the expression in parentheses is true, and to 0 if it is false. $\widetilde{V_i}$ stands for the difference vector of $V_i - V_j$ for all $j \neq i$ V_j, and similarly, $\widetilde{\varepsilon_i}$ stands for the corresponding error term differences (more details in Train, 2009 for example). $\phi(\widetilde{\varepsilon_i})$ is the density function of the distribution of the error term differences:

$$\phi(\widetilde{\varepsilon_i}) = \frac{exp\left[-\frac{1}{2}(\widetilde{\varepsilon_i})'\widetilde{\Omega_i}^{-1}(\widetilde{\varepsilon_i})\right]}{\sqrt{(2\pi)^{n-1}|\widetilde{\Omega_i}|}} \tag{2.15}$$

The elements of the covariance matrix ($\widetilde{\Omega_i}$) of the error term differences are normalized and estimated. There are several considerations that must be taken into account when specifying the probit model (see Daganzo (1979), for example). For computational ease it is possible to assume that the variances are equal across the alternatives. However, with a full covariance matrix it is possible to accommodate any pattern of correlation or heteroskedasticity (Train, 2009). Hausman and Wise (1978) suggest that computational ease is only one aspect for choosing an appropriate covariance matrix, but it is more important to have a plausible behavioural underpinning of it. They argue that the correlation between two errors depends on how far or close are the corresponding alternatives are in their measured characteristics. Their covariance structure is a viable solution to the violations of the independence of any irrelevant alternatives. Following similar behavioural assumptions, Yai et al. (1997) apply a specific covariance structure to capture any overlapped relationship between alternatives in route choice. In this case there is only one parameter estimated in the covariance matrix, which captures how strongly the routes that overlap correlate. Computational expensiveness can also be addressed with covariance matrix structuring. The approach of Bolduc (1992) enables the approximation of general correlation structures in large choice sets. This has been used by Bolduc et al. (1996), for example, to study doctors' location choice when policies are directed to a more balanced distribution of new physicians. To study abortion policies, Alvarez and Brehm (1995) used a scaled heteroskedastic probit to capture respondent heterogeneity.

As probit is one of the most general DCMs (Daganzo, 1979), it is possible to apply it to several different problems, which results in a rich variety of covariance matrix structures in existing literature. This called for rigorous methods to establish the identifiability of the parameters.

2.3.2. IDENTIFIABILITY AND DISTINGUISHABILITY

In existing literature it has been established that identifiability and distinguishability are crucial criteria for a parametric model in order to draw conclusions from the data (e.g. Rothenberg, 1971; Walter and Pronzato, 1996). It is especially important when the parameters are used to reconstruct meaningful concepts, such as preference, that cannot be measured directly. As probit models often use meaningful parameters in the covariance matrix structure to capture theory-driven relationships, it has been extensively studied in the last few decades, with respect to identifiability.

In terms of identifiability, there are two kinds of issues: theoretical and empirical. Theoretical identifiability is concerned with whether the model specification is sufficient

to identify the parameters, while empirical identifiability focuses on whether the data is proper for the model to estimate the parameters (Chiou and Walker, 2007). In this study, we address both issues: theoretical identifiability with analytic derivations, and empirical with Monte Carlo experimentation. Furthermore, we examine a closely related concept of distinguishability to see whether process parameters in a model indeed capture a unique underlying process. In the case of indistinguishable processes, we see that two structurally different processes (which may use different sets of parameters) lead to the exact same observable outcome, and we cannot draw a conclusion about which one was the true data-generating process.

In the following section, we use terminology used by Walter and Pronzato (1996). In a model there are a finite number of structures that can describe the input-output relationship observed in the data at hand. The structures provide an explanation of how a respondent ended up choosing a particular alternative, knowing the attributes of the alternatives. For instance, in a process model, one structure could assume that the output (i.e. the choice) is a result of a single updating step, while another structure could assume that the output is a result of an accumulation of preference over time. The analyst's goal is always to find the best structure and estimate its parameters. The structure that is considered to be the best may depend on what the goal of the modelling is: to understand, predict or control the behaviour of a system. If $\zeta_i(p_i) \equiv \zeta_j(p_j)$ means that structure ζ_i with parameters p_i and structure ζ_j with parameters p_j generate the same input-output combination for any input, the following applies:

- *Identifiability* of a structure means that $\zeta(p_i) \equiv \zeta(p_j)$ if, and only if, $p_i = p_j$.
- *Distinguishability* of structures means that for almost any p_i there is no p_j, such that $\zeta_i(p_i) \equiv \zeta_j(p_j)$.

In the context of DCMs, there are several different ways to establish whether a model is identifiable (Walker et al., 2007). One is by examining the information matrix (the expected second derivatives of the log-likelihood), which can only be applied after a model has been estimated if there is no closed form solution for the probabilities (e.g. Rothenberg, 1971; Walker, 2002). Another method was described by Train (2009), where the identifiability of parameters depends on whether they can be computed from the elements of the normalized covariance matrix of utility differences. A third method is by examining the Jacobian matrix of the covariance matrix of utility differences, following the steps laid out by Bunch (1991) for probit models. In this paper we focus on the method described by Train (2009), as it can always be used and we examine four special cases. The steps are the following:

1. Take the covariance matrix of utility differences;
2. Normalize it;
3. Retrieve the estimable parameters from the elements of the normalized covariance matrix of utility differences.

2.4. RESULTS: EQUIVALENCE, IDENTIFIABILITY AND DISTINGUISHABILITY

In this section we present our results in the following order. In section 2.4.1 we present the specifications needed in a DFT model so that it is equivalent to a heteroskedastic[6] probit model. Then, in section 2.4.2, we examine each of these cases by applying the analytical procedure derived for the probit model identifiability by Train (2009). After that, in section 2.4.3 we present our approach to establish distinguishability. Finally in section 2.4.4, we use Monte Carlo experiments to see if DFT models in general (i.e. also beyond the cases where they are equivalent to probit) are able to recover the data generating parameters in an unbiased way.

2.4.1. EQUIVALENCE OF DFT AND PROBIT MODELS

First we show the general conditions that must hold in order to get equivalent DFT and probit models. As both of them use the integral of the multivariate normal distribution's probability density function to compute choice probabilities, they can be formulated as:

$$Pr[i] = \int_0^\infty f(X)dX \tag{2.16}$$

where $f(X)$ stands for the probability density function of the multivariate normal distribution. Variable X in the DFT model is

$$X^{DFT} \sim N(\widetilde{\xi}_i, \Lambda_i) \tag{2.17}$$

where $\widetilde{\xi}_i$ is the vector of differences between the expected preference value of alternative i and that of each of the other alternatives. Λ_i is the corresponding covariance matrix (for details on its formulation, see Section 2.2).

In probit X is

$$X^{probit} \sim N(\widetilde{V_i}, \Omega_i^*) \tag{2.18}$$

where $\widetilde{V_i}$ is the vector of differences between the systematic utility of alternative i and that of each of the other alternatives. Ω_i^* is the corresponding covariance matrix (for details, see Section 2.3).

Therefore when

$$\widetilde{\xi}_i = \widetilde{V_i} \tag{2.19}$$

and

$$\Lambda_i = \Omega_i^* \tag{2.20}$$

the two models are equivalent. The condition on the covariance matrices (equation 2.20) can be obtained with structural assumptions on the probit's covariances. Probit provides

[6] Note that heteroskedacticity here (contrary to many DCM applications, where the variance of the unobserved factors vary across alternatives) means that the covariance matrix varies across choice scenarios.

a flexible framework in which structured covariance matrices are often used to capture interdependencies between alternatives (see Section 2.3), and DFT's behavioural assumptions result in such covariances (see Section 2.2). Therefore we can impose the same structure as that of DFT on the elements of the probit covariance matrix. This structural restriction on probit results in a different covariance matrix in each choice scenario (as the covariance matrix of DFT depends on the attribute differences of alternatives), therefore a probit model that is equivalent to DFT is heteroskedastic.

As for the mean, the well-known utility difference of probit (right-hand-side of equation 2.19) can be written in the linear-additive form of the betas multiplied by the corresponding attribute differences, or formally:

$$V_i - V_j = \sum_m \beta_m (x_{im} - x_{jm}) \forall\, j \neq i \tag{2.21}$$

In the following subsections we show that the left-hand-side of equation (i.e. condition) 2.19, in several cases (see Table 2.1), takes the form of a vector with elements:

$$\xi_i - \xi_j = \pi \sum_m w_m (x_{im} - x_{jm}) \forall\, j \neq i \tag{2.22}$$

π stands for a *scale term*, (that includes the psychological parameters and time) that we multiply the weighted sum of the attribute differences by. For the derivations to find which specifications can be written in this form (equation 2.22), see Appendix 2.A. Note that here we ignore the initial preference value (P_0) as in several applications it is not estimated. Its effects are discussed in Section 2.5. As the probabilities are calculated using the multivariate normal distribution function, we can eliminate this scale term from the mean by multiplying the covariance matrix by the scale term's inverse square, π^{-2}. This means, that in the special cases that we focus on, the mean of the MVN (equation 2.17) takes the linear additive form of attribute differences multiplied by their weights (similar to probit: equation 2.21), and the covariance matrix of the MVN (equation 2.17) is structured by the assumed underlying process of attention wandering, time, and psychological parameters.

Below, we introduce the four cases of DFT which correspond to probit models with linear additive utility and structured, heteroskedastic covariance matrix. These four cases are the following: 1) when there is only one timestep (i.e. updating is omitted), 2) when the sensitivity parameter is relatively high, 3) when there is a constraint on the memory parameter, or 4) when there are two alternatives in the choice set. If any of these four conditions apply, a DFT model is equivalent to a heteroskedastic, structured covariance probit model. Table 2.1 shows the four cases and indicates what they mean for the four parameters.

Table 2.1: *The different cases of DFT examined in depth. t is the number of timesteps, ϕ_1 is the sensitivity, ϕ_2 is the memory parameter and N is the number of alternatives. α is an arbitrarily small number, D is the distance between alternatives. In Cases 1 and 3 some process parameters drop out of the model (denoted by N/A). N is an integer, all other variables have real values. Parentheses denote open sets.*

	t	ϕ_1	ϕ_2	N
Case 1: One timestep	1	N/A	N/A	$N \geq 2$
Case 2: High sensitivity	$\forall t \in (1,\infty)$	$\phi_1 > -\frac{\ln \alpha}{D^2}$	$\forall \phi_2 \in (0,1)$	$N \geq 2$
Case 3: Zero memory decay	$\forall t \in (1,\infty)$	N/A	0	$N \geq 2$
Case 4: Two alternatives	$\forall t \in (1,\infty)$	$\forall \phi_1 \in \mathbb{R}+$	$\forall \phi_2 \in (0,1)$	2

The following subsection introduces the scale terms for the special cases shown in Table 2.1. In the first case ($t = 1$, Section 2.4.2) we elaborate on the derivation steps, while in the rest of the cases we only present the final result for π for the sake of brevity. The steps we use are the same as in the first case, which are the following: we take a special case of DFT (restricted number of timesteps, sensitivity or memory parameter), take the difference in preference values ($\widetilde{\xi}_i$), and bring it to the form of equation 2.22, so that the scale term can be pointed out. Furthermore, we also present the identifiability analysis in each case. Note that for the last case, when there are only two alternatives, the analytic derivation brings no conclusive result (derivations can be found in Appendix 2.C). Therefore the identification of this case is further studied in the Monte Carlo experiments (2.4.4).

2.4.2. IDENTIFIABILITY IN DFT'S SPECIAL CASES

The identifiabilty steps of Train (2009) translates as follows when used in DFT models. First we need to transform our DFT model by multiplying the covariance matrix of the preference value differences by the scale term's inverse square. This way, all parameters will be exclusively in the covariance, except for the weights that appear in the preference value differences, in a similar way to betas in probit utilities. If θs are the estimable parameters, Train argues that in order to set the scale of utility, we need to normalize one of these θs by dividing all θs by it. We show that in DFT this step is unnecessary, as the normalization takes place through normalizing the weights (for proof see Appendix 2.B). Once we have the model of interest converted into utility-differences (hereafter referred to as preference-value differences, in accordance with DFT terminology), and it is normalized (it is by definition), we can examine the elements of the covariance matrices. In three alternative cases[7] we have three θs that can be identified (one more than in a conventional probit):

$$\Lambda = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} \\ \cdot & \theta_{2,2} \end{bmatrix}$$

and we examine how these θs relate to the estimable parameters, i.e. whether the psychological parameters and timesteps can be retrieved from these θs.

[7] We use three alternative cases for illustration purposes, extension to more alternatives is straightforward.

CASE 1: ONE TIMESTEP

The first special case is when there is only one timestep. This case was referred to as being nested within the probit model by Berkowitsch et al. (2014). As we can omit the feedback matrix, equation 2.7 reduces to:

$$\xi_1 = \mu \tag{2.23}$$

and 2.9 to:

$$\Omega_1 = \Phi \tag{2.24}$$

μ stands for the expected valence, that can can be written as

$$\mu_i = \sum_m w_m \left(X_{im} - \frac{1}{N-1} \sum_{n \neq i} X_{nm} \right) \tag{2.25}$$

for the ith alternative. N is the number of alternatives in the choice set. Equation 2.5 represents the expected preference value differences between the alternatives, or in other words a vector with elements:

$$\begin{aligned}
\xi_i - \xi_j = \mu_i - \mu_j &= \sum_m w_m \left(X_{im} - \frac{1}{N-1} \sum_{n \neq i} X_{nm} \right) - \sum_m w_m \left(X_{jm} - \frac{1}{N-1} \sum_{n \neq j} X_{nm} \right) \\
&= \sum_m w_m \left((X_{im} - X_{jm}) - \frac{1}{N-1} (X_{jm} - X_{im}) \right) \\
&= \sum_m w_m \left((X_{im} - X_{jm}) + \frac{1}{N-1} (X_{im} - X_{jm}) \right)
\end{aligned}$$

$$\xi_i - \xi_j = \sum_m w_m \left(1 + \frac{1}{N-1} \right) (X_{im} - X_{jm}) \tag{2.26}$$

which, with

$$\pi = \left(1 + \frac{1}{N-1} \right) \tag{2.27}$$

takes the form of (2.22). Therefore, it is equivalent to a heteroskedastic, structured covariance probit model. In this case, DFT and probit differ in terms of structure, however no additional parameters (for memory, sensitivity or time) are necessary for them to be equivalent.

In order to establish identifiability in this special case, we take the vectorized form of Λ (see Equations 2.6 and 2.9-2.11 form Section 2.2):

$$\bar{\Lambda} = \pi^{-2} \cdot [(LCM \otimes LCM)\bar{\Psi} + (L \otimes L)\bar{s}] \tag{2.28}$$

$\bar{\Psi}$ includes the weight parameters ($w_1, w_2, ..., w_{m-1}$, m assumed to be 2), $\bar{s}$ includes the error term's variance (σ). π contains the number of alternatives (N, in this illustration,

being 3) and none of the estimable parameters in this case.

This covariance matrix can be written in the following expanded form:

$$\bar{\Lambda} = \begin{pmatrix} \frac{8\sigma}{9} - (w_1 - 1)\, w_1 \left(X_{1,2}\right)^2 & \frac{4\sigma}{9} - (w_1 - 1)\, w_1 \left(X_{1,2}\right)\left(X_{1,3}\right) \\ \cdot & \frac{8\sigma}{9} - (w_1 - 1)\, w_1 \left(X_{1,3}\right)^2 \end{pmatrix}$$

where

$$X_{i,j} = \left(x_{i,1} - x_{i,2} - x_{j,1} + x_{j,2}\right)$$

There is only one parameter to be identified by this covariance matrix, and that is σ which, for example, can be expressed by the first estimable parameter $\theta_{1,1}$ as:

$$\sigma = \frac{9 \cdot (\theta_{1,1} + (w_1 - 1)\, w_1 \left(X_{1,2}\right)^2)}{8} \tag{2.29}$$

Therefore the model is identifiable.

CASE 2: HIGH SENSITIVITY

The second special case is when the feedback matrix of DFT is diagonal. In the feedback matrix parametrisation developed by Hotaling et al. (2010) and used for transport data analysis in several cases (by e.g. Hancock, Hess, and Choudhury, 2018), this means that the sensitivity parameter is relatively high (i.e. $e^{-\phi_1 D^2}$ is very close to zero). As D is the distance between alternatives in the multi-attribute space, the sensitivity parameter's size (whether or not it can be considered "high") depends on the data. However, if we establish that it is high, the following applies.

For any number of timesteps and alternatives, the scale term is:

$$\pi_t = \frac{1 - (1 - \phi_2)^t}{\phi_2} \left(1 + \frac{1}{N - 1}\right) \tag{2.30}$$

Multiplying the covariance matrix by π_t^{-2} results in a structured covariance probit model. The covariance matrix includes parameters for time and memory. In this case, the covariance matrix takes the form of:

$$\bar{\Lambda}_t = \pi_t^{-2} \cdot [(I - S \otimes S)^{-1} (I - S^t \otimes S^t) \times \left((LCM \otimes LCM)\bar{\Psi} + (L \otimes L)\bar{s}\right)] \tag{2.31}$$

which can be expanded to:

$$\bar{\Lambda}_t = \begin{pmatrix} \frac{\phi_2(((\phi_2-1)^2)^t-1)(8\sigma-9(w_1-1)w_1(X_{1,2})^2)}{9(\phi_2-2)((1-\phi_2)^t-1)^2} & \frac{\phi_2(((\phi_2-1)^2)^t-1)(4\sigma-9(w_1-1)w_1(X_{1,2})(X_{1,3}))}{9(\phi_2-2)((1-\phi_2)^t-1)^2} \\ \cdot & \frac{\phi_2(((\phi_2-1)^2)^t-1)(8\sigma-9(w_1-1)w_1(X_{1,3})^2)}{9(\phi_2-2)((1-\phi_2)^t-1)^2} \end{pmatrix}$$

This can be written in the form of a constant term (across the sample) multiplied with a matrix that includes the error term variance, the weights and the attributes of alternatives in a structured form:

$$\bar{\Lambda}_t = \frac{\phi_2\left(1+\left(1-\phi_2\right)^t\right)}{\left(2-\phi_2\right)\left(1-\left(1-\phi_2\right)^t\right)}\begin{pmatrix} \frac{8\sigma-9(w_1-1)w_1\left(X_{1,2}\right)^2}{9} & \frac{4\sigma-9(w_1-1)w_1\left(X_{1,2}\right)\left(X_{1,3}\right)}{9} \\ \cdot & \frac{8\sigma-9(w_1-1)w_1\left(X_{1,3}\right)^2}{9} \end{pmatrix} \tag{2.32}$$

The multiplicative term contains the memory parameter and the timesteps. If we name this term z, we can derive that several combinations of ϕ_2 and t can result in the same z, therefore, these two parameters are not jointly identifiable. Equation 2.33 shows t as a function of ϕ_2, z is constant. As this relationship is not dependent on the data, we can state that the parameter combinations generated by this relationship will give the same input-output combination for any data.

$$t = \frac{\ln\left(\frac{z\phi_2-2z+\phi_2}{z\phi_2-2z-\phi_2}\right)}{\ln\left(1-\phi_2\right)} \tag{2.33}$$

To corroborate this finding, we have plotted the log-likelihood as a function of t and ϕ_2 in Figure 2.2. We find that along several combinations of t and ϕ_2 (along the red line) the log-likelihood is flat, and these (t,ϕ_2) pairs satisfy relation 2.33.

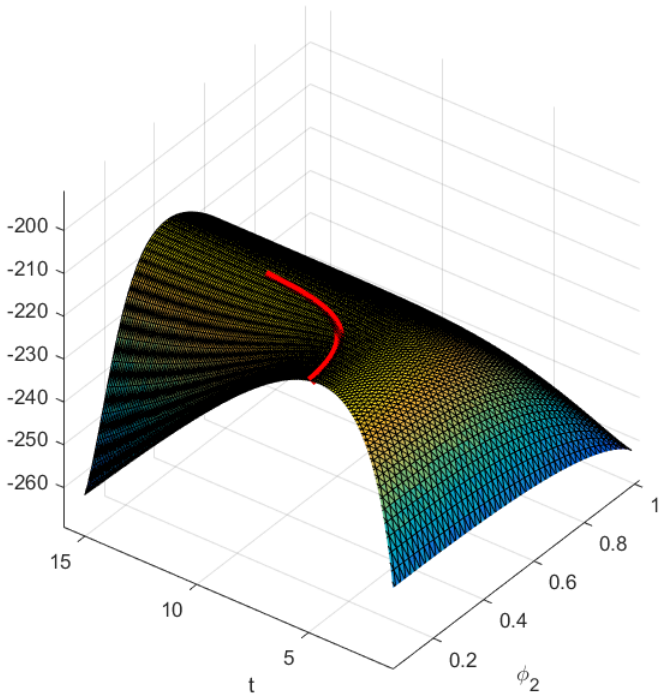


Figure 2.2: *The two horizontal axes show timesteps and memory parameters, and the vertical axis shows the log-likelihood generated by these combinations, all other parameters being equal. The choice data of this specific plot is based on randomly generated attributes between 0 and 1, 2 alternatives and 2 attributes. The choices are generated with DFT ($w_1 = 0.6899745$; $w_2 = 0.3100255$; $\phi_1 = 3$; $\phi_2 = 0.1$; $s = 1$; $t = 5$). The log-likelihood is calculated with the following parameters being fixed: $w_1 = 0.6899745$; $w_2 = 0.3100255$; $\phi_1 = exp(100)$; $s = 1$.*

This shape of the log-likelihood as a function of t and ϕ_2 holds for any data, as long as the estimated sensitivity in the model being examined can be considered high. For different data sets, the scales on the axes might differ, but the shape and the flatness is a characteristic that is independent of the data. Therefore, when the sensitivity is high, the memory and timestep parameters are not jointly identifiable. This issue, however, is most likely to be an empirical issue. Previously, DFT models have not been estimated with a sensitivity parameter *fixed* to a high value, it only becomes one through estimation. Therefore the identifiability issue in this special case is of an empirical nature; and not a theoretical nature.

CASE 3: ZERO MEMORY DECAY

The third special case is when the feedback matrix is an identity matrix. Using the parametrisation of Hotaling et al. (2010), that means that the memory parameter is zero ($\phi_2 = 0$). This also means that the feedback matrix's eigenvalues are equal to 1, therefore the geometric matrix series formula cannot be used in equations 2.7 and 2.9. Instead, the for-

mulas for ξ and Ω (equations 2.7 and 2.9) reduce to:

$$\xi_t = \sum_{k=0}^{t-1} S^k \mu = t \cdot \mu \tag{2.34}$$

$$\Omega_t = \sum_{k=0}^{t-1} [S^k \Phi S^{k'}] = t \cdot \Phi \tag{2.35}$$

We note that these formulas are the same as in the one timestep case, only multiplied by scalar t. The scale term therefore is:

$$\pi_t = \left(1 + \frac{1}{N-1}\right) \cdot t \tag{2.36}$$

When the feedback matrix is an identity matrix, its eigenvalues are 1. Using the parametrisation of Hotaling et al. (2010), this means that the memory decay is zero, and the vector of covariance matrix elements is:

$$\bar{\Lambda}_t = \pi_t^{-2} \cdot t \times \left((LCM \otimes LCM)\bar{\Psi} + (L \otimes L)\bar{s}\right) \tag{2.37}$$

which can be written in an extended form as:

$$\bar{\Lambda}_t = \frac{1}{t} \begin{pmatrix} \frac{8\sigma - 9(w_1-1)w_1(X_{1,2})^2}{9} & \frac{4\sigma - 9(w_1-1)w_1(X_{1,2})(X_{1,3})}{9} \\ . & \frac{8\sigma - 9(w_1-1)w_1(X_{1,3})^2}{9} \end{pmatrix}$$

From these three estimable θs, it is straightforward to express σ and t in a similar way to the one timestep case (2.4.2):

$$t = \frac{(w_1 - 1)w_1(X_{1,2}X_{1,3} - X_{1,2,}^2)}{\theta_{1,1} - \theta_{1,2}} \tag{2.38}$$

and

$$\sigma = \frac{9}{8}(w_1 - 1)w_1 \left(\frac{\theta_{1,1}}{\theta_{1,1} - \theta_{1,2}} X_{1,2}X_{1,3} - \frac{\theta_{1,2}}{\theta_{1,1} - \theta_{1,2}} X_{1,2}^2\right) \tag{2.39}$$

We conclude that the model is identifiable in this case.

2.4.3. DISTINGUISHABILITY IN DFT'S SPECIAL CASES

In order to test distinguishability we compared the covariance matrices we had previously analysed for identifiability. The one timestep case is distinguishable from all the others (i.e. the only way to get equivalent covariance matrices in high sensitivity, zero memory decay or two alternative cases, iff $t = 1$). Similarly, we find that the two alternatives case is distinguishable from the others, as the data-dependent D can only be eliminated if there is only 1 timestep, the memory parameter is zero, or when the sensitivity is relatively high. The last two special cases to compare are the high sensitivity and the zero memory decay cases. Comparing equations 2.37 and 2.31 and their expanded

forms we can see that both of them are essentially the same matrix, one multiplied by $\frac{1}{t}$ the other multiplied by $\frac{\phi_2(1+(1-\phi_2)^t)}{(2-\phi_2)(1-(1-\phi_2)^t)}$.

If we now denote the zero memory decay model's timestep parameter with capital T, and make the two multiplicative terms equal, we find that

$$T = \frac{(2-\phi_2)\left(1-(1-\phi_2)^t\right)}{\phi_2\left(1+(1-\phi_2)^t\right)} \tag{2.40}$$

This means, that for any (t,ϕ_2) combination in a high sensitivity model, there is an equivalent zero memory decay model with T timesteps, that generates exactly the same choice probabilities for any input data.

We can solve this for t as well:

$$t = \frac{\ln\left(\frac{2-\phi_2-T*\phi_2}{2-\phi_2+T*\phi_2}\right)}{\ln(1-\phi_2)} \tag{2.41}$$

similar to equation 2.33. To sum up the two consequences of these two formulas (equations 2.40 and 2.41):

- For any combination of ϕ_2 and t in a high sensitivity model, there is a T, that will result in an equivalent zero memory decay model.

- For one zero memory decay model with timestep parameter T, there are several combinations of ϕ_2 and t that will result in equivalent models in a high sensitivity case.

Figure 2.3 shows the (t,ϕ_2) combinations with the corresponding Ts that satisfy equation 2.41.

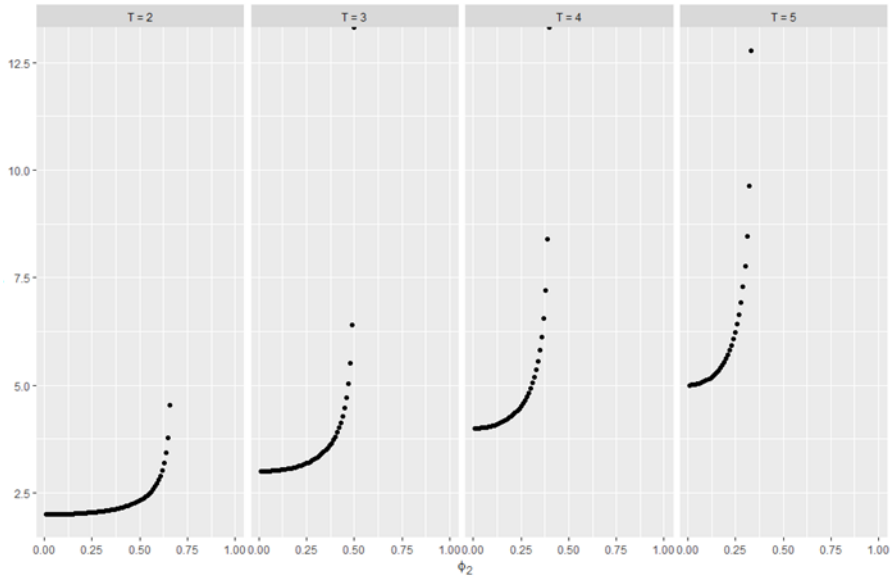


Figure 2.3: *In each facet we can see that for a zero memory decay model with timestep $T = 2,3,4,5$, which combinations of ϕ_2 and t in a high sensitivity model give the same output. On the x-axis we have the memory (ϕ_2), on the y-axis the timestep parameter (t). The observational equivalence of these combinations is not dependent on the data we use. This corresponds to Figure 2.2, where the log-likelihood function is flat along the exponential shape, which we can see here as well.*

We also examine an empirical example (for details, see Appendix 2.D), which can be summarized as follows. We take an estimated model from the literature (Hancock, Hess, and Choudhury, 2018, Table 3, model 4), that potentially fits our high sensitivity case (section 2.4.2) and generate choice probabilities with the estimated parameters (the relevant psychological parameters being $\phi_1 = 142.6043, \phi_2 = 0.1835, t = 112.2185$). Then we generate two other parameter sets: one based on equation 2.33 and one based on equation 2.40 (all the non-included parameters are kept the same), and calculate the choice probabilities based on these parameter-sets too. We confirm that for almost the whole dataset, the generated choice probabilities are the same for the three distinct parameter-sets. This means that a parameter-set that has high sensitivity, and other parameter-sets, which correspond to "better memory", or even "perfect memory" (i.e., $\phi_2 = 0$), and less timesteps during deliberation, are indistinguishable.

2.4.4. IDENTIFIABILITY AND DISTINGUISHABILITY IN NON-RESTRICTED DFT MODELS

In the previous section we showed that identification and distinguishability issues arise for two out of the four special cases of DFT. For special case 4, however, we were unable to draw a final conclusion, analytically (see Appendix 2.C for details). In this section we study whether DFT models exhibit identification problems in the more general case, where no parameters are restricted, thus including special case 4.

In order to do this, we conduct Monte Carlo simulations. That is, we generate numerous synthetic choice data sets using DFT data-generating processes (DGPs), and estimate the non-restricted DFT model. The data sets for these analyses were created in the following way. To obtain the choice tasks, we used two approaches: (1) a full factorial design and (2) a standard score normalisation. For each data set, we generated choices using a DFT DGP. In view of the inherent randomness of this process, 100 choice realisations were generated for each data set. Table 2.E.1 summarises the set-up and parametrisation of the Monte Carlo data sets. This analysis was conducted using 2 alternatives data sets and 3 alternatives data sets.

After having created the data sets (and the realisations of the choices), we estimated non-restricted DFT models, using the Apollo software (Hess and Palma, 2019). As DFT model estimation results can be sensitive to the set of starting values in the estimation, for each data set, we estimated, a series of DFT models using 200 sets of starting values. These starting values were drawn uniformly from pre-set ranges (see Table 2.E.2 in Appendix 2.E). We stored the 200 estimations for each of the 100 realisations of choices, so that we could randomly select one successful estimation. In our analysis, we used this single estimation result per dataset, which gave us a set of 100 model estimation instances to work with (one set for the case with 2 alternatives, and one for the case with 3 alternatives). Based on these estimations, we were able to test the statistical difference between the true underlying values on which we had based the realisations we had created, and the corresponding estimates. In order to do this, we used t-tests in the following way: we took the 100 estimates and their empirical standard error (i.e., the standard deviation from the mean estimate), and applied the formula: *t-statistic* = *mean(estimates-true value)/sd(estimates)*. Based on this t-test we report the t-value, p-value and the 95% confidence interval in Table 2.2 and 2.3. We also report the number of times the true value was in the 75% confidence interval of the estimator (denoted *count* in Table 2.2 and 2.3). For the final log-likelihood, we examined the *difference* between the true values from the DGP and the value obtained from the estimations. Note that, since the true value of the log-likelihood is different in each realisation (unlike the true value of the underlying parameters), the true value column for the difference in log-likelihoods is set to 0 in Tables 2.2 and 2.3.

In the following two subsections, we present our results. Specifically, section 2.4.4 reports the t-tests to establish whether or not the unrestricted DFT model's parameters can be recovered without bias. In cases where parameters cannot be recovered without bias, this points towards identifiability issues. After that, in section 2.4.4 we conduct a meta-analysis to acquire a better understanding of which features of the data are associated with successful or unsuccessful DFT estimations. Specifically, we look at whether we can find correlations between specific features of the data and the number of times estimation fails.

STUDYING BIAS IN DFT-PARAMETERS USING T-TESTS

Table 2.2 and 2.3 report the recovered and true parameters. To see whether or not model parameters are recovered without bias (i.e. whether or not the estimated parameters are significantly different from their true value), we apply t-tests. If the t-statistic is larger

than 1.96, we reject the hypothesis that the mean of the sampling distribution is equal to the true value at a 5% significance level. Besides applying t-tests to the model parameters, we also apply them to the log-likelihood of the model. After all, for each data set (and realisation of the choices) the true log-likelihood is known. Therefore, by comparing the true and estimated log-likelihoods, we can see the extent that the true model has recovered. Finding that the model fit is statistically different from the true one, serves as an indication that the model estimations might have failed, in the sense that they have most likely ended up in local optima. Table 2.2 shows the results for the two alternative data sets; Table 2.3 shows the results from the 3 alternative data sets.

Table 2.2: *T-tests for all estimated parameters in a 2 alternatives non-restricted DFT model.*

Parameter	True value	Estimate (mean across 100 estimations)	Empirical standard error (relative to mean)	t-statistic (mean estimate relative to true value)	p-value	Confidence interval	Count (out of 100 estimations)
t	1.5	2.084	0.141	4.152	0.0001	(1.805, 2.363)	83
ϕ_1	0.1	1.099	0.268	3.729	0.0003	(0.567, 1.631)	71
ϕ_2	0.3	0.086	0.103	−2.077	0.040	(−0.118, 0.290)	78
s	1	1.401	0.132	3.033	0.003	(1.139, 1.664)	79
β_2	0.7	0.698	0.011	−0.138	0.890	(0.676, 0.721)	75
β_3	0.5	0.496	0.011	−0.387	0.700	(0.473, 0.518)	73
LL-difference	0	−2.386	0.167	−14.331	0	(−2.717, −2.056)	*N/A*

Table 2.3: *T-tests for all estimated parameters in a 3 alternatives non-restricted DFT model.*

Parameter	True value	Estimate (mean across 100 estimations)	Empirical standard error (relative to mean)	t-statistic (mean estimate relative to true value)	p-value	Confidence interval	Count (out of 100 estimations)
t	2	2.179	0.084	2.135	0.035	(2.013, 2.344)	75
ϕ_1	0.2	0.520	0.042	7.527	0	(0.435, 0.604)	68
ϕ_2	0.6	0.260	0.012	−28.727	0	(0.237, 0.284)	38
s	1	0.407	0.137	−4.325	0.00004	(0.135, 0.679)	55
β_2	0.7	0.719	0.026	0.760	0.449	(0.669, 0.770)	66
LL-difference	0	22.820	4.242	5.379	0.00000	(14.403, 31.238)	*N/A*

Looking at the results in Table 2.2 and Table 2.3, we can make two observations. Firstly, the results are consistent across the two tables, in the sense that in both tables process parameters[8] are significantly different from their true value while the taste parameters[9] are not statistically different from their true values. Secondly, looking at the log-likelihood, we see they are significantly different from their true values. This suggests that a substantial number of estimations failed to recover the true model. Altogether, and also in light of the analytical results derived for the special cases, we conclude that there

[8]The estimated parameter (t) of Table 2.2 and Table 2.3 relates to the timesteps described in DFT theory (t_{theory}) in the following way: $t_{theory} = 1 + exp(t)$. This ensures that the DFT timesteps are always larger than 1 (Hancock, 2019).

[9]In the estimation the βs are related to the weights in the following way: $w_i = \frac{exp(\beta_i)}{\sum_j^J exp(\beta_j)}$, where J is the number of alternatives. This ensures that the weights add up to 1 in the estimation.

is strong evidence that the non-restricted DFT model also suffers from identifiability issues.

STUDYING THE ESTIMATION FAILURE OF DFT-MODELS: A META-ANALYSIS

Next, we conducted a meta-analysis to acquire a better understanding of which features of the data can be associated with successful or unsuccessful DFT estimations. To do so, we used a number of features or characteristics of data sets (such as the experimental design and number of alternatives) and of the DGP (such as the type of DFT) as explanatory variables, which we used to explain variation in: (1) the estimation success (fail / successful) and (2) the number of parameters for which the standard errors obtained were infinitely large.

For this analysis we created another 200 data sets, in which we varied the experimental design (full factorial, random attribute, standard score normalised), the number of alternatives (2,3,4), the number of attributes (2,3,4) and the value of three psychological parameters (i.e. sensitivity, memory decay and timestep parameters) in the underlying DFT DGP process. The steps of our methodology are as follows. First we created three kinds of data sets: full factorial, random attributes between 0 and 1, random attributes between 0 and 10. We created these three kinds of data sets for the case of choice sets containing 2, 3, and 4 alternatives, and for the case of 2, 3 and 4 attributes per choice alternative. For each of the resulting data sets, we also created a standard score normalised version. Next, we simulated the choices based on different DFT parameters, and computed the rho-squared for each data set. Then, from this set we selected 200 data sets that each had a rho-squared between 0.2 and 0.4. Table 2.4 summarises how many times the features described above appear in these 200 data sets. After having created these 200 data sets, we used them to estimate non-restricted DFT models. As DFT model estimations are known to be sensitive with respect to starting values, for each data set we used 200 sets of starting values, giving us a total of 40,000 estimation instances altogether. The starting value ranges which we uniformly drew from each time, are shown in Table 2.E.2 in Appendix 2.E.

Table 2.4: *Summary of the different features of the data sets (in bold) and the count of how many times they appear in the 200 datasets.*

number of alternatives	number of attributes	choice task type	Standard score normalisation	t	ϕ_1	ϕ_2
2 : 89	**2** : 76	**full factorial** : 78	**FALSE**: 95	**0.3** : 31	**0.1** : 58	**0.1** : 57
3 : 72	**3** : 85	**random between 0 and 1** : 52	**TRUE** : 105	**0.8** : 21	**0.2** : 26	**0.3** : 41
4 : 39	**4** : 39	**random between 0 and 10** : 70		**1.0** : 13	**0.5** : 22	**0.5** : 14
				1.5 : 34	**0.8** : 40	**0.6** : 43
				1.7 : 17	**1.6** : 23	**0.8** : 45
				2.0 : 84	**2.0** : 31	

Table 2.5 shows the regression results. These results provide a mixed picture regarding whether or not there is a systematic relationship between the explanatory variables (i.e. the features of the DFT DGP and the experimental design) and outcome variables (i.e. estimation failures, infinite standard errors, and very high standard errors). The number of alternatives in the choice set seems to be associated with a higher number of estimation failures. However, in cases where the model has converged, then the number of alternatives is associated with a lower number of infinite standard error estimates

and also with smaller standard errors of psychological parameters (see Table 2.E.3 in Appendix 2.E). An increasing number of attributes seems to be associated with fewer estimation failures, but has no significant effect on the number of infinite standard error estimates. Larger values for the sensitivity and memory decay parameters both seem to be associated with fewer estimation failures.

Table 2.5: *Multivariate regressions with two dependent variables: general estimation failures (representing 58.1% of our sample) and infinite standard errors (16.6%, which represent identification problems due to the flatness of the log-likelihood function). The constant represents the baseline of full factorial design with 2 alternatives and 2 attributes. Standard errors of estimated effects between brackets.*

	Dependent variable:	
	Number of general failures	Number of infinite standard error estimates
	(1)	(2)
3 alternatives	28.926^{***}	-29.075^{***}
	(1.447)	(4.044)
4 alternatives	36.856^{***}	-32.894^{***}
	(1.641)	(4.588)
3 attributes	-10.486^{***}	-5.701
	(1.503)	(4.201)
4 attributes	-16.110^{***}	6.383
	(1.778)	(4.970)
t	3.430^{***}	-4.469
	(1.146)	(3.204)
ϕ_1	-2.013^{**}	-7.679^{***}
	(0.851)	(2.378)
ϕ_2	-9.693^{***}	-33.262^{***}
	(2.448)	(6.843)
Random attributes from the range of 0-1	12.002^{***}	-7.332
	(1.611)	(4.504)
Random attributes from the range of 0-10	8.968^{***}	-0.773
	(1.423)	(3.979)
Standard score normalisation	11.337^{***}	-12.565^{***}
	(1.339)	(3.742)
Constant	94.818^{***}	86.733^{***}
	(2.774)	(7.755)
Observations	200	200
R^2	0.886	0.383
Adjusted R^2	0.880	0.350
Residual Std. Error (df = 189)	8.060	22.530
F Statistic (df = 10; 189)	146.865^{***}	11.726^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

2.5. CONCLUSION

This paper highlights and explores identifiability and distinguishability issues in Decision Field Theory (DFT) models. Our study is motivated by the fact that DFT models are routinely applied as so-called process models, which presumably give insight into the process of decision-making. When such models are estimated based only on out-

come data (i.e. the choice made), this begs the question whether process parameters, such as memory effects, can be inferred in the process of model estimation. Our paper sheds light on this issue by using a combination of analytical and simulation-based techniques. First, we show that four types of specifications of the DFT model are in fact equivalent to structured covariance probit models. We use this equivalence to study the parameter identifiability of DFT models' parameters, based on well-established results from the field of micro-econometrics (classical discrete choice theory). Specifically, we show that when the DFT model's sensitivity parameter is very high, its memory and time parameters are not jointly identifiable. Furthermore, we establish that the high sensitivity and the zero memory decay specifications of DFT are indistinguishable. Using Monte Carlo simulations, we show that the psychological parameters of the unrestricted (generic) specification of the DFT model are biased and that converged models often come with large and infinite standard errors for parameter estimates.

Our main methodological contribution lies in deriving the conditions under which the DFT model is equivalent to probit models, which allows for rigorous analytical methods to examine the identifiability of parameters in several cases. We believe this approach holds potential to be applied to other process models or alternative models proposed in mathematical psychology literature, helping to incorporate such models into the toolbox of choice modellers in general and travel behaviour researchers in particular. The main conclusion of our application of this probit-equivalence based technique, is that when estimating DFT models, it is very important to choose an appropriate model set-up which guarantees a solution to identification issues, or to collect data beyond choice observations, (concerning attention-wandering and decision time, for instance) to avoid misguided behavioural inferences.

Our findings are well aligned with results and intuitions that have already been found and discussed in existing literature. Below, we discuss these earlier insights, and we specify how our results are connected to them, highlighting the contributions of this paper. First we discuss the probit-equivalence-related results, then the identifiability investigations.

Berkowitsch et al. (2014) stated that the probit model is nested within the DFT model, when there is only one timestep. This case can be considered as an identity feedback matrix case, which is constrained by $t = 1$. As such, this specification satisfies the generic conditions we derive in Appendix 2.A (equations 2.48 and 2.49). Moreover, we formally derive which structure needs to be imposed on the covariance matrix for the two models to be equivalent. Busemeyer et al. (2006) pointed out the similarity between RUM models and DFT for a special case, where there are only two alternatives and the feedback matrix is identity. Our work extends this finding by showing that these two conditions are sufficient to prove equivalence between DFT and the RUM probit model. Furthermore, we find that the equivalence condition reported in our Appendix 2.A is satisfied by a fourth case: when the feedback matrix is diagonal but not necessarily identity. This is the case when the sensitivity parameter is relatively high and the distance between competing alternatives does not play a role. Our probit-equivalence findings derived in Appendix 2.A are generic in the sense that they are applicable to any feedback matrix

(not necessarily in the form of equation 2.2). To study the identifiability of the parameters we applied this generic result for the specification most often used in transportation literature (equation 2.2). This results in four special cases that allow for a formal identifiability analysis based on analytical derivations.

We find that if the sensitivity parameter is relatively high, the memory and timestep parameters are unidentifiable. A similar finding also appears in recent literature: Hancock (2019) (Chapter 6, 2.2) found, based on conceptual reasoning, that for two alternatives the sensitivity parameter loses its meaning, and that therefore the process will only depend on the memory and time, which then cannot be identified jointly. We show that this applies to multinomial cases as well, if the feedback matrix is diagonal (i.e. the sensitivity parameter is relatively high). This is important, because such a matrix can be an outcome of an estimation, without imposing it in advance (e.g. Hancock, Hess, and Choudhury, 2018, Table 3, Model 4). This identifiability problem is therefore primarily an empirical problem. This can be confirmed with the standard empirical identifiabilty test; setting the sensitivity to a very high value results in a singular Hessian at the maximum loglikelihood. Recent empirical investigations in current literature (Chapter 3 in Hancock, 2019) also showed that in some cases a non-restricted DFT generates similar results to a scaled multinomial logit (MNL) model, where the scale is a function of time. This is in line with our theoretical result that the zero memory decay model and the high sensitivity model are indistinguishable. In both these cases a more complicated version of a DFT model can be reduced to a simpler one with fewer parameters, where the covariance matrix is scaled. The scale can capture how deterministic the choice process is, but it is not possible to establish whether this is a result of more deliberation time or better memory. Our analytical result proves this for two special cases of DFT, while the empirical result of Hancock (2019) also extends to the domain of MNL, and indicates that a non-restrictive DFT might also exhibit the identification problem of memory and time.

In order to tackle the identifiability problem, there are several solutions that can be applied in the DFT framework. Based on the wide variety of DFT models proposed in existing literature, we provide an overview on the specifications that can ensure that the identification issues discussed above will not arise. They have all been used before in existing literature, but, with the exception of number 3, their use has not been connected to identifiability.

1. A zero memory decay model where only the timesteps are estimated. This specification was used by Hancock (2019), to demonstrate that the psychological parameters do not always result in an improvement in model fit. Our results show that this specification also serves as a solution to identification issues, when sensitivity is relatively high. This specification comes with the assumption that the decision maker has perfect memory, and that previous preference states matter just as much as the current one.

2. Assuming that timesteps go to infinity. Berkowitsch et al. (2014) used this specification when estimating their model, to avoid computationally intensive simulations in DFT estimation. Our results suggest that this assumption will also eliminate identification issues when the memory and timestep parameters are not

jointly identifiable. Behaviourally, this specification presumes that decision makers make a choice once their preferences have converged.

3. Including an initial preference state (which implies the behavioural assumption that the decision maker had an initial preference value towards the alternatives in the choice set). Hancock (2019) (Chapter 6) argues the identifiability of this specification in the context of binary choices (which implies in their study that the sensitivity parameter is not playing a role). Our results show that very high sensitivity can also lead to an identifiability problem, and that including an initial preference state ensures that the DFT model cannot be written in the form of equation 2.22. As such this identifiability issue can be avoided.

4. Scaling of the attributes. When done the right way, this eliminates the issue of relatively high sensitivity, as the relative magnitude depends on the attribute differences of alternatives. Scaling can take several different forms (for an overview see Chapter 4, 3.5 by Hancock (2019)). It has been suggested as a technique to gain better model fit and to avoid the necessity of a priori knowledge on whether an attribute has a positive or negative effect on the preference value (Hancock, 2019). We showed that identifiability issues can occur due to the interaction between the distance between alternatives and the sensitivity parameter, and that therefore scaling the attribute levels is also a technique to eliminate identification issues. It is important to note however, that even scaled attributes can lead to high sensitivity estimates, as it is the *relative* size of distance and sensitivity that matters.

These solutions each represent different underlying processes; as such, the behavioural conclusions that would be drawn from the resulting model specifications can be very different. This is especially important to keep in mind when one has limited data (for instance when the only observation is the final choice) and when the primary goal is not to find the best model fit or prediction, but rather to actually interpret the parameters and draw behavioural inferences. In this case, our results warrant for caution when estimating the DFT model and interpreting its parameters.

To illustrate this, consider the use of the DFT model to study moral decision-making processes. As evidenced by a growing amount of literature on social routing (e.g. van Essen et al., 2020), ethical aspects of road safety (see references below), and the moral dimension of travel mode choices (Matthies et al., 2006), these types of decisions are enjoying increasing levels of attention in the transportation research community. In morally sensitive situations, the trade-offs between the attributes of alternatives are often not as straightforward as in situations without a clear moral salience. For instance, in the case of a taboo trade-off situation in a road safety context (Chorus et al., 2018), involving a trade-off between so-called sacred (e.g. human lives) and secular (e.g. travel time) attributes, we expect that the distance[10] between attributes is very important. As such, the

[10] In a taboo trade-off case the sign of the attribute differences matters as well: trading human lives for travel time gains is taboo, while accepting longer travel times to save lives is not. This sets the taboo trade-off apart from the most commonly used Eucledian distance function. Note that the DFT model allows for non-Eucledian distance functions (e.g. Hotaling et al., 2010).

DFT-model's sensitivity parameter, which captures the importance of attribute distance on decision-making, is expected to be an important determinant of decision making, and understanding its identifiability in a DFT model is crucial.

Another example involving ethical decision making would be the study of empirically observed behavioural phenomena in the context of moral dilemmas; such dilemmas have gained popularity in the field of transportation since the re-invention of the so-called trolley problem to study the preferred behaviour of autonomous vehicles in collision situations (Awad et al., 2018): it is known that rapid decision making is associated with more deontologic (rule-based) ethics[11], while longer deliberation times are associated with more consequentialist ethical theories[12] (Suter and Hertwig, 2011). In such cases it is very important, when studying decision-making using DFT models, to disentangle the effect of decision time and memory.

These behavioural notions also urge further exploration of the identifiability of DFT models. In particular, the identifiability of the general DFT model (i.e. without imposing any restrictions on parameters) needs to be further investigated, to gain insight into what steps need to be taken to ensure that the unrestricted DFT model is identifiable. This thorny problem should, preferably, be approached using both empirical and simulated (process) data, while at the same time building on the analytical results provided in this paper. As a first step, collecting empirical data on attention-wandering and deliberation times will help develop suitable DFT-models (e.g. by testing different transformations on the parameters in the estimation or by developing alternative parametrisations for the feedback matrix). Then, to test whether the parameters of the resulting empirically supported models can be recovered without bias, the analytical steps laid out in this paper can be applied, together with analyses based on simulated data.

APPENDIX

2.A. DERIVATION OF PROBIT-LIKE FORMULAS

In order to find DFT specifications where conditions 2.19 and 2.20 hold, we need to examine the left-hand-side of condition 2.19. The aim is to get to equation 2.22, so that the model can be connected to the RUM specification of probit models. For this, we start with equation 6b in Hancock, Hess, and Choudhury (2018) assuming $P_0 = 0$, $S^* = (I - S)^{-1}(I - S^t)$ (due to this formulation, we leave subscript t out from the following derivations), and N is the number of alternatives. These reduce preference value equation to:

$$\xi = S^* \cdot \mu \tag{2.42}$$

which in matrix form is:

[11] In deontology an action can be considered right or wrong depending on whether it fits into a set of rules, and not based on its consequences.

[12] In consequentialism an action can be considered right or wrong based on its consequences. This is considered to be the main decision rule in economics, but it is also relevant in ethics in moral decision making.

2

$$\xi = \begin{bmatrix} S^*_{11} & S^*_{12} & \dots & S^*_{1N} \\ S^*_{21} & S^*_{22} & & \\ \vdots & & \ddots & \\ S^*_{N1} & & & S^*_{NN} \end{bmatrix} \times \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}$$

The generic element of the preference differences (i.e. the difference between alternative i and j) is:

$$\xi_i - \xi_j = (S^*_{i1} - S^*_{j1})\mu_1 + \dots + (S^*_{iN} - S^*_{jN})\mu_N \tag{2.43}$$

Substituting equation 2.25 as the generic element of vector μ,

$$\xi_i - \xi_j = (S^*_{i1} - S^*_{j1}) \sum_m w_m (x_{1m} - \frac{1}{N-1} \sum_{n \neq 1} x_{nm}) + \dots + (S^*_{iN} - S^*_{jN}) \sum_m w_m (x_{Nm} - \frac{1}{N-1} \sum_{n \neq N} x_{nm}) \tag{2.44}$$

This can be written concisely as:

$$\xi_i - \xi_j = \sum_m w_m \sum_{k=1}^{N} \left((S^*_{ik} - S^*_{jk}) - \frac{1}{N-1} \sum_{l \neq k} (S^*_{il} - S^*_{jl}) \right) x_{km} \tag{2.45}$$

From this equation we gain two conditions that ensure probit-equivalence:

- That the coefficient (multiplicative term) of x_{im} is equal to the opposite of the coefficient of x_{jm} (i.e. attribute m of alternative j, when we examine the difference between alternative i and j) for all m
- The coefficient of x_{km} when $k \neq i, j$ is zero for all m.

The first condition can be formalized as:

$$(S^*_{ii} - S^*_{ji}) - \frac{1}{N-1} \sum_{l \neq i} (S^*_{il} - S^*_{jl}) = -\left((S^*_{ij} - S^*_{jj}) - \frac{1}{N-1} \sum_{l \neq j} (S^*_{il} - S^*_{jl}) \right) \tag{2.46}$$

which can be simplified to:

$$\left(1 - \frac{1}{N-1}\right)(S^*_{ii} - S^*_{ji} + S^*_{ij} - S^*_{jj}) - \left(\frac{2}{N-1}\right) \sum_{l \neq i \neq j} (S^*_{il} - S^*_{jl}) = 0 \qquad \forall i, j, l \tag{2.47}$$

If we assume a symmetric matrix with the same elements on the diagonal (which is the case in all DFT applications so far), then the above condition reduces to:

$$\sum_{l \neq i \neq j} (S^*_{il} - S^*_{jl}) = 0 \qquad \forall i, j, l \tag{2.48}$$

meaning the off-diagonal elements of the matrix must be equal.

The second condition, written as a formula is:

$$\left((S^*_{ik} - S^*_{jk}) - \frac{1}{N-1}\sum_{l\neq k}(S^*_{il} - S^*_{jl})\right) = 0 \qquad \forall k \neq i,j \tag{2.49}$$

Which is satisfied once all the off-diagonal elements are equal. In the parametrisation we use in this paper, this means either there must be only two alternatives, or the feedback matrix must be diagonal. The reason for this is that the off-diagonal elements are dependent on the attribute distances (therefore the data). The only way they all become equal for any data in a multi-alternative case, is that the sensitivity parameter is very high (relative to the data), or the memory decay is zero.

2.B. NORMALIZATION

In probit models, we need to normalize the covariance matrix in order to have an identifiable model. In general the problem is that the multivariate normal probability distribution gives the same probability for the left and right hand side of the following equation:

$$\Phi(0|V,\Sigma) = \Phi(0|kV,k^2\Sigma) \tag{2.50}$$

Where V is the systematic utility difference between the two alternatives and Σ is the corresponding covariance matrix. In the following we use the 2 alternative, 2 attribute example for illustration. Let us call the following two equations 'model 1':

$$V = \beta_1 X_1 + \beta_2 X_2 \tag{2.51}$$

$$\Sigma = \Omega \tag{2.52}$$

where βs are the taste parameters and Xs are the attribute differences between the two alternatives. Ω is the estimated covariance without further specification.

Following on from equation 2.50, we can specify 'model 2' to be equivalent with 'model 1':

$$V' = \beta'_1 X_1 + \beta'_2 X_2 = k(\beta_1 X_1 + \beta_2 X_2) \tag{2.53}$$

$$\Sigma' = \Omega' = k^2\Omega \tag{2.54}$$

This shows, that if the estimable parameters of 'model 2' take the values listed below, the resulting distribution function is equivalent to that of 'model 1'.

$$\beta'_1 = k\beta_1$$
$$\beta'_2 = k\beta_2$$
$$\Omega' = k^2\Omega$$

Without further restrictions, this implies that there are an infinite number of $(\beta_1, \beta_2, \Omega)$ combinations that give the exact same choice probabilities, therefore the model is unidentifiable. The probit model handles it by fixing $\Omega = 1$, then all βs are uniquely defined.

In the following, we show that using the DFT restrictions (two kinds of specifications in particular) on weight parameters will ensure that there is only one k that gives

the same choice probabilities, and that is $k = 1$. Therefore the estimable parameters are uniquely defined.

The choice probability in DFT can also be expressed as in equation 2.50.

When the weights are the estimated parameters so that $w_1 + w_2 = 1$, thus $w_2 = 1 - w_1$ (in this case only w_1 is estimated), the following should hold:

$$w_1' = k w_1$$

$$w_2' = k w_2 = k(1 - w_1)$$

also, the condition that $w_1' + w_2' = 1$ must hold, therefore:

$$w_2' = 1 - w_1' = 1 - k w_1$$

merging the latter two equations we find:

$$k(1 - w_1) = 1 - k w_1$$

$$k = 1$$

If the taste parameters (named ws), and the estimable parameters (named βs) in DFT are not the same, and

$$V = w_1 X_1 + w_2 X_2 \tag{2.55}$$

$$\Sigma = \Omega \tag{2.56}$$

where the weights are expressed as

$$w_1 = \frac{exp(\beta_1)}{exp(\beta_1) + exp(\beta_2)} \tag{2.57}$$

and

$$w_2 = \frac{exp(\beta_2)}{exp(\beta_1) + exp(\beta_2)} \tag{2.58}$$

equations 2.57 and 2.58 imply that $w_1 + w_2 = 1$, which means the model is normalized and the weights are uniquely identifiable.

2.C. IDENTIFICATION IN CASE 4: TWO ALTERNATIVES

The fourth special case of DFT-probit equivalence is when there are only two alternatives. In this case the scale term is dependent on all psychological parameters and the timesteps as well.

$$\pi_t = \frac{1 - \left(1 - \phi_2(1 - e^{-\phi_1 D^2})\right)^t}{\phi_2(1 - e^{-\phi_1 D^2})} \cdot 2 \tag{2.59}$$

The scale term is dependent on the number of alternatives, the distance between the attributes of alternatives (D), sensitivity (ϕ_1), memory (ϕ_2) and also timesteps (t). The covariance matrix consists of a single θ, which is the following:

$$\bar{\Lambda_i} = \theta_i = \frac{\phi_2\left(e^{D_i\phi_1}-1\right)\left(\left(e^{-2D_i\phi_1}\left(\phi_2-\left(\phi_2-1\right)e^{D_i\phi_1}\right)^2\right)^t-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)}{2\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)\left(\left(\phi_2\left(e^{-D_i\phi_1}-1\right)+1\right)^t-1\right)^2} \tag{2.60}$$

for $i \in 1,..,n$, where n is the size of the data. Although the above equation contains four unknown variables $(\phi_1, \phi_2, t, \sigma)$, due to the heteroskedasticity we cannot eliminate the parameter identifiability based on the number of equations. Following up on our previous results (high sensitivity case, where t can be expressed as a function of ϕ_2), we solve the equation for t. In order to express t as a function of the other parameters, first let us reformulate the the following term from the numerator:

$$\left(\left(e^{-2D_i\phi_1}\left(\phi_2-\left(\phi_2-1\right)e^{D_i\phi_1}\right)^2\right)^t-1\right) = \left(\phi_2 e^{-D_i\phi_1}-\left(\phi_2-1\right)\right)^{2t}-1 \tag{2.61}$$

and the following term from the denominator:

$$\left(\left(\phi_2\left(e^{-D_i\phi_1}-1\right)+1\right)^t-1\right)^2 = \left(\left(\phi_2 e^{-D_i\phi_1}-\left(\phi_2-1\right)\right)^t-1\right)^2 \tag{2.62}$$

Let A be defined as:

$$A = \phi_2 e^{-D_i\phi_1}-\left(\phi_2-1\right) \tag{2.63}$$

By substituting A and using the well-known identity, $a^2-b^2=(a+b)(a-b)$, we can reformulate equation 2.61 as:

$$A^{2t}-1 = (A^t+1)(A^t-1) \tag{2.64}$$

and equation 2.62 as:

$$(A^t-1)^2 \tag{2.65}$$

Substituting the above two formulas, and simplifying the numerator and denominator by (A^t-1), equation 2.60 becomes:

$$\theta_i = \frac{\phi_2\left(e^{D_i\phi_1}-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)(A^t+1)}{2\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)(A^t-1)} \tag{2.66}$$

Rearranging this to express A^t we find:

$$A^t = \frac{\phi_2\left(e^{D_i\phi_1}-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)+\theta_i\cdot 2\cdot\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)}{\theta_i\cdot 2\cdot\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)-\phi_2\left(e^{D_i\phi_1}-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)} \tag{2.67}$$

from which it follows that:

$$t = \log_A\left(\frac{\phi_2\left(e^{D_i\phi_1}-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)+\theta_i\cdot 2\cdot\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)}{\theta_i\cdot 2\cdot\left(\left(\phi_2-2\right)e^{D_i\phi_1}-\phi_2\right)-\phi_2\left(e^{D_i\phi_1}-1\right)\left(\sigma-2\left(w_1-1\right)w_1\left(X_{1,2,i}\right)^2\right)}\right) \tag{2.68}$$

We find that t is dependent on D_i, which is not constant across the data. Although the analytical derivation does not give evidence for a theoretical identification problem,

due to the high non-linearity of the model we do not expect it to be identifiable. Hence, for the full model with 2 and 3 alternatives, we use simulations to further examine identifiability (section 2.4.4).

2

2.D. EMPIRICAL EXAMPLE

The high sensitivity case is not a very common issue in published literature. In this section we investigate an estimated model, where the estimated sensitivity parameter is unusually high (i.e. above hundred). In Hancock, Hess, and Choudhury (2018) Table 3, model 4, we see that $\phi_1 = 142.6043$, and the two other psychological parameters are $\phi_2 = 0.1835$, $t = 112.2185$. After standard score normalization of the data[13], we generate DFT choice probabilities based on the above parameters. Then, to establish whether there is an identification-issue, we test whether other parameter-sets result in the same choice probabilities. To find such parameter-sets, we use equation 2.33 and 2.40.

First, we examine a high sensitivity case with the memory decay parameter set approximately to half its size (i.e. people have "better memory" when deliberating). We set $\phi_2' = 0.09$, and apply equation 2.41 to get $t' = 10.72$. We generate DFT choice probabilities with this new parameter-set ($\phi_2' = 0.09$ and $t' = 10.72$, everything else kept the same as before), and examine the difference compared to the choice probabilities generated by the original parameter-set.
We see that the mean difference in choice probabilities (generated by the original and the newly calculated parameter-set) is 0.00162883 (less than 1 percentage point). Although the highest difference we see in the data is 12 percentage point, out of the total of 3492 data points, in 3380 instances the difference is less than 1 percentage point. This means basically, that for the most part of the data, the original parameter-set and the one corresponding to "better memory" and "less timesteps spent on deliberation" generate the same choice probabilities. Examining the 112 data points where the choice probabilities have larger difference (than 1 percentage point), we find that the average Eucledian distance between the alternatives is 25 times smaller compared to the lower-difference (than 1 percentage point) part of the data. This is because if the distance (D) is very small (i.e. close to 0), then $\exp(-\phi_1 * D^2) \approx 1$. This illustrates why we define high sensitivity as *relatively high sensitivity* in section 2.4.2; the smaller the distance between alternatives, the higher sensitivity parameter is 'needed' for the model to become unidentifiable.

Next, we test the distinguishability from the zero-memory decay model, applying equation 2.40, and generating DFT choice probabilities with $\phi_2'' = 0$ and $t'' = 9.899183$. The mean choice probability difference from the original estimated model is 0.001776593 (less than 1 percentage point), and out of 3492 data points, in 3373 the difference is less than 1 percentage point.

We can conclude, that although it is not a perfectly-high-sensitivity case, for a substantial part of the data (96.6% of the total sample), the generated choice probabilities are almost the same, whether we assume there is no memory decay at all and the deci-

[13] Swiss route choice data, Axhausen et al., 2008

sion is made quickly, or when there is some memory decay and more time is spent on deliberation.

2.E. SIMULATION RESULTS

Tables 2.E.1, 2.E.2 and 2.E.3 show information about the Monte Carlo simulations done in Section 2.4.4.

Table 2.E.1 shows the features and underlying parameters of the two DGPs used for the t-tests. Table 2.E.2 shows the starting values that were used in the estimation procedures.

Table 2.E.3 shows the result of different multivariate regressions on the standard errors (that did not fail nor went to infinity) of DFT parameters (extension of the meta-analysis of model failures of Section 2.4.4). From the R-squares we can see that the explanatory power of these models is relatively low.

Table 2.E.1: *Features and true values used in the data generating processes (DGPs).*

DGP attribute	2-alternative DGP	3-alternative DGP
DGP type	full factorial	full factorial
levels	$(2,2,4)$	$(2,4)$
standard score normalisation	✓	✓
number of choice tasks	2160	2016
n	2	3
m	3	2
t	1.5	2
ϕ_1	0.1	0.2
ϕ_2	0.3	0.6
s	1	1
β_1	0.2	0.2
β_2	0.7	0.7
β_3	0.5	N/A

Table 2.E.2: *Starting values uniformly drawn from the corresponding ranges for each parameter.*

Parameter	Starting value range
t	$(0,1)$
ϕ_1	$(0,1)$
ϕ_2	$(0,1)$
s	$(0,1)$
β_1	fixed at true value: 0.2
β_2	$(0,1)$
β_3	$(0,1)$
β_4	$(0,1)$

Table 2.E.3: *Multivariate regressions on non-failed estimations, the dependent variable being the standard errors of DFT parameters (seen in columns) and explanatory variables are features of the underlying data (seen in the rows). The constant represents the baseline of full factorial design with 2 alternatives and 2 attributes.*

	Dependent variable:			
	se_timesteps	se_phi1	se_phi2	se_error
	(1)	(2)	(3)	(4)
3 alternatives	−25.924*	−84.634**	−51.239***	−2.613***
	(15.162)	(32.794)	(18.131)	(0.980)
4 alternatives	−18.843	−61.699*	−40.747**	−2.329**
	(17.153)	(37.111)	(20.354)	(1.112)
3 attributes	−25.318	−50.865	−25.737	−2.624**
	(15.881)	(34.559)	(19.006)	(1.028)
4 attributes	−33.981*	−1.123	6.345	−2.483**
	(18.550)	(39.681)	(22.186)	(1.199)
t	15.308	−44.097*	−19.110	0.743
	(12.115)	(25.994)	(14.564)	(0.783)
ϕ_1	16.112*	−3.704	−12.886	1.050*
	(8.826)	(18.831)	(10.509)	(0.572)
ϕ_2	−28.987	27.318	20.104	−2.154
	(25.759)	(54.934)	(30.797)	(1.657)
Random attributes from the range of 0-1	8.954	−17.644	−12.113	0.816
	(17.098)	(36.078)	(20.226)	(1.103)
Random attributes from the range of 0-10	6.825	−71.257**	−42.798**	0.269
	(14.999)	(32.197)	(17.989)	(0.972)
Standard score normalisation	−24.437*	26.312	30.599*	−1.192
	(14.091)	(29.784)	(16.655)	(0.913)
Constant	26.046	166.168***	85.829**	3.450*
	(28.713)	(62.122)	(34.027)	(1.856)
Observations	185	182	180	186
R^2	0.072	0.100	0.117	0.105
Adjusted R^2	0.019	0.047	0.065	0.054
Residual Std. Error	81.510 (df = 174)	172.187 (df = 171)	95.947 (df = 169)	5.285 (df = 175)
F Statistic	1.355 (df = 10; 174)	1.896** (df = 10; 171)	2.240** (df = 10; 169)	2.063** (df = 10; 175)

Note: *p<0.1; **p<0.05; ***p<0.01

2.F. RELATION TO ORDER AND RANK CONDITIONS

Rank and order conditions (Bunch, 1991; Walker et al., 2007) were also suggested in the literature as ways to find identification issues in probit models.

The order condition states that in a probit model's covarince matrix there are maximum $\frac{J(J-1)}{2} - 1$ identifiable parameters, J being the number of alternatives. In the probit-equivalent DFT sepecifications this is not applicable as the covariance matrix is structured: it varies across the data.

The rank condition states that in a probit model's covarince matrix there are maximum $Rank(Jacobian(vecu(\Omega_\Delta)) - 1$ identifiable parameters (Walker et al., 2007). In the probit-equivalent DFT sepecifications this condition should be modified because the scale is already set by the weight normalization (Appendix 2.B); thus 1 should not be subtracted at the end. The rank condition can be used to confirm our findings. For instance, the rank of the jacobian of the vectorized covariance matrix in equation 2.32 is 2, while 3 parameters should be estimated. This is a direct result of the multiplicative term involving two parameters in equation 2.32. As the rank condition takes into account the structure of the covariance matrix (and not just the number of alternatives), it is also suitable for identification analysis in DFT.

REFERENCES

Alvarez, R. M., & Brehm, J. (1995). American ambivalence towards abortion policy: Development of a heteroskedastic probit model of competing values. *American Journal of Political Science*, *39*, 1055–1079.

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

Axhausen, K. W., Hess, S., König, A., Abay, G., Bates, J. J., & Bierlaire, M. (2008). Income and distance elasticities of values of travel time savings: New swiss results. *Transport Policy*, *15*(3), 173–185.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2014). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*, *143*(3), 1331.

Bolduc, D. (1992). Generalized autoregressive errors in the multinomial probit model. *Transportation Research Part B: Methodological*, *26*(2), 155–170.

Bolduc, D., Fortin, B., & Fournier, M.-A. (1996). The effect of incentive policies on the practice location of doctors: A multinomial probit analysis. *Journal of labor economics*, *14*(4), 703–732.

Bunch, D. S. (1991). Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, *25*(1), 1–12.

Busemeyer, J. R., Jessup, R. K., Johnson, J. G., & Townsend, J. T. (2006). Building bridges between neural models and complex decision making behaviour. *Neural Networks*, *19*(8), 1047–1058.

Busemeyer, J. R., Rieskamp, J. et al. (2014). Psychological research and theories on preferential choice. *Handbook of Choice Modeling. Cheltenham: Edward Elgar Publication*, 49–72.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences, 23*(3), 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432.

Chiou, L., & Walker, J. (2007). Masking identification of discrete choice models under simulation methods. *Journal of Econometrics*, *141*(2), 683–703.

Chorus, C. G., Pudāne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: A discrete choice model and empirical analysis. *Journal of choice modelling, 27*, 37–49.

Daganzo, C. (1979). *Multinomial probit: The theory and its application to demand forecasting*. Elsevier.

Hancock, T. O. (2019). *Travel behaviour modelling at the interface between econometrics and mathematical psychology* (Doctoral dissertation). University of Leeds.

Hancock, T. O., Hess, S., & Choudhury, C. (2018). Incorporating response time in a decision field theory model. *The Transportation Research Board (TRB) 97th Annual Meeting*.

Hancock, T. O., Hess, S., & Choudhury, C. F. (2018). Decision field theory: Improvements to current methodology and comparisons with standard choice modelling techniques. *Transportation Research Part B: Methodological*, *107*, 18–40.

Hausman, J. A., & Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica: Journal of the econometric society*, 403–426.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Hey, J. D., Lotito, G., & Maffioletti, A. (2010). The descriptive and predictive adequacy of theories of decision making under uncertainty/ambiguity. *Journal of risk and uncertainty*, *41*(2), 81–111.

Hotaling, J. M., Busemeyer, J. R., & Li, J. (2010). Theoretical developments in decision field theory: Comment on Tsetsos, Usher, and Chater (2010).

Matthies, E., Klöckner, C. A., & Preißner, C. L. (2006). Applying a modified moral decision making model to change habitual car use: How can commitment be effective? *Applied Psychology*, *55*(1), 91–106.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Noguchi, T., & Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, *132*(1), 44–56.

Raab, M., & Johnson, J. G. (2004). Individual differences of action orientation for risk taking in sports. *Research quarterly for exercise and sport*, *75*(3), 326–336.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, *39*(3), 577–591.

Scheibehenne, B., Rieskamp, J., & González-Vallejo, C. (2009). Cognitive models of choice: Comparing decision field theory to the proportional difference model. *Cognitive science*, *33*(5), 911–939.

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

Szép, T., van Cranenburgh, S., & Chorus, C. G. (2022). Decision field theory: Equivalence with probit models and guidance for identifiabilty. *Journal of Choice Modelling*, 100358.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

van Essen, M., Thomas, T., van Berkum, E., & Chorus, C. G. (2020). Travelers' compliance with social routing advice: Evidence from SP and RP experiments. *Transportation*, *47*(3), 1047–1070. https://doi.org/10.1007/s11116-018-9934-z

Walker, J. (2002). Mixed logit (or logit kernel) model: Dispelling misconceptions of identification. *Transportation Research Record*, *1805*(1), 86–98.

Walker, J., Ben-Akiva, M., & Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (NECLM) models. *Journal of Applied Econometrics*, *22*(6), 1095–1125.

Walter, E., & Pronzato, L. (1996). On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and computers in simulation*, *42*(2-3), 125–134.

Yai, T., Iwakura, S., Morichi, S., et al. (1997). Multinomial probit with structured covariance for route choice behavior. *Transportation research. Part B, Methodological*, *31*(3), 195–207.

3

IDENTIFICATION OF PREFERENCES UNDER OBFUSCATING BEHAVIOUR

Part I of this thesis focuses on parameter identifiability and recoverability in novel Discrete Choice Models. This chapter, the second study of Part I, concerns a model that was designed to capture a decision-maker's intention to obfuscate their preferences. Obfuscation can be highly relevant in several situations involving morality. The identifiability of preferences in such a model is crucial to interpreting the model's parameters based on behavioural theory.
This chapter uses Monte Carlo simulations to examine whether preferences can be recovered 1) in the original obfuscation model and 2) in an extended version of the obfuscation model proposed in this study, called sequential obfuscation. In sequential obfuscation, the decision-makers make their decisions assuming that not only their current choice but their previous choices were also observed. This study examines preference recoverability under varying levels of obfuscation intention and a varying number of choice tasks. Section 3.1 introduces the relevance and related literature. Section 3.2 introduces the original obfuscation model formally. Section 3.3 presents the identifiability results in the original obfuscation model under varying levels of obfuscation. Section 3.4 presents the formalization of sequential obfuscation, and section 3.5 shows the identifiability results under varying levels of obfuscation and varying number of choice tasks. Section 3.6 concludes with practical implications and potential future research avenues.

This chapter builds on, and partly repeats, the paper entitled 'Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence' by Chorus et al. (2021). My role in that paper, of which I am a co-author, was to establish whether the preference weights and the obfuscation parameter are jointly identifiable; the corresponding analysis can be found in Appendix B of the paper, which serves as the base for section 3.3 of this chapter. For readability reasons, this paper also presents the theoretical base for the obfuscation model (which is the work of the paper's first author, Caspar Chorus); section 3.2 is based on and partly repeats section 2 of the paper by Chorus et al. (2021). For readability reasons, I do not cite the paper throughout these sections.

3.1. INTRODUCTION

A fundamental assumption on which traditional Discrete Choice Models (DCMs) are built is that preferences echo through the choices made by decision-makers. This means that when we see someone opting for a red over a blue car when every other attributes of the product are the same, we can conclude that the decision-maker prefers red to blue. This crucial assumption allows DCMs to estimate preference weights (i.e. relative importance of different attributes) from observed choices. However, in several situations, especially when morality is involved, this assumption may not hold. People have various incentives to suppress their true underlying preferences, such as avoiding judgment or shame, protecting their privacy, or allowing themselves some 'wiggle room' in case they have to explain their actions later. Suppose a politician has to decide which welfare program to discontinue: one that is targeted to help people living on minimum wage, one that is targeted to help the retired, or one that is designed to be less transparent and it is not clear to the public, who the beneficiaries are. In this situation politicians have an incentive to choose less transparency and thus the third alternative. Choosing such strategy allows politicians to minimise blame and empirical evidence for such behaviour was found related to Swedish welfare policy cutbacks (Lindbom, 2007). Obfuscating behaviour can be furthermore relevant in online behaviour: users often alter their behaviour when they know they are being monitored and targeted by different organizations. For instance, to avoid being banned from Twitter, hate speech is often hidden behind coded language (Dunn et al., 2018). In interpersonal situations, such as lending money to a family member, behaviours that obfuscate one's capacity to help are often triggered to avoid awkwardness or social pressure (Wherry et al., 2019). In situations where trading goods are considered to be taboo, such as compensated adoption or bribery, various strategies emerged that facilitate exchange without the obvious revealing of intentions. These involve finding third parties or using gift exchange in order to avoid social judgement, or even legal consequences (e.g.,Rossman, 2014; Schilke and Rossman, 2018).

Based on these behavioural notions, the recently developed obfuscation model (Chorus et al., 2021) relaxes the fundamental assumption of preferences echoing through one's actions. The model is designed to capture obfuscating behaviour, which is when people do not want to reveal their true underlying preferences but rather hide them to some extent. The obfuscation model concerns three agents, and having a clear distinc-

tion between them is crucial in this study.

1. The *decision-maker* (DM) wants to hide their true preferences about the different attributes in a choice task from an onlooker.

2. The *onlooker* may or may not exist. Their relevance is that the DM tries to obfuscate their preferences from the onlooker; therefore, DM makes their decision partially based on what DM assumes about the onlooker's prior beliefs and thinking process. In other words, the ground truth about an onlooker's thinking is irrelevant; what matters is what the DM believes about it. Therefore, from now on, "the onlooker does not update their priors" strictly speaking means "the DM believes that the onlooker does not update their priors".

3. The analyst tries to estimate the preferences of the DM and may or may not calculate with their obfuscation intention. The analyst differs from the onlooker, as the onlooker uses Bayesian inference to guess the preferences of the DM, while the analyst estimates them with maximum likelihood estimation. The DM does not consider the analyst in their decision making.

Figure 1 shows the agents, how they relate to one another, and the sections of this chapter that concern their perspectives or behaviours.

The aim of this study is twofold; first, it aims to relax a behavioural assumption that is not always realistic, namely that the decision-maker believes only one of their choices is observed. Second, it aims to establish identifiability of parameters, meaning that the parameters can be recovered when estimated with correct assumptions. Identifiability is crucial for drawing behavioural conclusions from a theory-driven model.

The originally developed obfuscation model (presented in detail in section 3.2) assumes that the decision-maker believes only their current choice is observable to the onlooker. This assumption holds in several cases when a single individual has to make only one decision at a time, for instance, which charity to donate to at an event. Even when more similar decisions are observable, the assumption is still realistic if they have no common attribute. For instance, a politician is voting about several unrelated policy packages: one concerning tax distributions, another gender equality, and a third about foreign policy. However, in several cases, decisions may be made sequentially, and an observer has the opportunity to update their beliefs about the decision-maker after every decision. For example, suppose one donates to different charities on several events. In that case, other attendees can get a hint after every pledge about the decision-maker's preferences, and the decision-maker is probably aware of this. Politicians can also be observed for several years, and all their tax policy votes can inform their observers about their true preferences. Furthermore, Discrete Choice Models are often used with data where multiple choices of an individual are observed, and discrete choice experiments typically present several choice tasks sequentially to respondents. Thus, several applications of DCMs, where obfuscation can play a role, concern decisions made sequentially, where the single choice assumption does not hold. In this chapter, I propose an extension to the model that takes this into account. In this extension, when an agent sequentially obfuscates, they consider that an observer sees not only their current choice

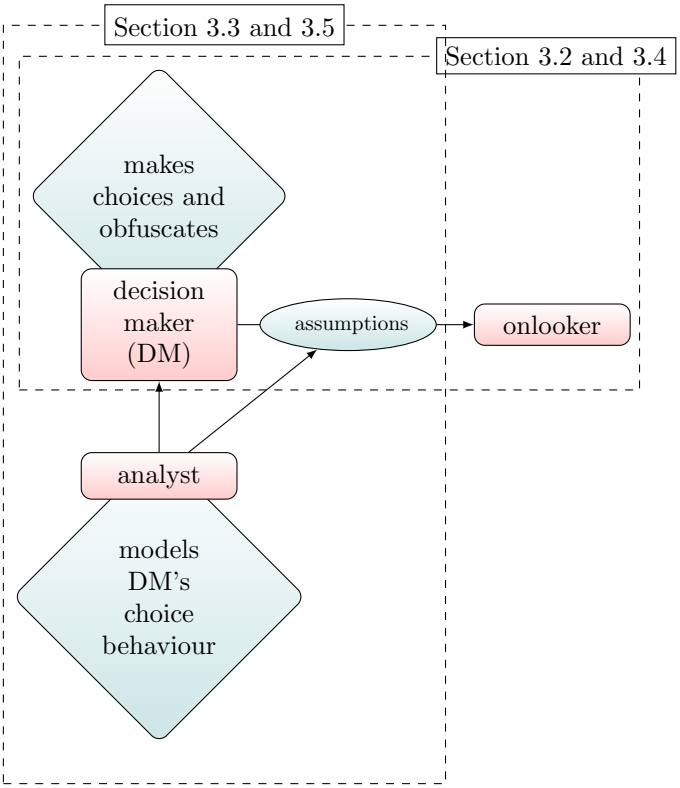


Figure 1: *The three agents and their relations. Arrows depict when an agent considers another. The decision-maker believes an onlooker is observing them and tries to hide their preferences from them. The analyst is only concerned with the decision-maker: it is irrelevant whether the onlooker exists; the only thing that matters to them is what the decision-maker believes about them that influences their choices.*

but also the ones beforehand. Therefore, if someone made a choice that might be "too revealing" of their underlying preferences, they have a chance to offset it in subsequent decisions. For example, when someone does not want to seem too 'unethical' (choosing a taboo too often) or 'frugal' (going for the cheapest alternative too often), or 'ungenerous' (distributing benefits for mostly selfish rather than charitable purposes).

To interpret and draw behavioural conclusions from a theory-driven choice model's parameter estimates, identifiability of the model's parameters is crucial. Identifiability allows the analyst to estimate the true underlying preferences accurately. As various decision-making strategies that involve hiding true preferences, such as making random choices, or deceiving onlookers, do not allow an analyst to recover true preferences from observed choices, the question arises: does obfuscation behaviour allow for it? Is it possible to recover the extent to which a decision-maker obfuscates jointly with the

attribute weights when the decision-maker tries to obfuscate them? To test parameter identifiability in both the original and the sequential obfuscation model, I use a Monte Carlo experiment (e.g., Ben-Akiva et al., 2002). First, I generate synthetic data on choice tasks and choices, with varying levels of obfuscation and varying number of choice tasks completed by one individual. Then, I estimate the obfuscation model and test if the parameters can be accurately recovered. The results suggest that the obfuscation model's parameters can be recovered without bias. However, obfuscating behaviour affects the estimates' standard error: the more one wants to obfuscate, the less certainty an analyst has about their true preferences. The magnitude and patterns (how the standard errors change with varying levels of obfuscation or number of choice tasks) of this uncertainty depend on whether the obfuscation strategy builds on previous choices.

The outline of this thesis chapter is the following. Section 3.2 presents the original obfuscation model, section 3.3 analyses its identifiability with Monte Carlo simulations. Section 3.4 presents the sequential obfuscation, and 3.5 its identifiability, again using Monte Carlo simulations. Section 3.6 discusses the interpretation of the results and further research directions. As it is shown on Figure 1, section 3.2 and 3.4 describes a single decision-maker's thought process under obfuscation and sequential obfuscation accordingly. Taking the analyst's perspective into account, section 3.3 and 3.5 presents the corresponding identifiability analyses.

3.2. THE OBFUSCATION FRAMEWORK

In this section, I first present the formalization of a single obfuscating decision-maker's behaviour in the original (or single choice) obfuscation framework. It is important to note that in this section, the perspective of an analyst analyzing choices is not yet adopted; hence, I do not discuss any econometric considerations. Those will be the topic of section 3.3. This section presents the process that goes through the decision-maker's (DM's) mind and the corresponding mathematical formulations or illustrations.

Consider a decision-maker who's task is choosing an alternative from set A containing J alternatives: $\{a_1, ...a_j ...a_J\}$. Set G contains K attributes (or goals, or criteria) based on which the alternatives can be assessed: $\{g_1, ...g_k ...g_K\}$. The extent to which the decision-maker cares about each attribute g_k is denoted by weights β_k. Assume for ease of communication, but without loss of generic applicability, that β_k takes one value from a set of possible values (i.e. $\beta_k \in \{0, 1, 2, ..., M\} \quad \forall k$). If the decision-maker does not care about a particular attribute, the associated weight equals zero; increasing values reflect the increasing importance of the attribute; a weight of M reflects that the attribute is of the highest possible importance to the decision-maker. Matrix X (size of $K \times J$) contains scores denoted by x_{kj}, which reflect how alternative j scores on attribute k; the non-negative attribute-weights imply that higher scores are preferred over lower ones. The aggregated utility associated with alternative j is $u_j = \sum_{k=1}^{K} u_{jk}$, where $u_{jk} = \beta_k \cdot x_{kj}$

Note that this aggregation reflects a classical linear-additive multiattribute utility approach; other aggregation procedures may also be considered. Denote the K-dimensional vector containing the weights of all attributes as β, which defines the decision-makers preferences. The decision-maker's beliefs are defined as follows:

1. An onlooker is watching their decision.

2. The onlooker observes the alternatives (A), their attributes (G), and how each alternative scores on each attribute (X); they have the same perception of these vectors and matrix as the agent themselves.

3. The onlooker has uninformative prior probabilistic beliefs $P(\beta)$ about the preference weights of different attributes. The onlooker knows that each weight is an element from the set $\{0,1,2,...,M\}$. The onlooker's multidimensional uninformative prior thus consists of probabilities of size $\frac{1}{(M+1)K}$ for each of the $(M+1)K$ possible states of the world, where each state is characterized by a realization of each of the K weights β_k.

4. The onlooker observes one choice by the decision-maker from A and uses that observation to update their beliefs about the preference weights (β) into posterior probabilities; they do so using Bayes' rule. The onlooker's posterior probabilities, after having observed the decision-maker's choice for alternative a_j, are given by[1]:

$$P(\beta|a_j) = \frac{P(a_j|\beta)P(\beta)}{\sum_{\beta \in B} P(a_j|)P(\beta)} \tag{3.1}$$

Here B represents the domain of β (i.e., it contains all $(M+1)K$ possible states of the world), and $P(a_j|\beta)$ is given by the well-known logit-formulation[2] (Luce, 1959; McFadden, 1973), which stipulates that the probability of choosing an action given a set of preferences increases when the utility of that action (which is a function of the decision-maker's preferences and the action's scores) increases.

$$P(a_j|\beta) = \frac{\exp(\sum_{k=1}^{K} u_{jk})}{\sum_{l=1}^{J} \exp(\sum_{k=1}^{K} u_{lk})} \tag{3.2}$$

In the following sub-sections, a model of a 'hybrid' agent is presented, who attempts to choose in line with their preferences while at the same time trying to avoid the onlooker learning those underlying preferences. A 'preference-aligned' decision-maker (i.e. one who believes they are not being observed or simply ignores onlookers) applies their preferences to each alternative, giving:

$$u_j = \sum_{k=1}^{K} u_{jk} + \varepsilon_j = \sum_{k=1}^{K} \beta_k \cdot x_{kj} + \varepsilon_j \tag{3.3}$$

for alternative j; they then choose the alternative with the highest aggregated utility. An obfuscating decision-maker considers that the onlooker quantifies the remaining uncertainty (i.e., after having observed his choice for a_j) in terms of Shannon entropy (Shannon, 1948):

[1] Note that although an onlooker might not necessarily knows the decision-maker's decision rule, the decision-maker believes the onlooker updates their beliefs based on equation 3.1.

[2] In theory, $P(a_j|\beta)$ could be given by any random utility model. Relying on the logit-formula keeps the computation time relatively low, even when two intertwined thought processes are modelled at the same time.

$$H_j = -\sum_{\beta \in B} P(\beta|a_j)\log(P(\beta|a_j)) \tag{3.4}$$

The obfuscating agent chooses the alternative which maximizes entropy[3]: $argmax_{j=1...J}\{H_j\}$. A hybrid decision-maker's behaviour is driven by a combination of preference-oriented behaviour and entropy maximization, which may be represented by a utility maximization process where the utility of an alternative is:

$$U_j = u_j + \gamma \cdot H_j \tag{3.5}$$

This model has solid behavioural intuition: it represents a decision-maker who wishes to fulfil their preferences (through u_j) but willing to give up some preference-related utility if this preserves his privacy by prohibiting the onlooker from learning his preferences (through $\gamma \cdot H_j$).

3.3. IDENTIFICATION IN THE SINGLE CHOICE OBFUSCATION MODEL

In this section, I move to the perspective of parameter identification by a decision analyst in the context of a dataset containing choices resulting from (possible) obfuscation-based choice behaviour by a set of decision-makers. I generate a set of choice tasks; then, I use Monte Carlo simulation to generate choice tasks with varying levels of obfuscation intention. Then on each dataset containing choices, I estimate the obfuscation model, first ignoring the obfuscation, then taking it into account. The details on the data generation process and method can be found in section 3.3.1, and the corresponding results in section 3.3.2.

3.3.1. DATA GENERATION AND METHODOLOGY

The situation I consider is one where decision-makers, onlookers and decision analysts have the following behaviours:

- The decision-maker chooses from a set of three alternatives j that are defined in terms of their scores x on two attributes; he may be concerned with obfuscation and preference-aligned behaviour. More specifically, the decision-maker maximizes random utility, and his utility function for alternative j is specified as $U_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \gamma H_j + \varepsilon_j$, where $\beta_1 = 1$ and $\beta_2 = 2$. That is, an alternative's utility consists of the sum of i) a weighted summation of the alternative's scores on the two attributes and their corresponding attribute weights, the second attribute being twice as important to the decision-maker as the first one; ii) the entropy of the alternative multiplied by an obfuscation weight γ; iii) an iid Extreme Value Type I error term with variance equalling $\frac{\pi^2}{6}$. If the obfuscation weight equals zero, the model collapses to a standard linear additive random utility maximization based logit model.

[3] But note that while the obfuscating agent chooses based on entropy maximization, they are assumed to believe – see equation 3.2 – that the onlooker does not consider the possibility that the decision-maker might obfuscate; in other words, the decision-maker believes that the onlooker believes that his (the decision-makers) choices are purely preference-aligned.

- The onlooker may be a real person or a mere mental representation in the decision-maker's mind (think of the 'moral persona' invoked in Adam Smith's writings). Note that even when the onlooker is real, it is not their actual behaviour that is of interest, but rather the decision-maker's beliefs regarding their behaviour. These beliefs are as follows: the onlooker inspects the observed choice and attempts to infer DM's attribute weights β_1,β_2. The onlooker does so using the Bayesian learning scheme presented in equations 3.1 and 3.2 presented in section 3.2. We adopt the same settings as in the example presented in that section for ease of exposition. That is, the choice set contains three alternatives, there are two attribute weights, and the onlooker is uncertain about which element of the set $\{0,1,2\}$ represents the decision-maker's weight for any particular attribute. (Note that we tested several variations of these attribute weights, leading to similar results.) Before observing the choice, the onlooker assigns an uninformative prior probability of 1/9 to each of the following nine states of the world (see equations 3.6 and 3.7).

- The decision analyst receives a dataset containing 7500 choice observations, consisting of 15 choices made by 500 decision-makers (note that we checked that our conclusions also hold for considerably smaller datasets, e.g. containing 500 cases). Each decision-maker has the same attribute weights (i.e., $\beta_1 = 1$ and $\beta_2 = 2$ as mentioned above) but is confronted with a different choice task: attribute values x_{j1} and x_{j2} (for $j \in \{1,2,3\}$) were randomly[4] –across alternatives and choice tasks– drawn from the interval $[0,1]$. Throughout the Monte Carlo analyses, the obfuscation parameter γ is systematically varied but kept constant across decision-makers. The analyst identifies parameters employing maximum likelihood estimation. Two cases can be distinguished: first, the analyst is 'naive' and believes that the decision-maker's utility function is characterized as $U_j = \beta_1 x_j 1 + \beta_2 x_{j2} + \varepsilon_j$. The analyst does not consider that the decision-maker might have been trying to obfuscate an onlooker. Second, the analyst is 'prepared', allowing for the possibility that the decision-maker might have been trying to obfuscate an onlooker while not knowing if and to what extent this is the case. In this case, the analyst assumes the utility function which was described further above: $U_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \gamma H_j + \varepsilon_j$. Here, the analyst attempts to jointly estimate attribute weights and the obfuscation parameter.

The main question that the Monte Carlo experiment attempts to answer can be put as follows: in case $\gamma > 0$ (i.e., when decision-makers have attempted to obfuscate their attribute weights from a real or imagined onlooker), would the analyst still be able to identify the obfuscation parameter, which gives the degree of obfuscation, jointly with β_1 and β_2, which give the true attribute weights which the decision-makers have attempted to hide from the onlooker? Before presenting our results, one critical remark needs to be made: entropy H_i is a function of the decision maker's beliefs regarding uncertainty in the mind of the onlooker. As such, entropy is based on the decision-maker's beliefs about

[4]Note that this data contains dominant alternatives in some choice sets too. A dominant alternative in a single choice scenario does not allow for trade-offs, thus obfuscating bahaviour, however in a dataset where there is substantial amount of choice scenarios which are not containing dominant alternatives, the estimation is not adversely affected.

how the onlooker will use an observed choice to update a prior distribution regarding his preferences (attribute weights) into a posterior distribution. From the analyst's viewpoint, this entropy is a data point computed based on the choice task before the process of model estimation; it is not a function of the analyst's estimates of the attribute weights. In other words, in the process of model estimation (i.e., the process of finding the maximum likelihood attribute weights and entropy parameter), the entropy itself is invariant. It should also be noted that these analyses presuppose that the analyst is aware of the decision-maker's beliefs about the onlooker's priors and about how the onlooker would update those based on the choice made by the decision-maker. This fairly restricted assumption should be relaxed in future research to explore the identifiability of the obfuscation model under more lenient conditions, for example, using continuous distributions for the attribute weights specified over a larger domain of possible values. I use the newly developed R-package Apollo (Hess and Palma, 2019) for the analyses; the code can be found at https://github.com/szepteodora/obfuscation_identification.

Note that the synthetic data set used for the identification analyses is based on simulated choice probabilities for three alternatives for each of the 7500 decision tasks. The alternatives' attribute values vary across decision-makers in the synthetic dataset, but each decision-maker is assumed to have the same preferences and obfuscation-related beliefs. A formulation for the simulated probability that a particular decision-maker, faced with a choice set, chooses a particular alternative (hence the notation omits a subscript for decision-makers) is presented below. For ease of communication, the notation slightly differs from the one used directly above. The symbol β is now used for the parameter the analyst will estimate. The symbol $\widetilde{\beta}$ is used for the parameter which indirectly (i.e., through the entropy which the decision-maker believes exists in the mind of the onlooker) determines the decision-maker's behaviour, but is not estimated by the analyst. Another small addition in notation concerns the use of s to denote a possible state of the world. The decision-maker believes that the onlooker assigns a prior probability of 1/9 to each of the following nine possible states of the world:

$$\widetilde{\beta} = \begin{cases} \widetilde{\beta}_1^1 = 0 & \widetilde{\beta}_2^1 = 0 \\ \widetilde{\beta}_1^2 = 1 & \widetilde{\beta}_2^2 = 0 \\ \widetilde{\beta}_1^3 = 2 & \widetilde{\beta}_2^3 = 0 \\ \widetilde{\beta}_1^4 = 0 & \widetilde{\beta}_2^4 = 1 \\ \widetilde{\beta}_1^5 = 1 & \widetilde{\beta}_2^5 = 1 \\ \widetilde{\beta}_1^6 = 2 & \widetilde{\beta}_2^6 = 1 \\ \widetilde{\beta}_1^7 = 0 & \widetilde{\beta}_2^7 = 2 \\ \widetilde{\beta}_1^8 = 1 & \widetilde{\beta}_2^8 = 2 \\ \widetilde{\beta}_1^9 = 2 & \widetilde{\beta}_2^9 = 2 \end{cases} \tag{3.6}$$

$$P(\widetilde{\beta}^s) = \frac{1}{9} \quad \forall s \tag{3.7}$$

The decision-maker believes that the onlooker assigns the following choice probability to alternative A from a set of three alternatives $\{A, B, C\}$, given a particular state of the world ($\widetilde{\beta}^s$) and given the attribute scores (which are also observed by the onlooker):

$$P(A|\widetilde{\beta}^s) = \frac{\exp(\widetilde{\beta}_1^s x_{1A} + \widetilde{\beta}_2^s x_{2A})}{\sum_{l\in\{A,B,C\}} \exp(\widetilde{\beta}_1^s x_{1l} + \widetilde{\beta}_2^s x_{2l})} \tag{3.8}$$

This implies that the decision-maker believes that the onlooker believes that the decision-maker maximizes utility and does not obfuscate. The decision-maker also believes that upon seeing their choice for, for example, alternative A, the onlooker will update their prior probabilities $P(\widetilde{\beta}^s)$ as to which state of the world prevails into posterior probabilities $P(\widetilde{\beta}^s|A))$ using Bayes' formula:

$$P(\widetilde{\beta}^s|A) = \frac{P(A|\widetilde{\beta}^s)P(\widetilde{\beta}^s)}{\sum_{k\in\{1,\dots,9\}} P(A|\widetilde{\beta}^k)P(\widetilde{\beta}^k)} \tag{3.9}$$

Here, $P(A|\widetilde{\beta}^s)$ is as given in equation 3.8, and $P(\widetilde{\beta}^s)$ is as given in equation 3.7. Given these beliefs held by the decision-maker, their belief concerning the entropy in the mind of the onlooker, after the onlooker has observed a choice for alternative A equals:

$$H_A = -\sum_{s\in\{1,\dots,9\}} P(\widetilde{\beta}^s|A)\log P(\widetilde{\beta}^s|A) \tag{3.10}$$

The decision-maker's choice behaviour (e.g. the probability that he chooses alternative A from a set of A,B,C) is governed by the following logit formula, which includes goal-directed utility as well an entropy related term:

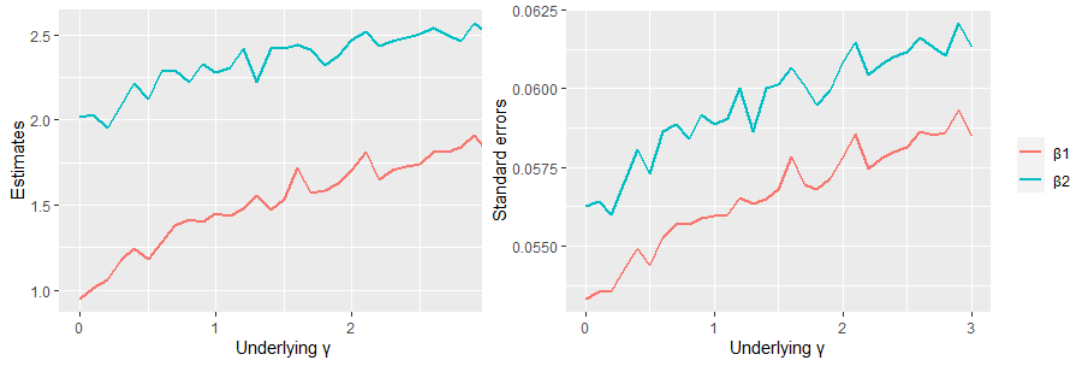
$$P_A = \frac{\exp \beta_1 x_{1A} + \beta_2 x_{2A} + \gamma I_A}{\sum_{i\in\{A,B,C\}} \exp \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma I_i} \tag{3.11}$$

Here, entropy terms H are computed as given in equation 3.10. Similarly, choice probabilities for alternatives B and C are obtained. Based on these choice probabilities, choices are simulated for 500 virtual decision-makers, each making 15 choices given particular attribute values for all three alternatives. The analyst then uses this data set containing 7500 choices for model estimation. It is important to repeat there that only parameters β_1, β_2, γ are estimated by the analyst in the stage of model estimation. In contrast, $\widetilde{\beta_1}^s$ and $\widetilde{\beta_2}^s$ which are embedded in the entropy terms (through equations 3.6–3.10) are pre-defined (see equation 3.6), and they are not estimated. In other words, the entropy term in 3.11 is computed prior to estimation, based on each observation's attribute levels, and subsequently used as fixed input (i.e., 'data') in the stage of model estimation.

3.3.2. RESULTS

As a starting point for the next analyses, I confirm the obvious intuition that if the analyst is naive and the decision-makers' $\gamma = 0$ (i.e., they do not obfuscate), the true attribute weights are recovered without any problem. Next, another obvious intuition can be confirmed: if the analyst is 'naive' and if the decision-makers' $\gamma > 0$ (i.e., they do obfuscate to a certain level), the estimates for the attribute weights become biased and increasingly so as γ gets bigger. Figure 2 shows the estimates and standard errors for $\gamma \in [0,3]$.

This finding was expected, as there is a mismatch between the utility function of the decision-makers, and the one assumed by the analyst. This result's straightforward and

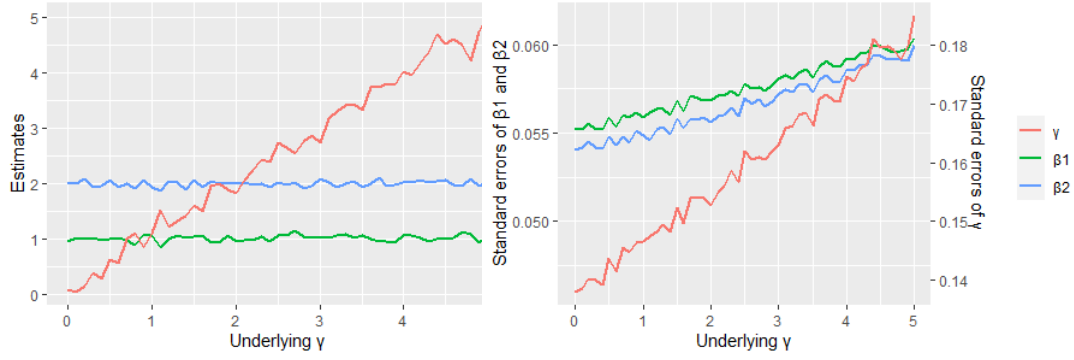


(a) *(b)*

Figure 2: *Estimates and standard errors obtained by the 'naive' analyst under varied levels of obfuscation (γ) when the true underlying preferences are $\beta_1 = 1$ and $\beta_2 = 2$. Increased level of obfuscation results in larger bias and larger standard errors.*

intuitive implication is that when decision-makers obfuscate and the analyst is unaware of that –thus does not specify it in the estimated choice model– estimation results will be biased.

Now I turn to the case where decision-makers obfuscate, and the analyst is 'prepared', that is, they allow for the possibility that decision-makers obfuscate but do not know whether or not and to what extent this has happened (Figure 3a and 3b). Figure 3a shows that in this case, the true attribute weights and the obfuscation parameter are jointly being recovered by the analyst without noticeable bias, even when the obfuscation parameter is large.



(a) *(b)*

Figure 3: *Estimates and standard errors obtained by the 'prepared' analyst under varied levels of obfuscation (γ) when the true underlying preferences are $\beta_1 = 1$ and $\beta_2 = 2$. The preference weights and the obfuscation parameter can be recovered without bias, but increased levels of obfuscation results in larger standard errors.*

In other words, from the choices made by obfuscating decision-makers, the prepared analyst can infer the presence and degree of obfuscation and the true attribute weights that the decision-makers attempted to hide from the onlooker. Figure 3b shows the standard errors of the estimates of the attribute weights (and that of the obfuscation parameter): these again increase as a function of the size of obfuscation parameter γ. This confirms the intuitive notion that a prepared analyst can spot obfuscation behaviour and simultaneously recover the true attribute weights of an obfuscating decision-maker, but with an increasing lack of precision as obfuscation becomes more pervasive.

3.4. SEQUENTIAL OBFUSCATION

In the obfuscation theory, there is a key assumption made by the DM that I aim to relax; that the onlooker takes into account *only* one (i.e. their current) choice when guessing their preferences (assumption number 4 in section 3.4). In real life, we might observe many choices of someone before inferring their preferences. We often have several observations from the same person in choice modelling either from an experiment or revealed preference data. Therefore, it is a logical extension of the obfuscation model to incorporate the possibility of sequential obfuscation to increase the model's behavioural realism. In the hereby proposed *sequential obfuscation*, assumption number 4 in section 3.4 is replaced by the following assumption.

4. The onlooker observed the decision-maker's past choices and already has an updated idea about the decision-maker's preference weights (β).

Formally this means two extensions to the process described in section 3.2.

- Equation 3.2 is multiplied with a term (called *likelihood*) to modify the question "assuming β^s, what is the probability of alternative i", to "assuming β^s, what is the probability of the sequence of the past observable choices and alternative i".
- The priors are updated continuously as the individual makes new choices. The posteriors in the first choice will become the priors in the second choice, or more generally, the posteriors in nth choice are the priors in the $(n+1)$th choice.

Equation 3.12 shows the formalized likelihood function. If the decision-maker made N choices, then the probability of observing the outcome-sequence $Y = y_1, ..., y_n$ assuming their preference corresponds to s state of the world (i.e. β^s) is

$$L(Y|\beta^s) = \prod_{n=1}^{N} P(y_n|\beta^s). \tag{3.12}$$

The sequential obfuscation model therefore uses the following equation instead of equation 3.2, when the DM faces the $N+1$th choice task:

$$P(a_j|\beta^s)_{N+1} = \frac{\exp(u_{j,s})}{\sum_{l=1}^{J}\exp(u_{l,s})} \prod_{n=1}^{N} P(y_n|\beta^s) \tag{3.13}$$

where y_n is the choice made in the nth choice task.

The posterior, entropy and final utility is then calculated based on this extension, formally using equations 3.1, 3.4, and 3.5. Then, after each decision, the posterior becomes the new prior for the next observation. This updating process formally is:

$$P(\beta^s)_{N+1} = P(\beta^s|y_n)_N \tag{3.14}$$

Note, however, that the DM assumes each update the onlooker makes is naive, meaning that the onlooker is not assumed to calculate with the obfuscation-intention (the onlooker ignores i.e., γ), only with previous choices.

ILLUSTRATIVE FIGURES

To illustrate what sequential obfuscation entails, compared to the original (from now on, for ease of communication: single choice) obfuscation model, I plot the entropies and prior updates of the above described thought process. At this stage, I consider a single decision-maker making 300 choices. There are three alternatives and two attributes randomly drawn from $[0,1]$ in each choice task. The possible states of the attribute weights (in the onlooker's, or more specifically, the decision-maker's mind) can be:

$$B = \begin{cases} \beta_1^1 = 0 & \beta_2^1 = 0 \\ \beta_1^2 = 1 & \beta_2^2 = 0 \\ \beta_1^3 = 2 & \beta_2^3 = 0 \\ \beta_1^4 = 0 & \beta_2^4 = 1 \\ \beta_1^5 = 1 & \beta_2^5 = 1 \\ \beta_1^6 = 2 & \beta_2^6 = 1 \\ \beta_1^7 = 0 & \beta_2^7 = 2 \\ \beta_1^8 = 1 & \beta_2^8 = 2 \\ \beta_1^9 = 2 & \beta_2^9 = 2 \end{cases}$$

from which set 8 (i.e. $\beta_1^8 = 1, \beta_2^8 = 2$) is the true underlying preference of the DM and $\gamma = 1$. First, I plot the entropies for 2 scenarios: one where the DM assumes only one of their choices is observed (and accordingly, single choice obfuscates; see equations 3.1-3.5), and one where DM assumes at every decision, that their past sequence of choices were observed (and accordingly sequentially obfuscates; see equations 3.12-3.14).

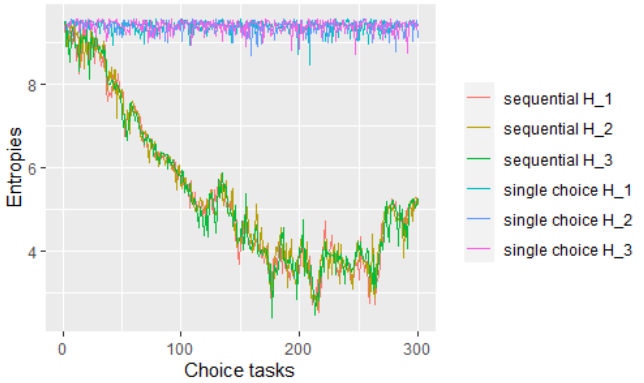


Figure 4: *Entropies of the 3 alternatives through 300 choice tasks, assuming single choice and sequential obfuscation.*

The entropies for each alternative show the same pattern: in the single choice case, they are stationary, while in the sequential case, there is a downwards trend. For both the single choice and sequential obfuscation, the decision-maker assumes that the onlooker updates their belief about the DM's betas but does it in a naive way (i.e. not calculating with γ). So the single choice obfuscator thinks the onlooker is naive and does not update their belief. The sequential obfuscator thinks the onlooker is naive but updates their belief after every observed decision.

Figure 4 shows that the sequential entropy tends to decrease as the number of choices made increases. The single choice entropies stay approximately the same. This is intuitive as, after many choices, the onlooker has more information than with one choice. This also means that the obfuscation part of the utility (γH_i) is decreasing; therefore, attribute weights are relatively more important in driving the decision.

In order to see how the sequential obfuscation handles potential onlookers, I now turn to the prior updates. I compare the sequential obfuscation strategy to the single choice obfuscation in the following way. I generate choices made using single choice and sequential obfuscation for the same choice tasks. Then, by plotting how a potential onlooker would update their beliefs, we can gain insight into what motivates a decision-maker to follow a sequential strategy over the single choice obfuscation when they believe all their choices are observed. The updating process means that after each choice, the onlooker has a belief about each of the nine beta sets; how likely it is that they are the true preferences of the DM. Figure 5 shows the evolution of posteriors under single choice obfuscation.

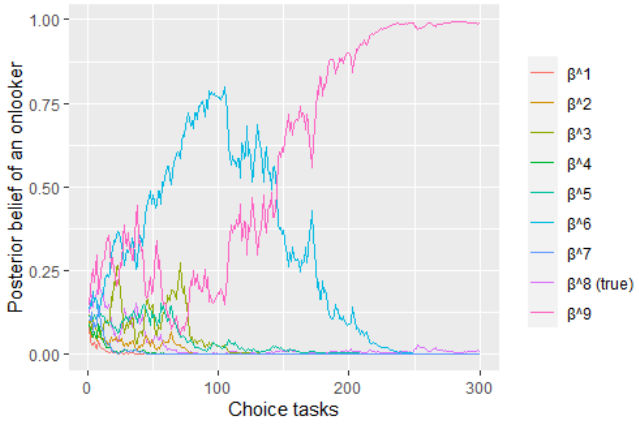


Figure 5: *Posterior updates of an observer if the decision-maker's strategy is single choice obfuscation ($\gamma = 1$).*

Next, figure 6 shows the same choice tasks, but with the decision-maker applying sequential obfuscation.

Figure 6: *Posterior updates of an observer if the decision-maker's strategy is sequential obfuscation ($\gamma = 1$).*

We can observe a general tendency in the long run, namely that 'taking over' is more frequent in sequential obfuscation[5]. This means that the onlooker often changes their belief about which is the most likely β, and very often, they would guess wrong. However, figure 5 also shows that with single choice obfuscation, the onlooker's beliefs potentially converge to the wrong βs. This effect is closer to deception, and if the decision-maker believes the onlooker updates their beliefs in such a way, in this example, sequential obfuscation masks their preferences without leading to deception. In the short-run, however (after 10-20 choices, which are typical in discrete choice experiments), we do not

[5]This finding is supported with running the simulations using different seeds and different sizes of γ.

see convergence (to 0 or 1). Testing the sensitivity to different priors, more specifically, setting the true weights' priors to values between 0.1 and 0.9 with steps of 0.1, we can see that in this particular setup, there is no significant difference between the belief-evolution, regardless of the initial priors. This means that initial beliefs in some situations can quickly dissolve[6].

3.5. IDENTIFICATION UNDER SEQUENTIAL OBFUSCATION

This section examines whether preference weights can be recovered under sequential obfuscation. Subsection 3.5.2 tests this under varied levels of obfuscation intention and under varied number of choice tasks. In both cases, I test the effects of sequential obfuscation strategy on discrete choice model estimates when 1, the analyst is unaware of obfuscation (i.e. 'naive') and 2, when the analyst models obfuscation with correct assumptions (i.e. 'prepared').

3.5.1. DATA GENERATION AND METHODOLOGY

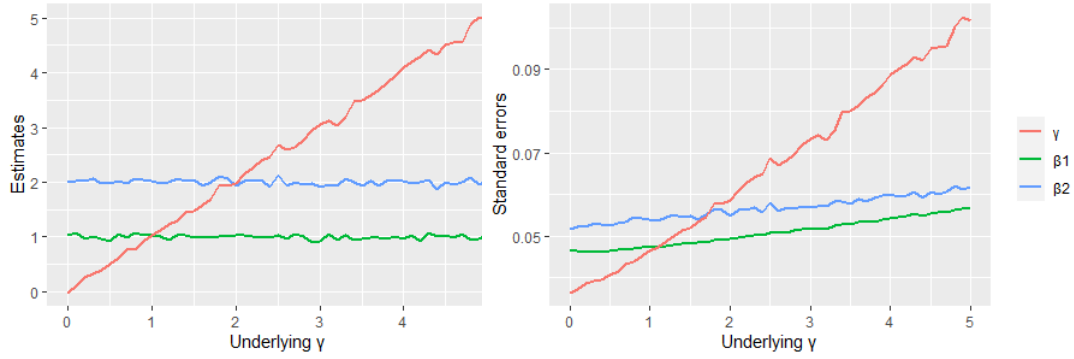
In this section, I use the same data generating process as described in section 3.3, with the following two modifications:

- The obfuscation strategy the decision-maker applies is sequential instead of a single choice.
- When the number of choice tasks varies (subsection 3.5.2), the number of individuals also varies to get datasets of approximately 7500 data points.

EXPECTATIONS REGARDING IDENTIFIABILITY ANALYSIS

Throughout this chapter, there is a key difference between the onlooker and the analyst: the onlooker observes past choices and uses Bayesian updates, while the analyst uses Discrete Choice Models (DCMs). Furthermore, the DM ignores the analyst, while the onlooker is not. When estimating the DCM, the analyst uses all observations from the DM simultaneously, when all choices are already made. If the DM were to obfuscate the analyst (and not the onlooker), they would need to use a completely different strategy. However, as we have seen in 3.3, the obfuscation strategy affects the analyst's estimates too. If the analyst has accurate assumptions about the underlying behaviour, they can recover the parameters accurately but with low certainty (i.e. high standard errors). I test the preference weight identifiability for the sequential obfuscation too, and in section 3.5 I also examine (besides how increasing levels of obfuscation affect parameter estimates) how an increased number of choice tasks affect estimation outcomes. As we have seen from figures 4 the sequential entropies have a lower magnitude in the sequential case than in the single choice case. This could mean that identifiability from an analyst's point of view is easier (i.e. more negligible bias or standard errors expected) in the sequential case than in the single choice case.

[6] Note that with other choice tasks, betas or gamma prior-sensitivity might be stronger.



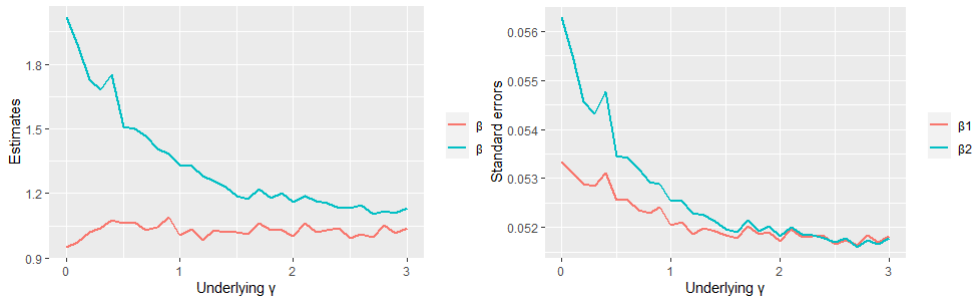
(a) *(b)*

Figure 8: *Figure a) shows the estimation results when the obfuscation parameter (γ) is also estimated and the true $\beta_1 = 1$ and $\beta_2 = 2$. Figure b) shows the corresponding standard errors.*

3.5.2. RESULTS

VARYING γ

NAIVE ANALYST



(a) *(b)*

Figure 7: *Figure a) shows the estimation results when the obfuscation is ignored and the true $\beta_1 = 1$ and $\beta_2 = 2$. Figure b) shows their corresponding standard errors.*

The estimates are biased, and the preference weights converge to a similar level. The standard errors, on the other hand (contrary to the single choice obfuscation case), decrease. Thus, ignoring the sequential obfuscation results in biased preference weight estimates and high certainty about them.

PREPARED ANALYST

Similar conclusions can be drawn when the analyst estimates the weights as in the single choice obfuscation case. The standard errors increase as the intention to obfuscate also increases, meaning the uncertainty about the estimates from the analyst's side is growing.

Comparing the standard errors of the sequential case and the single choice case (section 3.3) shows that the single choice case's estimates start with higher standard errors. However, as γ increases, the sequential case's standard errors become higher. This means that when one has a little intention to obfuscate, considering that only their current choice is observed generates higher uncertainty in an estimated model. However, if there is a larger preference for obfuscation, considering that choices are observed sequentially will generate higher uncertainty in an estimated model.

Figure 9: *Comparison of the standard errors generated by the two types of obfuscation under varying levels of obfuscation intention.*

VARYING THE NUMBER OF CHOICE TASKS

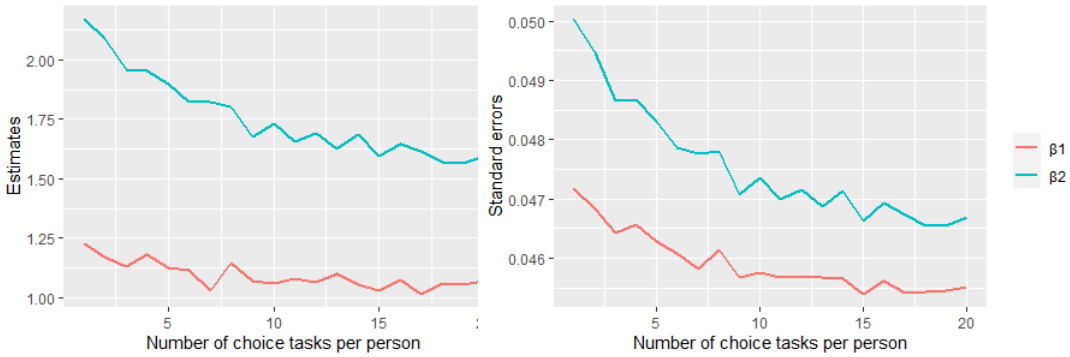
As the number of choice tasks increases, two effects play a role that potentially impact parameter estimation. First, the DM has more chance to obfuscate; if they made a decision that by design or chance was too revealing, they could offset it in subsequent choice tasks. Second, the more occasions they make decisions that reveal their preferences about specific attributes and their intention to obfuscate, the more data the analyst has; thus, the analyst can estimate with more accuracy.

Therefore, the recoverability of parameters under a varying number of choice tasks made by one individual is examined below.

I confirm but do not plot that the single choice obfuscation results in a horizontal line without trend. This means it does not matter whether 1 individual makes 20 choices or 20 individuals make 1 choice task. This is intuitive. For single choice obfuscation, the number of choice tasks is irrelevant.

NAIVE ANALYST

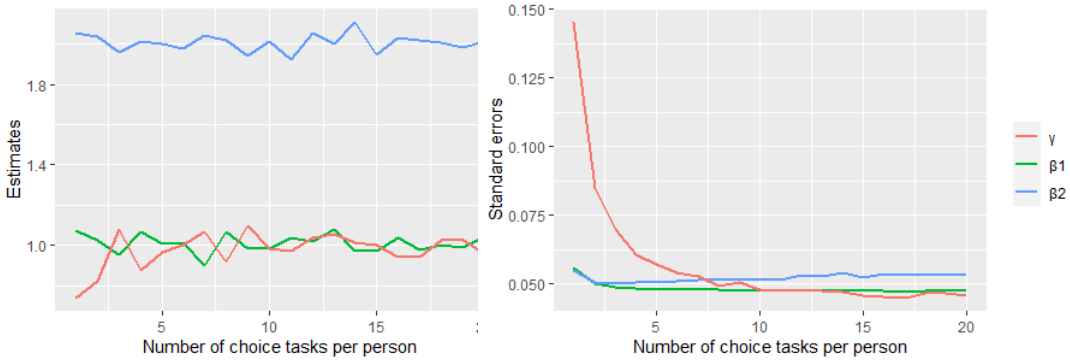
Figure 10 shows the estimates and standard errors for sequential obfuscation when the analyst ignores the obfuscating behaviour. The estimates are increasingly biased, and standard errors decrease.



(a) *(b)*

Figure 10: *Figure a) shows the estimation results when the obfuscation is ignored under sequential obfuscation (true $\beta_1 = 1$ and $\beta_2 = 2$). Figure b) shows their corresponding standard errors.*

PREPARED ANALYST



(a) *(b)*

Figure 11: *Figure a) shows the estimation results when the obfuscation parameter is also estimated under sequential obfuscation (true $\beta_1 = 1$ and $\beta_2 = 2$). Figure b) shows their corresponding standard errors.*

Figure 11 shows the estimates and standard errors with a varying number of choice tasks. As the number of choice tasks increases, the standard error of the obfuscation parameter decreases, first rapidly, then somewhat plateaus before the 20th choice task. Testing the robustness of this result, I generate the same plots for larger gammas (i.e. 3,5,7) and find the standard error plateaus in each case before the 20th choice task. Thus,

with sequential obfuscation, the highest possible level of certainty is reached within the typical number of choice tasks asked from one respondent in classical discrete choice experiments[7] (i.e. lower than 20).

3.6. DISCUSSION

Any obfuscating behaviour bears the question, whether the true preferences can be identified when modelling decision-making. Similar (yet not the same) strategies, such as making random choices or deception, prohibit the analyst to learn the true preferences of the decision-maker. In this study, I examined the identifiability of preference weights and obfuscation parameters in the recently developed obfuscation model (Chorus et al., 2021) and in its extension (i.e. sequential obfuscation) proposed in this chapter. There are four main conclusions of section 3.3 and 3.5 summarized in table 1.

	single choice DM	**sequential DM**
naive analyst	1. Increased levels of γ increase bias in estimates as well as their standard errors (section 3.3.2).	2. Increased levels of γ results in larger bias and lower standard errors. Increased number of choice tasks results in larger bias and lower standard errors (section 3.5.2)
obfuscation-expecting analyst	3. All preference weights (including γ) are recoverable, but as γ increases, the analyst is less certain about the estimates (i.e. standard errors increase, section 3.3.2).	4.All parameters can be recovered without bias. Increased levels of γ results in higher standard errors. Increased number of choice tasks results in lower standard errors for γ (section 3.5.2).

Table 1: *Summary of the results of examining preference weight identifiability under different conditions.*

When comparing the outcomes of the two obfuscation strategies, we can find the following.

- If the analyst disregards the obfuscation behaviour when it is actually present, the estimates will be biased. The higher the obfuscation intention is, the larger the bias will be. In the single choice obfuscation, the increased levels of obfuscation result in increasing standard errors (less certainty about the estimates). In contrast, in case of sequential obfuscation, the standard errors decrease (more certainty about the biased estimates).

- If the analyst takes obfuscation into account and has correct assumptions, both the single choice and sequential obfuscation model's parameters can be recovered without bias. When the two obfuscation strategies are compared, we see that single choice generates larger standard errors for minor obfuscation intention, but

[7] Although there is no theoretical limit on how many questions can be asked, respondent fatigue from too many choice questions can affect the data adversely (e.g., Johnson et al., 2013)

sequential does for higher obfuscation intention. Thus, the analyst is more obfuscated (note that the decision-maker's goal is not to obfuscate the analyst, but an onlooker, however, this has an effect on the analyst, too) by the single choice obfuscator when the obfuscation is less important and more by the sequential obfuscator when it is more important.

- Presenting the DM with more choice tasks also results in biased estimates with increasing certainty when the analyst is naive (similarly to the increasing level of obfuscation). For a prepared analyst, the certainty about the obfuscation intention grows as they see more choices by the DM. This finding is aligned with the standard notion of identifiability: obtaining more choices per individual helps identification (e.g., Cherchi and de Dios Ortúzar, 2008).

Contrary to the single choice model, the sequential obfuscation strategy creates a 'false' certainty, as larger bias can be related to lower standard errors. At this point, I discuss briefly how two other decision-making strategies, namely random choice and variety-seeking, can be mistaken for the entropy-based obfuscation strategies discussed in this chapter (Chorus et al., 2021). Random choice means that the decision-maker randomly picks an alternative without considering any attributes or their own preferences. Variety-seeking means that the decision-maker prefers alternatives that were not chosen before. These strategies could be mistaken for obfuscation, particularly in repeated choice contexts. However, their econometric implications differ from that of the obfuscation strategies. Estimating a model specified based on the alternative's attributes would make the analyst conclude that all attribute weights are zero, with high certainty (i.e. small standard errors); this would essentially entail deception, not obfuscation. The obfuscation behaviour presented in Chorus et al. (2021) and section 3.2 leads to very different conclusions: accurate or biased estimates (depending on whether the analyst prepares for obfuscation or not accordingly) with low certainty (i.e. high standard errors). The obfuscation strategy presented in section 3.4 of this chapter also leads to somewhat different conclusions: accurate or biased (but non-zero) estimates (depending on whether the analyst prepares for obfuscation or not accordingly) with high certainty (i.e. low standard errors). Due to the high certainty, sequential obfuscation is closer to deception from an analyst's point of view than the single choice strategy. However, the attribute weights are more informative, even though they are biased, than if they were zero.

This chapter established what effects do two obfuscation strategies have on discrete choice model estimation when the analyst is unaware and when they are aware of it. It needs to be emphasized that the analyses and conclusions presented in this section are to be interpreted with care: although they show that, in principle, the obfuscation behaviour of decision-makers need not prohibit the choice modeller from estimating their models without bias, further work is needed to show that these interpretations indeed hold in general, as opposed to only in the context of the carefully constructed Monte Carlo simulation exercise on synthetic data which was presented here.

Future research regarding obfuscation behaviour modelling should relax practical

assumptions of the model, such as the analyst knowing what possible preference weights are considered by the decision-maker (i.e., what DM believes the onlooker believes about their possible preference weights). This assumption (and its potential relaxation) also has relevance in accommodating heterogeneity in obfuscation. Heterogeneity in the obfuscation model goes beyond different preference weights; decision-makers can have different assumptions about their onlookers and what possible states of the world they consider. Chorus et al. (2021) presented the first empirical evidence for entropy-based obfuscation modelling; in laboratory settings, decision-makers displayed such obfuscation behaviour when incentivized to do so. Several situations in real life, such as political voting potentially trigger obfuscation, which can serve as a base for further empirical investigations.

Other, more complex obfuscation strategies may also be explored with discrete choice modelling tools. For example, decision-makers knowing in advance what decisions they have to make, may obfuscate their preferences for a whole sequence of choice tasks. This behaviour can be relevant, for example, when politicians vote on several different bills, or when someone strategically collects or ignores information before a choice is made. Such behaviour allows decision-makers to construct their own choice sets in a way that it obfuscates their preferences best. For instance, when people avoid online content that encourages them to welcome and help refugees (Freddi, 2015), they can say (to others or themselves) they were not aware that help is needed or they did not know what the problem's magnitude was. A similar example can be when governments privatize organizations that later found to be doing illicit actions, (such as the newly privatized Australian Wheat Board exchanging payments for wheat export with Saddam Hussein's Iraqi regime), the government can argue they were unaware of the organization's actions (Leong and Howlett, 2017; McConnell et al., 2008). These strategies allow the a decision-maker to construct or modify the choice task in which they have to decide, thus could be modelled simultaneously with the decision itself.

REFERENCES

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges, 2002*, 431–470.

Cherchi, E., & de Dios Ortúzar, J. (2008). Empirical identification in the mixed logit model: Analysing the effect of data richness. *Networks and Spatial Economics*, *8*(2), 109–124.

Chorus, C. G., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. *Mathematical Social Sciences*, *109*, 28–44.

Dunn, A. G., Mandl, K. D., & Coiera, E. (2018). Social media interventions for precision public health: Promises and risks. *NPJ digital medicine*, *1*(1), 1–4.

Freddi, E. (2015). Do people avoid morally relevant information? evidence from the refugee crisis. *Exploiting the unexpected in inflow of refugees to Sweden during*.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Johnson, F. R., Lancsar, E., Marshall, D., Kilambi, V., Mühlbacher, A., Regier, D. A., Bresnahan, B. W., Kanninen, B., & Bridges, J. F. (2013). Constructing experimental designs for discrete-choice experiments: Report of the ispor conjoint analysis experimental design good research practices task force. *Value in health*, *16*(1), 3–13.

Leong, C., & Howlett, M. (2017). On credit and blame: Disentangling the motivations of public policy decision-making behaviour. *Policy Sciences*, *50*(4), 599–618.

Lindbom, A. (2007). Obfuscating retrenchment: Swedish welfare policy in the 1990s. *Journal of Public Policy*, *27*(2), 129–150.

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis, new york, ny: John willey and sons.

McConnell, A., Gauja, A., & Botterill, L. C. (2008). Policy fiascos, blame management and awb limited: The howard government's escape from the iraq wheat scandal. *Australian journal of political science*, *43*(4), 599–616.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior.

Rossman, G. (2014). Obfuscatory relational work and disreputable exchange. *Sociological Theory*, *32*(1), 43–63.

Schilke, O., & Rossman, G. (2018). It's only wrong if it's transactional: Moral perceptions of obfuscated exchange. *American Sociological Review*, *83*(6), 1079–1107.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Wherry, F. F., Seefeldt, K. S., & Alvarez, A. S. (2019). To lend or not to lend to friends and kin: Awkwardness, obfuscation, and negative reciprocity. *Social Forces*, *98*(2), 753–793.

II

Enriching Discrete Choice Models with morality data

4

MORAL ASPECTS OF TRAVELERS' INTENTIONS TO PARTICIPATE IN A SOCIAL ROUTING SCHEME

While Part I examined novel models' identifiability issues when only choice data is available, Part II of this thesis focuses on the use of additional data to identify antecedents of moral decision-making. This chapter uses an empirical study to investigate how general moral foundations, contextual moral motivations, and different moral incentives affect decision-making in a social routing context.
This study uses a stated choice experiment with two treatments, standard morality surveys and contextual moral motivation questions and model whether decision-makers intend to join a social routing scheme or not. Section 4.1 introduces the motivation for this empirical study, and section 4.2 presents the theoretical differentiation between the two social routing schemes and how they relate to morality. Section 4.3 introduces the data collection and methodology in detail. Section 4.4 shows the results and discusses the interpretation. Section 4.5 concludes with practical implications and policy recommendations.

This chapter is based on the paper entitled 'Give and take: Moral aspects of travelers' intentions to participate in a social routing scheme' by Teodóra Szép, Tom van den Berg, Nicolas Cointe, Aemiro Melkamu Daniel, Andreia Martinho, Tanzhe Tang, and Caspar Chorus[1] currently under revision.

4.1. INTRODUCTION

In and around most cities in the urbanized world, governments are struggling with congested road infrastructures. As is by now well recognized, an important key to a solution to this problem lies in influencing traveler behavior in such a way that travel demand is more evenly spread across available network capacity. Such a 'system optimal' distribution of traffic would generate large gains in accessibility and travel times, compared

to the user equilibrium that arises when all travelers behave independently without any form of coordination. Various approaches have been tried to reduce this so-called 'price of anarchy' and move towards a better distribution of travel demand, including regulation, information provision, pricing and other incentive schemes (e.g., Albalate and Fageda, 2019; de Palma et al., 2018; Knockaert et al., 2012; Liu et al., 2017; Noordegraaf et al., 2014). Unfortunately, these policies all suffer from a trade-off between effectiveness and public acceptance: the most effective schemes (regulation and pricing) are unpopular among the general public (Gu et al., 2018; Krabbenborg et al., 2020), while the schemes that enjoy higher levels of acceptance among the public (such as information provision and soft incentive schemes) are considerably less effective in redistributing traffic (Chatterjee and Mcdonald, 2004).

The idea of social routing schemes is believed to combine relatively high levels of effectiveness and acceptance. The idea behind such schemes is that car users voluntarily agree, every once in a while, to choose a different route (or departure time, or even travel mode) with a somewhat higher travel time than their normal route, for the benefit of the system at large. A recent flurry of research has explored such policies (Djavadian et al., 2014; Klein et al., 2018; Koller, 2021; Kröller et al., 2021; Mariotte et al., 2021; van Essen et al., 2020; Van Essen et al., 2016), and the general consensus is that they indeed have the potential to deliver sizable gains in travel times by inching closer towards a system optimal distribution of traffic (Chen et al., 2021; Çolak et al., 2016; Eikenbroek et al., 2021; Van Essen et al., 2019). Many open questions remain, though; this paper aims to help find answers to some of those.

Particularly, we address an aspect of social routing schemes that hitherto has not received the attention it deserves: the role of morality, as in the 'moral personality' of travelers and the 'moral motivations' behind their choices, and how these interact with the characteristics and framing of the social routing system. Our first contribution to the literature lies in the use of a widely established morality scale (Moral Foundations Questionnaire or MFQ; Graham et al., 2009) to measure the moral orientation of travelers. The MFQ provides a multi-dimensional picture of someone's morality. We show how the various dimensions of travelers' moral personalities are associated with their stated intentions to join a social routing scheme. In addition to the MFQ, which measures deep-seated moral values and convictions, we also use more specific morality-related questions targeted at the particular choice situation at hand – whether or not to join a social routing scheme. Whereas the MFQ measures, in a general and abstract sense, moral personality, the more specific questions measure contextual moral motivations. Recent literature suggests that the former measurements are more stable but less predictive of actual behavior, than the latter (Kroesen and Chorus, 2018); we set out to explore, amongst other things, whether this holds for the context of social routing and we aim to identify which personality- and motivation-related aspects help determine travelers stated intentions in this morally sensitive context.

Our second contribution lies in the way the social routing scheme is characterized and framed towards travelers. In addition to conventional characteristics (such as the

difference between the travel time of the social alternative and that of someone's normal travel alternative), we distinguish between so-called 'sacrifice-based' and 'collective good' schemes. The former asks the traveler to make a personal sacrifice – in terms of a longer travel time – for the greater good, while the latter asks the traveler to join a collective endeavour that will result in lower travel times for themselves as well as for others in the long run. Given that traffic is often conceptualized as a matter of collective action, it is somewhat surprising that studies into travellers' acceptance of social routing, so far, have not investigated the collective action nature of such a scheme in detail. Theoretically, we offer a new, collective good- perspective on social routing that seems more aligned with the nature of traffic.

Our third contribution lies in highlighting, using a combination of conceptual and empirical analysis, the interactions between travelers' moral personality and moral motivations on the one hand and the framing of the social routing scheme (sacrifice-based versus collective good) on the other hand. We show how different dimensions of a traveler's moral personality and motivations influence in different ways their stated intention to join a scheme, depending on how the scheme is framed. Research in moral psychology helps us to interpret these interactions in meaningful ways.

The empirical part of our study is based on a large-scale, two-wave data collection effort, consisting of a stated intention experiment and a morality survey. Resulting data are analyzed using a series of advanced discrete choice models. Together, these empirical analyses allow us to tentatively derive implications for policy makers as to the role of moral aspects in the optimal development and implementation of social routing schemes.

The remainder of the paper is structured as follows: section 'Theoretical background' describes the fundamental differences between the two types of social routing schemes considered in this study and the moral motivations they aim to trigger. Section 'Data and methodology' presents the stated intention experiment and the morality survey and touches upon data collection and sample aspects. Section 'Empirical results and interpretation' presents the model estimations and interprets them in light of theories and notions from moral psychology. Section 'Conclusions and implications' summarizes the main takeaways from our study, translates these into tentative policy recommendations, and suggests avenues for further research on the topic.

4.2. THEORETICAL BACKGROUND

To move people towards making the social choice in their route choice and reach a system optimum, at least three different motivations appear in the literature: self-interest, altruism, and free ride avoidance, or more specifically, fairness. The first of these has received wide attention so far in the literature. The second motivation has come up in the context of investigating to what extent people are willing to choose the social route when there are no external incentives but only the right information. For example an alternative route would be provided and it would be indicated that, although the suggested route may be longer for the individual personally, it will contribute to saving travel time

on the collective level. The third motivation of free ride avoidance and the different effects it has compared to altruism in a social routing system, seems to have received little attention so far. In the following, we argue that the motivation of free ride avoidance or fairness can be used to build a 'collective good based' social routing system, and that it can be a viable way to move people towards choosing the social route and reach the system optimum without relying on external incentives.

To get a better understanding of an individual's moral motivation to contribute to a collective good and not free ride it seems helpful to consider normative accounts of free riding given by moral philosophers. Cullity (1995) explains that the free rider gives herself an objectionable preferential treatment "in allowing herself not to pay for goods that she either does or ought to realize are worth paying for, and that she only receives because others are moved by the same realization to pay". This amounts to unfairness. Giving yourself a preferential treatment in collective good situations that cannot be reasonably justified is wrong even when no-one is directly harmed. Building on this, we establish that in order to target the free riding avoidance motivation in a social routing system we have to (1) make it clear to travelers that the system is a collective good, where everyone has to contribute to achieve success and (2) make sure that there is no reasonable justification for free riding.

The first condition can be met with information provision. In order to meet the second condition, there must be a fair distribution of costs and benefits among beneficiaries (otherwise those who contribute but do not benefit, or benefit significantly less than others, could object against the scheme) and the individual costs should not be higher than the individual benefits in the long run (otherwise the scheme is not worth to participate in at all). In situations where the above two conditions are met, free riding should be regarded as unfair because one profits from other people's contributions without contributing oneself and there is no reason to justify it. According to Cullity (1995) this is unfair even when the scheme has been imposed on someone involuntarily. As long as the scheme is fair and participation is overall beneficial, one should pay their share for the benefits provided by the scheme. Based on this normative account, it can be expected that when a collective good situation is as described above, considerations of fairness -in the sense of not wanting to unjustly profit from other people's contributions- may play a role in the decisions that people make. On the other hand, when the above conditions are not met, we have an altruistic, or 'sacrifice-based' scheme at hand. If there is no fair allocation of the benefits or benefits for the individual in the long run, a contributing individual is most probably driven by an altruistic motivation like that of 'care': sacrificing one's own good for the benefit of others. Not joining a sacrifice-based scheme may indicate a lack of care, but does not amount to unfairness. On the flip side, participating in a collective good scheme does not amount to altruism, but rather being fair.

Considering these two distinct moral motivations and taking into account that a participant of a social routing scheme will participate over a longer period of time consisting of recurring longer trips, it may be expected that a collective good scheme is the more viable system. In this system the individual profits oneself from the generated collective

good in the long run, while non-participation remains unfair. A sacrifice-based scheme that runs solely on information provision without assurance of individual benefit, can only count on people's altruism or care. It seems questionable, at the least, whether this could sustain long term participation. Arguably, in this context, the motivation of fairness harbours a stronger social and normative force than that of altruism.

Some of the recent literature also investigated different motivations in social routing systems, but mostly relying on information-level differences (Klein and Ben-Elia, 2018; van Essen et al., 2020). Following our theoretical framework, many of the information based social routing systems that are considered in the literature should be categorized as sacrifice-based schemes. For instance, van Essen et al. (2020) conceptualizes the social choice as one that entails "personal travel time sacrifice for the benefit of others" (p.1048). The design of the stated choice experiments does not include the recurrence of trips and thus lacks the assurance of individual benefit in the long run -making the motivation it triggers altruistic without testing its (lack of) sustainability in the long run. The revealed choice experiment of this study does include a recurrence of trips and also a principle of distributing the costs (28 participants are asked to drive the longer route to work for two days a week). Though this gets closer to a collective good scheme, the crucial assurance of individual benefit over the long run that needs to be in place is lacking and -importantly- there is no experience of profiting from other people's contributions when not complying -i.e. free riding.

Other literature does more or less construct the social routing scheme as a collective good scheme but miss, or at least do not thematize, the specific moral motivation that plays a role here. Klein and Ben-Elia (2018), for instance, do explicitly take the recurrence of trips and the fair distribution of costs and benefits into account in their investigation of social routing systems. They argue that if these conditions are met and the individual benefits in the long run there is no need to rely on the unreliable motivation of altruism. The scheme should be regarded as a repeated game in which it is in people's own self-interest to cooperate and produce the collective good. However, as the authors make clear, when the group size in a repeated game increases -like in a social routing scheme- cooperation becomes less likely. In their experiment they test whether cooperation can be sustained through triggering an 'intrinsic motivation' by providing the information that following the recommended routes will lead to shorter average travel time for everyone in the end. However, what the 'intrinsic motivation' exactly entails here is not explicated. If it still refers to a form of self-interest it does not suffice for compliance. First of all, it is questionable whether in reality the individual gain in travel time is noticeable for the individual herself, especially given the variation of travel time due to random everyday incidents (van Essen et al., 2020). Secondly, and more fundamental, pure self-interest within a collective good scheme leads to free-riding. Hence, a moral motivation must be assumed here that is not made explicit nor is further conceptualized: fairness or free-ride avoidance. Assuming that freeride avoidance plays a role here instead of mere self-interest also solves the first problem: even if travel gains are not noticeable for participants who sometimes drive longer routes, this seems less plausible for free riders who always take the shorter route. At least they should notice a reduc-

tion in congestion. As this amounts to profiting without contributing, the motivation of free ride avoidance can, theoretically, still play a role in steering individuals to compliance while self-interest cannot. The social routing system as a collective good scheme with the specific moral motivation of fairness or free-ride avoidance -though sometimes partly or implicitly assumed in the literature- has, so far, not been explicitly conceptualized nor empirically investigated and been compared to the more frequently relied on sacrifice-based scheme.

In our study -building on our conceptual framework- we explicitly focus on these distinct moral motivations of altruism and free-ride avoidance. Although theoretically speaking, free-riding is primarily a violation of the principle of fairness and not making sacrifices for others a lack of altruism or care, the question on what basis individuals in real life make these choices is empirical. In our empirical study we therefore focus on the above described moral motivational differences and their aspects such as how much contribution is asked from the individual and how others behave under the social routing system. The following section describes our experimental approach and methodology.

4.3. DATA AND METHODOLOGY

Our experimental approach has two main parts: a stated intention experiment (first wave) and a morality survey (second wave). For the first wave we designed the stated intention experiment the following way: participants are asked whether they would join a social routing scheme with specific attributes. The response is binary, yes or no. The attributes of the social routing scheme are:

- number of days, out of 10, on which the commuter will be asked to use the social route (levels can be 2 and 4 days out of ten),
- average additional travel time, representing the number of minutes the social route is slower than the non-social alternative (levels can be 3 and 7 minutes),
- total travel time saved in the system over 10 days if the commuter participates in the scheme (levels can be 40 and 75 minutes), and
- participation rate, which indicates the percentage of fellow road users that join the scheme (levels can be 20% and 80%).

Following our theoretical framework, we test the difference between sacrifice-based and collective good based incentives by embedding the above characteristics into two different schemes. The 'Sacrifice-based scheme' makes no mention about potential gain for the respondents themselves. The 'Collective good scheme' presents the decision as whether the respondent is willing to contribute to an outcome that is beneficial to all travelers, including the respondent. We highlighted that the overall travel time for the participating travelers is lower with the scheme than without it. We do not guarantee this in the Sacrifice-based scheme. Aside from this, the two presented schemes are the same. Each participant is assigned to one treatment (Sacrifice-based versus Collective

good), and answers 16 questions with varying attributes within that scheme. As such, each participant evaluates all possible attribute level combinations in a full factorial design. Figure 1 shows an example choice task for both schemes. Note that this approach goes beyond a mere framing exercise: the two schemes are inherently different in the sense that the collective good scheme would be designed in order to make everyone better off, including those who regularly choose the social route. It is also made clear that because the scheme was implemented one's regular route to work has become faster. Hence, not participating amounts to free riding as described in section 4.2. In order to prevent an overload of information in the choice tasks, we did not add a statement on the fair distribution of costs and benefits. As benefits are probably unnoticeable, their differences across travelers are even smaller, therefore making sure there are no losers implies the system is more or less fair without going into complicated details of the benefit-distribution.

After the choice tasks we asked the respondents about their motivations when making their decisions. Respondents indicated on a Likert-scale from 1-5 how important the following motivations were for them when making their decisions:

- "To do something for my fellow road users",
- "Make sure that others don't profit from my personal contribution",
- "Help solve congestion for me and my fellow road users",
- "Make sure that I do not profit from other road users' contributions while not contributing myself",
- "Ensure that my own travel time is minimized".

In the second wave, we collected data on the moral character of respondents using the widely used Moral Foundations Questionnaire (MFQ). MFQ is built on the Moral Foundations Theory (MFT, Graham et al., 2009), which argues that at least five basic 'moral foundations' are the same across people and cultures. Moral characters only differ in the extent to which they value these basic foundations. Namely, these foundations are care / harm, fairness / cheating, loyalty / betrayal, authority / subversion, sanctity / degradation. We use the MFQ with 30 questions and statements where respondents choose to what extent they agree with a statement or to what extent something is crucial for them when making a moral decision.

We collected the data from a representative panel of the Dutch population in 2021, for the first wave in March, for the second wave in April[2]. Travelers who commute by car and are above 18 years old were recruited. Respondents first filled in the choice experiment, then two weeks later the MFQ. As the MFQ has two control questions, we use these to detect inattentiveness. Similarly to Viđak et al. (2020), we also use the following rule: those who reply 2 to 5 to question 6 (meaning it is from somewhat to extremely relevant to them whether or not someone was good at math when making a judgment of moral right and wrong) or 1 to 3 to question 21 (meaning they firmly to slightly disagree with

[2]The study was approved by the human research ethics committee (case number of application: 1039).

Imagine that the following social route system has been introduced in your traffic network. You are not yet a member of the social route system. Would you participate in a system with the following characteristics for six months?

- Out of every ten days, you will be asked to drive a short distance to or from work for **4 days.**
- This will cost you approximately 7 **minutes** extra travel time at a time.
- When you participate, the network benefits. The total reduction in travel time – calculated over ten days – is **75 minutes**. This is distributed across the network. (+)
- About **80%** of your fellow road users participate in the system.

(+) ***Note:*** *the total travel time of the network is the sum of all travel times of all road users added together. Therefore, the aforementioned decrease is not the travel time gain per individual road user but that of the network as a whole.* *

○ I would join

○ I would not join. I will always choose my regular route

Imagine that the following social route system has been introduced in your traffic network. **Your regular route to work has become faster as a result.** You are not yet a member of the social route system. Would you participate in a system with the following characteristics for six months?

- Out of every ten days, you will be asked to drive a short distance to or from work for **4 days.**
- This will cost you approximately 7 **minutes** extra travel time at a time.
- When you participate, the network benefits. The total reduction in travel time – calculated over ten days – is **75 minutes**. This is distributed across the network. (+)
- About **80%** of your fellow road users participate in the system.

Even if you participate in the social route system, you save travel time compared to the situation before the introduction of the system (calculated over the six months that you participate).

(+) ***Note:*** *the total travel time of the network is the sum of all travel times of all road users added together. Therefore, the aforementioned decrease is not the travel time gain per individual road user but that of the network as a whole.* *

○ I would join

○ I would not join. I will always choose my regular route

Figure 1: *Two example choice tasks of our stated choice experiment. The first column shows the sacrifice-based scheme, the second column shows the collective good scheme.*

doing good is better than doing bad) were excluded from the analysis. Our final sample consisted of 786 respondents (395 in the altruism frame and 391 in the collective good frame) and 12576 choice tasks. In this data that we used for our analysis, 46% of participants are female, the average age is 45.3, and the mean of their average trip to work is 27.8 minutes.

The choice experiment and morality survey were analyzed using Discrete Choice Models (DCMs, for an extensive overview, see Train, 2009). For benchmark, we use the binary logit model, a regression model where the dependent variable is binary, in this case whether or not someone joins the social routing scheme. The explanatory variables are the specifics of the scheme (additional travel time, number of days to drive the longer route, travel time benefit for all, and participation rate). The binary logit cannot account for random taste variation (i.e., taste differences that cannot be linked to observed determinants). In order to account for such random taste variation, we use panel mixed logit models which allow us to estimate not just one taste parameter for the population, but also a distribution for them. The following section shows our estimation results.

4.4. EMPIRICAL RESULTS AND INTERPRETATION

We first present a simple base model to set the stage for our analyses, see Table 1. We estimate a binary logit model on the combined data of the two schemes. We use the linear additive form for the utility. Each attribute weight includes a dummy indicator that takes a value of 1 for responses from the Collective good scheme and 0 for responses from the Sacrifice-based scheme. The utility specification can be found in the Appendix. We directly obtain all parameter estimates and standard errors for the Sacrifice-based scheme and the indicator terms (indicating the respective difference between the two schemes). Then we obtain the Collective good scheme estimates by adding the Sacrifice-based scheme's corresponding estimates and the differences. The standard errors are calculated using the Delta method.

This binary logit model predicts the stated intention to join a social routing scheme with particular attributes, as a linear function of the attribute values. We distinguish between the two schemes (Sacrifice-based versus Collective good), to explore whether sensitivities to attributes are specific to a particular scheme. It may be noticed that all parameters have the expected sign, with Number of days and Additional travel time being valued negatively and Network travel time benefit and Participation rate being valued positively. All but one parameter are significant at a 1% level: Network travel time benefit is not significant at conventional levels of confidence. The only significant difference between the two schemes is found for the attribute Additional travel time, which is valued more negatively in the Sacrifice-based frame than in the Collective good frame. This difference is intuitive, as the Collective good frame promises travelers that they will not be worse off compared to the situation without a social routing scheme in place, even though for particular days they may experience a slightly longer travel time than they would have, if they would not have joined the scheme. To get an idea of the implied sensitivity of the different attributes, we computed the predicted probability that a randomly sampled traveler would intend to join a social routing scheme with particularly

	Sacrifice-based	Collective good	Differences
	Est.(SE)	Est.(SE)	Est.(SE)
ASC_{SR}	1.949 (0.176)***	1.663 (0.159)***	-0.287 (0.237)
Number of days	-0.235 (0.025)***	-0.202 (0.024)***	0.033 (0.035)
Additional travel time	-0.272 (0.019)***	-0.220 (0.018)***	0.051 (0.027)***
Network travel time benefit	0.0003 (0.001)	0.0002 (0.001)	-0.0001 (0.002)
Participation rate	0.0049 (0.001)***	0.0055 (0.001)***	0.0006 (0.001)
Estimated parameters (k)	10		
McFadden ρ^2	0.058		
Final-loglikelihood	-8199.9		
Number of choices	12576		

Table 1: *Binary logit model of differences in the two schemes. The model is estimated on the combined data of both schemes. The corresponding systematic utility function of differences can be found in the Appendix. *, ** & ***, respectively represent significance at 10%, 5% 1% levels.*

unattractive versus particularly attractive attributes under a particular frame. Penetration rates vary between 31% and 75% for the Sacrifice-based frame, and between 36% and 74% for the Collective good frame. This suggests that the latter scheme is slightly more popular, which is in line with the observed empirical frequencies in the dataset.

As can be seen when inspecting McFadden's rho-squares, the model fit of this base model is rather poor compared to that of Table 2, suggesting that the incorporation of panel effects (i.e., acknowledging that choices made by one individual may be correlated) in combination with heterogeneity in tastes could lead to a more realistic model. The results of such a panel mixed logit model are presented in Table 2. The mixed logit model is also estimated on the combined data of the Sacrifice-based and Collective good schemes. However, in the mixed logit specification, we allow for differences between the two schemes in terms of mean and standard deviation estimates for each attribute, including the constant for joining the social route. More specifically, we include a dummy variable defined exactly as previously in the binary logit model and interact it with each attribute's mean and standard deviation. A significant difference in the mean estimates for an attribute indicates that the specific attribute has a different effect (on the decision to join the scheme) under the Collective good scheme and the Sacrifice-based scheme. On the other hand, a significant difference in the standard deviation estimates for an attribute informs about variations in the level of heterogeneity in the attribute's effect (on the decision to join the scheme) across the two schemes.

After exploring various distributional assumptions, all parameters are modeled with a normal distribution, which proved to lead to the most stable convergence. As a first observation, the model fit improves greatly, suggesting that as expected, panel effects and heterogeneity are important factors behind the choices made by participants. Signs and significance levels are the same as in the binary logit model (sensitivity to Network travel time benefit again being the only non-significant effect); additionally it is found that all

	Sacrifice-based		Collective good		Difference in mean	Difference in std.dev
	Est. (SE)	Std.dev (SE)	Est. (SE)	Std.dev (SE)	Est. (SE)	Est. (SE)
ASC_{SR}	7.485***	6.111***	6.191***	5.985***	-1.29***	-.126
	(.564)	(.436)	(.521)	(.458)	(.493)	(.116)
Number of days	-.984***	.975***	-.811***	.86***	.174	-.111
	(.117)	(.163)	(.092)	(.108)	(.143)	(.155)
Additional travel time	-1.123***	.891***	-.869***	.943***	.254***	.052
	(.086)	(.093)	(.076)	(.078)	(.094)	(.115)
Network travel time benefit	.004	.054***	.003	.035***	-.001	-.019**
	(.004)	(.008)	(.004)	(.004)	(.001)	(.009)
Participation rate	.022***	.034***	.023***	.038***	.002	.004
	(.003)	(.004)	(.003)	(.003)	(.004)	(.003)
McFadden ρ^2	.51					
Final-LL	-4258					
Estimated parameters (k)	20					
Number of choices	12576					

Table 2: *Mixed Logit model for the differences in the two frames. The model is estimated on the combined data of both schemes. A normal distribution is assumed for all explanatory variables[3]. The corresponding systematic utility function of differences can be found in the Appendix. *, ** & ***, respectively represent significance at 10%, 5% 1% levels.*

parameters come with high and significant levels of heterogeneity. This implies that the variation within the sample in terms of sensitivity to the attributes of the proposed routing scheme, is considerable. As in the binary logit model, we find that the sensitivity to Additional travel time is, again intuitively, greater for the Sacrifice-based scheme than for the Collective good frame.

To explore whether or not, in what ways and to what extent, moral personality as measured by the MFQ plays a role in explaining stated intentions for joining the social routing schemes, we interacted five morality-dummies with the constant that captures travelers' average inclination to join the social route. For the utility specification see Appendix. Note that the use of dummies was motivated by model stability considerations, as was the decision to not estimate a standard deviation for the constant simultaneously with the morality-interactions. Each dummy represents a particular moral dimension (Care, Fairness, Ingroup, Authority, Purity); whenever the individual would score at least 24 out of 30 points for a particular dimension, the corresponding dummy would take on the value of 1. Note that each dimension was measured by means of six questions, each having answer categories ranging from 0 (not at all relevant or strongly disagree) to 5 (extremely relevant or strongly agree). As such, each dummy is informative of whether or not someone scored very high on the corresponding moral dimension, implying that they believe that the particular moral foundation is key to their personal morality. Models were estimated using the Apollo package (Hess and Palma, 2019) in R; we used 16,000 MLHS draws (Hess et al., 2006) for the random parameters, after verifying that results were similar to models with half that number of draws. Results are presented in Table 3.

These outcomes can be summarized as follows: under the Sacrifice-based frame, whether someone strongly adheres to the Care foundation has a significant (at 1% level) and sizable positive effect on their intention to join the social routing scheme. Other moral dimensions do not have a significant effect. Under the Collective good frame, Fairness has a significant (at 5% level) and positive effect of moderate size, while Ingroup has a significant (at 5% level) and negative effect. Note that the effect of Ingroup is also negative, but not significantly so, under the Sacrifice-based frame. Authority and Purity do not have significant effects on stated intentions to join the social routing scheme, under either frame. We consider the differential effects of Care and Fairness under the two schemes an important and intuitive result: as conceptualized, a social routing scheme that is designed and framed as a sacrifice-based scheme in which travelers make personal sacrifices for other travelers, taps into the Care dimension of people's morality, making those that strongly adhere to this dimension particularly susceptible to joining the scheme. In contrast, the Care dimension does not seem to play a role when the scheme is designed and positioned as a collective good to which all are expected to contribute to the common good. In this frame, fairness is a leading factor, implying that travelers who strongly adhere to the Fairness dimension or morality are particularly likely to join the Collective good scheme. Given the nature of the two schemes, this result is intuitive, in the sense that not participating to a Collective good scheme (as opposed to not joining a Sacrifice-based scheme) may be considered as unfair: the scheme is beneficial to the traveller -joining the scheme would make the traveler better off than before the scheme was implemented-, not joining would thus amount to unjustly benefiting from other people's contributions without contributing oneself and, hence, a form of free riding. These empirical findings lend support to our theoretical exposition (presented in section 4.2) regarding the difference between sacrifice-based social routing schemes and collective good schemes, in terms of the moral motivations they tap into.

To grasp the negative effect of Ingroup on joining the social routing scheme (under both frames, but only significantly so under the Collective good scheme), it is good to look at the particular questions used to measure this dimension in the MFQ: these relate to loyalty to e.g. family, implying a distinction between in-group and out-group loyalty. It has been argued in another travel behavior context (van den Berg et al., 2020) that this particular definition and measurement of Ingroup morality could actually imply that those who score high on this dimension, are less willing to collaborate with or care for strangers outside their own in-group, as would be the case in joining a social routing scheme. As such, the negative association found in our experiment is in line with previous findings and interpretations.

It has been suggested that the moral values elicited by the MFQ are so general, abstract and deep-seated that they make poor predictors of concrete moral behaviors in real life (Kroesen and Chorus, 2018; van den Berg et al., 2020). Our results do find meaningful associations, which is probably partly due to the fact that our measured 'behaviors' are actually stated intentions in a rather abstract experiment setting. Nonetheless, we also explore associations with more contextually related moral motivations, which we operationalize by means of five questions. (note that these motivational questions were

	Sacrifice-based		Collective good	
	Est. (SE)	Std.dev (SE)	Est. (SE)	Std.dev (SE)
ASC_{SR}	5.490***		4.485***	
	(0.303)		(0.335)	
Number of days	-0.811***	0.861***	-0.698***	0.772***
	(0.077)	(0.083)	(0.076)	(0.094)
Additional travel time	-0.949***	0.737***	-0.755***	0.932***
	(0.062)	(0.080)	(0.067)	(0.091)
Network travel time benefit	0.001 (.009)	0.060***	0.005	0.051***
		(0.007)	(0.005)	(0.005)
Participation rate	0.017***	0.029***	0.021***	0.036***
	(0.003)	(0.003)	(0.003)	(0.004)
ASC_{SR} x Care	2.047***		0.479	
	(0.615)		(0.558)	
ASC_{SR} x Fairness	0.807		1.340**	
	(0.770)		(0.665)	
ASC_{SR} x Ingroup	-1.418		-2.845**	
	(1.276)		(1.415)	
ASC_{SR} x Authority	0.819		2.184	
	(1.023)		(1.683)	
ASC_{SR} x Purity	-1.083		1.486	
	(1.053)		(1.314)	
McFadden ρ^2	0.49		0.50	
Final-LL	-2215.5		-2165.9	
Estimated parameters (k)	14		14	
Number of choices	6320		6256	

Table 3: *Mixed logit models estimated separately for the two schemes, using MFQ interactions with the alternative specific constants, or in other words, the predisposition to join the social routing system (ASC_SR). See Appendix for the utility specification. The four scheme-specific attributes are assumed to have a normal distribution. Alternative specific constants are interacted with moral dummies (being 1 if the cumulative score is at least 24 out of 30 points). *, ** & ***, respectively represent significance at 10%, 5% & 1% levels.*

asked directly after the choice experiment in contrast with the moral foundation questions which were asked in the second wave administered two weeks later; this provides another reason why we would expect the answers of the motivational questions to correlate relatively strongly to the stated intentions to join a scheme) The resulting answers, on a Likert scale ranging from 1 to 5, are taken to be proxy measurements of five moral motivations for joining (or not) the social routing schemes presented in the experiment. We label these motivations as: Altruism ("To do something for my fellow road users"), Competition ("Make sure that others don't profit from my personal contribution"), Common good ("Help solve congestion for me and my fellow road users"), Fairness ("Make sure that I do not profit from other road users' contributions while not contributing myself"), and Individualism ("Ensure that my own travel time is minimized"). For each dimension, a dummy was created to identify those who strongly identify with a particular motivation, i.e. scored a 5 on the corresponding Likert scale. Models were estimated using 16,000 MLHS draws for the random parameters, after verifying that results were similar to models with half that number of draws.

Results are presented in Table 4 and can be summarized as follows: as expected, we over-all find stronger effects for these more contextual motivations, than we did for the generic moral personality dimensions. A clear distinction can be observed between the effects under the Sacrifice-based frame versus the Collective good frame: under the Sacrifice-based frame, Individualism is negatively associated with joining the social routing scheme and Altruism and Common good are positively associated with joining. These relations are intuitive. Competition and Fairness do not have a significant effect, although their signs are as expected. Under the Collective good frame, Individualism and Altruism are not associated with joining the scheme (but note that signs are as expected). Just like in the Sacrifice-based scheme, Common good is positively related to joining the scheme under the Collective good frame. Under this frame, Competition (negative) and Fairness (positive) are both significantly related to the intention to join the scheme, whereas these had no significant association under the Sacrifice-based scheme.

These differential associations between moral motivations and the stated intention to join the social routing system under the two distinct schemes, are in line with intuition as well as the conceptualizations presented further above. Since the Sacrifice-based scheme emphasizes the sacrifice made for other road users, this resonates with people whose motivation to join is driven by altruistic and common good related motivations, and it scares off those people with individualistic motivations (ensuring low travel times for themselves). In contrast, the Collective good scheme emphasizes the notion that also participants benefit from the scheme, attracting those for whom fairness and contributing to a common goal (fighting congestion) is important. The Collective good scheme does not scare off people with individualistic motivation as much (although the sign is, as expected, negative, it is not significant), despite the shortest travel time is always ensured with free riding. Those with competitive motivations are less likely than others to join a Collective good scheme, which is intuitive as the scheme is implicitly equitable in the sense of creating a more uniform distribution of travel times across participants by asking everyone to take turns and 'take one for the team'.

	Sacrifice-based		Collective good	
	Est. (SE)	Std.dev (SE)	Est. (SE)	Std.dev (SE)
ASC_{SR}	7.32***	6.27***	5.23***	5.10***
	(0.60)	(0.50)	(0.61)	(0.44)
Number of days	-0.95***	1.01***	-0.79***	0.97***
	(0.09)	(0.11)	(0.09)	(011)
Additional travel time	-1.12***	0.87***	-0.83***	0.90***
	(0.08)	(0.08)	(0.07)	(0.08)
Network travel time benefit	0.03 (0.04)	0.49***	0.03 (0.06)	0.31***
		(0.05)		(0.04)
Participation rate	0.21***	0.34***	0.24***	0.37***
	(0.03)	(0.03)	(0.03)	(0.03)
ASC_{SR} × Individualism	-2.65		-1.14 (0.85)	
	(0.85)***			
ASC_{SR} × Altruism	2.55		2.34 (2.11)	
	(1.20)**			
ASC_{SR} × Competition	-1.36 (1.49)		-3.68***	
			(1.30)	
ASC_{SR} × Common good	4.46***		4.75***	
	(0.84)		(0.82)	
ASC_{SR} × Fairness	0.87 (1.47)		4.97**	
			(1.96)	
McFadden ρ^2	0.51		0.52	
Final-LL	-2,133.16		-2,092.97	
Estimated parameters (k)	15		15	
Number of choices	6320		6256	

Table 4: *Mixed logit models estimated separately for the two schemes, using contextual moral motivation interactions with the alternative specific constants, or in other words, the predisposition to join the social routing system (ASC_{SR}). See Appendix for the utility specification. *, ** & ***, respectively represent significance at 10%, 5% & 1% levels.*

4.5. CONCLUSIONS AND DISCUSSION

Social routing schemes are touted as having the potential to reduce congestion while enjoying a relatively high level of public acceptance. These considerations have motivated a growing literature describing research efforts aimed at understanding travel behavior in the context of such schemes. The ultimate goal of such studies is to identify the most promising schemes in terms of their acceptance by travelers and their subsequent effects on network wide travel times. While the scientific community is nowhere near finding complete and reliable answers, much progress has been made. This paper contributes to this endeavour by focusing on an aspect of social routing schemes that hitherto has been underexplored: moral personality and moral motivations. Specifically, we looked into the interaction between the characteristics and framing of the scheme on the one hand, and travelers' moral personality and moral motivations on the other hand.

Using conceptual expositions and stated intention experiments, we shed light on these behavioral interactions: we hypothesize and empirically confirm that when the scheme is framed and designed as an altruistic effort (requesting personal sacrifices for the benefit of other travelers), mostly people who adhere to care related notions of morality are attracted to such a scheme. Contrary, a scheme that is designed and framed as a collective endeavour which would also benefit participating travelers (relative to the situation without a social routing scheme) attracts those who strongly adhere to moral notions related to fairness. These associations were found both at the level of generic personal morality as well as at the level of more targeted (to the specific context) moral motivations, implying robustness of these results. Interestingly, while moral personality and moral motivations turned out to significantly interact with the framing of the social routing schemes, the overall popularity of the schemes was about equal under the two frames – the Collective good frame only inducing slightly higher levels of stated intention to join the scheme.

We believe that the results presented in this paper have a relevance, albeit tentatively, for practitioners and policy makers. The main reason for being cautious here, is that our empirical results are obtained using a stated intention survey. Although there is growing evidence of the reliability and external validity of properly designed stated choice experiments, especially when they mimic situations that participants can easily relate to (Haghani et al., 2021; Rossetti and Hurtubia, 2020), we wish to note here that real life pilots are needed to further study the role of morality in the acceptance and effectiveness of social routing schemes. One differential outcome that we would expect from real life pilots is a larger difference between the overall willingness to join the sacrificed-based scheme and the collective good scheme. The particular downside of the former scheme of not benefiting from one's contribution over recurrent trips, as stipulated in section 4.2, is expected to be a stronger driving force in a real life setting compared to stated intentions. Nonetheless, what our findings do suggest is that morality plays a role in travelers' acceptance (willingness to join) social routing schemes and hence plays a role in defining their network level effects. Moreover, we find that different types of schemes appeal

to different types of road users, in ways that align with intuition and literature on moral decision making. Where a sacrifice-based scheme taps into notions of care, a collective good scheme taps into notions of fairness, broadly speaking. This leaves road authorities with a consequential decision to make: what type of social routing scheme, if any, should they implement?

The following considerations are relevant here: first, it should be noted that a sacrifice-based scheme is easier to implement than a collective good scheme, simply because the former does not need to live up to the promise of generating travel time gains – relative to the situation without a scheme – to participants. On the contrary, in order to be perceived as credible, a collective good scheme would need to ensure that most or all participants would benefit from joining the scheme even when occasionally being diverted to a slower route or less convenient departure time. This is not an easy task for a traffic authority, as it demands very careful forecasting and optimization; it is unclear whether the current state of technology would allow for such a tailored distribution of travel time benefits, although promising steps are made towards ever more sophisticated social routing schemes (Chen et al., 2021). It goes without saying that a scheme which claims that it also benefits participants, but in reality fails to live up to that promise, is doomed. The sacrifice-based scheme is much easier to implement as it makes no such promises.

The second consideration relates to the difficulty for sacrifice-based schemes in maintaining the loyalty of participants: our experiment affirms what other studies have found, in that willingness to join a scheme is determined by the size of the sacrifice – in terms of how often one is asked to choose the social alternative as well as in terms of the travel time difference with the usual alternative – and by the number of participants. This could easily set in motion a vicious circle of reduced willingness to participate: once a participating traveller becomes tired of making sacrifices for their fellow road users, they may be tempted to drop out. Once other participating travelers notice, they will become less likely to remain in the scheme (as the number of participants positively affects one's willingness to join). Furthermore, every traveler that leaves the scheme would trigger an increase in the sacrifices that need to be made by other participants to obtain the same system optimum, further eroding the willingness of the remaining travelers to continue their participation. Such a race to the bottom could easily and quickly lead to a depletion of altruistic motivations even amongst those with strong adherence to notions of care. No one wants to be the sole altruistic agent surrounded by a group of free riders. Such a destructive tipping point dynamic is less likely to occur when the scheme is set up as a collective routing scheme: in such a scheme, there is a much more limited incentive to drop out, as there are also personal benefits associated with participating. Furthermore, it seems reasonable to expect that free riding in the context of a fair social routing scheme has a higher chance of being frowned upon by others compared to just not being altruistic in a sacrifice-based social routing scheme, as stipulated in section 4.2. As such, we believe that social norms and peer pressure are more likely to sustain a collective good scheme than a sacrifice-based one.

As a third consideration, the distribution of costs and benefits across different 'moral types' should be considered by transport authorities. Our results suggest that a distinct group of care-oriented travelers would carry the largest burden of travel time sacrifices under a sacrifice-based scheme, whereas individualistic travelers would reap the benefits. Irrespective of how behaviorally sustainable such a distribution is in practice (see discussion above), the question should be asked whether society and its policy makers should actually be willing to accept such a situation. In contrast, the collective good scheme by design does not create a burden for any particular 'moral type': although participants to the scheme, driven mostly by fairness considerations, will benefit less than free riders, the former too will reap some benefits compared to the situation without a social routing scheme. As such, from a distributional justice perspective, a collective good frame may be preferred over a sacrifice-based scheme.

However, our results suggest that there is no silver bullet in the form of a social routing scheme that would be viable in the long run, in terms of travelers' willingness to participate, while at the same time being fair and easy to implement. This however, is a tentative conclusion drawn at this particular moment: as technology progresses and our understanding of traveler behavior – including moral aspects – advances, (partial) solutions to this conundrum may be found, building on the large and growing body of literature on the topic of social routing to which this paper contributes.

APPENDIX

4.A. FORMAL SPECIFICATIONS OF SYSTEMATIC UTILITIES

The systematic utility specification for the binary logit model of Table 1:

$$V_{SR,nt} = \sum_{k=0}^{K} (\beta_k + \delta_k F) x_{knt} \tag{4.1}$$

Where the k represents the four basic attributes and the average inclination to join the scheme and β_k their corresponding weights in the Sacrifice-based scheme. δ_k represents the difference between the estimate of parameter k for the Collective good scheme ($F = 1$) and the Sacrifice-based scheme ($F = 0$). Using this utility specification, the probability that individual n chooses to join the social routing scheme in choice occasion t is given by a binary logit model displayed in Table 1. The same systematic utility is used in Table 2, except $V_{SR,nt}$, β_k, and δ_k are not considered to be a single value, but a distribution instead.

The systematic utility specification for the morality interactions (Table 3 and 4) are also in a linear additive form:

$$V_{SR,nt} = \sum_{k=0}^{K} \beta_k x_{knt} + \sum_{m=0}^{M} \beta_m D_{mn} \tag{4.2}$$

Where m represents the moral foundations/motivations and β_m their corresponding weights. D_{mn} is a dummy variable which takes the value of 1 if individual n strongly adheres to foundation/motivation m. The weights of the moral foundations/motivations

are non-random, while the weights of the four basic attributes are normally distributed random variables. The weight for the average inclination to join the scheme is non-random for the moral personality (or MFQ scores) model (Table 3), and normally distributed random for the contextual moral motivation model (Table 4). This is due to stability reasons. These two specifications are estimated separately on the two schemes.

REFERENCES

Albalate, D., & Fageda, X. (2019). Congestion, road safety, and the effectiveness of public policies in urban areas. *Sustain. Sci. Pract. Policy*, *11*(18), 5092.

Chatterjee, K., & Mcdonald, M. (2004). Effectiveness of using variable message signs to disseminate dynamic traffic information: Evidence from field trails in european cities. *Transp. Rev.*, *24*(5), 559–585.

Chen, R., Leclercq, L., & Ameli, M. (2021). Unravelling system optimums by trajectory data analysis and machine learning. *Transp. Res. Part C: Emerg. Technol.*, *130*, 103318.

Çolak, S., Lima, A., & González, M. C. (2016). Understanding congested travel in urban areas. *Nat. Commun.*, *7*, 10793.

Cullity, G. (1995). Moral free riding. *Philos. Public Aff.*, *24*(1), 3–34.

de Palma, A., Proost, S., Seshadri, R., et al. (2018). Congestion tolling-dollars versus tokens: A comparative analysis. *Research Part B ...*

Djavadian, S., Hoogendoorn, R. G., Van Arerm, B., & Chow, J. Y. J. (2014). Empirical evaluation of drivers' route choice behavioral responses to social navigation. *Transp. Res. Rec.*, *2423*(1), 52–60.

Eikenbroek, O. A. L., Still, G. J., & van Berkum, E. C. (2021). Improving the performance of a traffic system by fair rerouting of travelers. *Eur. J. Oper. Res.*

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.

Gu, Z., Liu, Z., Cheng, Q., & Saberi, M. (2018). Congestion pricing practices and public acceptance: A review of evidence. *Case Studies on Transport Policy*, *6*(1), 94–101.

Haghani, M., Bliemer, M. C. J., Rose, J. M., Oppewal, H., & Lancsar, E. (2021). Hypothetical bias in stated choice experiments: Part i. integrative synthesis of empirical evidence and conceptualisation of external validity.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Hess, S., Train, K. E., & Polak, J. W. (2006). On the use of a modified latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice. *Trans. Res. Part B: Methodol.*, *40*(2), 147–163.

Klein, I., & Ben-Elia, E. (2018). Emergence of cooperative route-choice: A model and experiment of compliance with system-optimal ATIS. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 348–364. https://doi.org/10.1016/j.trf.2018.09.007

Klein, I., Levy, N., & Ben-Elia, E. (2018). An agent-based model of the emergence of cooperation and a fair and stable system optimum using ATIS on a simple road network. *Transp. Res. Part C: Emerg. Technol.*, *86*, 183–201.

Knockaert, J., Tseng, Y.-Y., Verhoef, E. T., & Rouwendal, J. (2012). The spitsmijden experiment: A reward to battle congestion. *Transp. Policy*, *24*, 260–272.

Koller, F. (2021). What determines the acceptance of socially optimal traffic coordination?: A scenario-based examination in germany. *Transp. Res. Part A: Policy Pract.*, *149*, 62–75.

Krabbenborg, L., Mouter, N., Molin, E., Annema, J. A., & van Wee, B. (2020). Exploring public perceptions of tradable credits for congestion management in urban areas. *Cities*, *107*, 102877.

Kroesen, M., & Chorus, C. G. (2018). The role of general and specific attitudes in predicting travel behavior–a fatal dilemma? *Travel behaviour and society*, *10*, 33–41.

Kröller, A., Hüffner, F., Kosma, Ł., Kröller, K., & Zeni, M. (2021). Driver expectations toward strategic routing. *Transp. Res. Rec.*, 036119812110064.

Liu, W., Li, X., Zhang, F., & Yang, H. (2017). Interactive travel choices and traffic forecast in a doubly dynamical system with user inertia and information provision. *Transp. Res. Part C: Emerg. Technol.*, *85*, 711–731.

Mariotte, G., Leclercq, L., Gonzalez Ramirez, H., Krug, J., & Bécarie, C. (2021). Assessing traveler compliance with the social optimum: A stated preference study. *Travel Behaviour and Society*, *23*, 177–191.

Noordegraaf, D. V., Annema, J. A., & van Wee, B. (2014). Policy implementation lessons from six road pricing cases. *Transp. Res. Part A: Policy Pract.*

Rossetti, T., & Hurtubia, R. (2020). An assessment of the ecological validity of immersive videos in stated preference surveys. *Journal of Choice Modelling*, *34*, 100198.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

van den Berg, T. G. C., Kroesen, M., & Chorus, C. G. (2020). Does morality predict aggressive driving? a conceptual analysis and exploratory empirical investigation. *Transp. Res. Part F Traffic Psychol. Behav.*, *74*, 259–271.

Van Essen, M., Eikenbroek, O., Thomas, T., & Van Berkum, E. (2019). Travelers' compliance with social routing advice: Impacts on road network performance and equity. *IEEE transactions on intelligent transportation systems*.

van Essen, M., Thomas, T., van Berkum, E., & Chorus, C. G. (2020). Travelers' compliance with social routing advice: Evidence from SP and RP experiments. *Transportation*, *47*(3), 1047–1070. https://doi.org/10.1007/s11116-018-9934-z

Van Essen, M., Thomas, T., van Berkum, E., & Chorus, C. G. (2016). From user equilibrium to system optimum: A literature review on the role of travel information, bounded rationality and non-selfish behaviour at the network and individual levels. *Transport reviews*, *36*(4), 527–548.

Viđak, M., Buljan, I., Tokalić, R., Lunić, A., Hren, D., & Marušić, A. (2020). Perception of organizational ethical climate by university staff and students in medicine and humanities: A cross sectional study. *Sci. Eng. Ethics*, *26*(6), 3437–3454.

5

MORAL IMAGES IN DISCRETE CHOICE MODELS: A NATURAL LANGUAGE PROCESSING APPROACH

The second study of Part II in this thesis uses language to enrich Discrete Choice Models in order to identify moral behavioural constructs. It does so using a novel approach. Natural Language Processing and free text data proved helpful in morality research. Therefore, connecting them to Discrete Choice Modelling to uncover moral motivations in decision-making can be a promising research avenue.
This study extracts moral dimensions from the decision-maker's verbal expressions and uses them as input in modelling moral decision making. The methodology is tested in a case study on voting in the European Parliament. Section 5.1 introduces the motivation and relevance of this research, section 5.2 presents the related literature. Section 5.3 introduces the general methodology. Section 5.4 contains the case study: the research background on political voting behaviour and the operationalization of the general methodology tailored to this context. Section 5.5 shows the modelling results and their interpretation. Section 5.6 concludes with further potential application fields, relevance for choice modelling and future research regarding the general methodology.

This chapter is based on the paper entitled 'Moral images in Discrete Choice Models: a Natural Language Processing approach' by Teodóra Szép, Sander van Cranenburgh, and Caspar Chorus submitted/currently under revision.

5.1. INTRODUCTION

Choice data is often used to infer people's underlying preferences about different products, policies or several other subjects. The field of discrete choice modelling focuses

on the mathematically rigorous analysis of decision making. Using data on observed choices, the analyst can derive people's preferences about different attributes in a choice task, such as price or quality. Most decisions in life potentially have a moral dimension, such as consumers considering worker conditions, fair trade, animal welfare or local community when making a purchase, doctors making trade-offs between health outcome and patient experience, or commuters considering how their travel practices affect other commuters or the environment. Morality can be defined as a set of principles that tells whether an action or state of the world is right or wrong.

Therefore, besides the traditional attributes, personality and moral values in particular also often play a significant role in many situations. Moral 'attributes' are substantially different from non-moral ones. Emotions, intuitions, and decision heuristics play a major role when contemplating trade-offs between them (Gigerenzer, 2010; Haidt, 2001; Sunstein, 2005). These processes are latent not only for the analyst but, in most cases, for the decision-makers too. Recent work regarding a broad range of latent variables, including latent moral motivations, shows that the joint identification of underlying preferences and other latent determinants of decision-making is a very challenging task (e.g. Vij and Walker, 2016).

Although progress is being made to advance the identification of such models based on choice data, one obvious potential solution has not received the attention it deserves: the use of additional text data to help identify latent behavioural constructs. One central argument for using text data in choice analysis is that the nuances that are present in free text often cannot be grasped with standard, closed-ended responses (Baburajan et al., 2020)[1]. This is even more relevant when the subjects are abstract and complex phenomena, such as moral values (Boyd et al., 2015). For two main reasons, free text data and language modelling show great promise for understanding how moral values relate to behaviour and choices. First, in the age of the internet and social media, a vast amount of text is generated every day, carrying plenty of information potentially useful for understanding morality and complex decision-making phenomena. Second, language models in the rapidly growing Natural Language Processing (NLP) field approach the human level of text understanding and allow us to quantify qualitative text data in several ways to understand moral values and behaviour better.

This paper proposes a method to combine choice- and text data to infer moral motivations in decision-making contexts. We show how this novel approach can lead to new, subtle insights regarding latent antecedents of moral choice, which would be very difficult – if not impossible – to obtain using traditional choice models based on observed choices only. To test and illustrate our proposed approach, we investigate the voting behaviour of Members of the European Parliament (MEPs). The rest of the paper is organized as follows: Section 5.2 provides a brief literature review related to our methodological approach. Section 5.3 describes our general methodology for creating moral

[1] In their recent work, Baburajan et al. (2022) find that although Topic Modeling is suitable to extract information from open-ended responses, discrete choice models estimated using closed-ended questions perform better than those using the open-ended questions.

images of texts, which can be used in various research contexts. Section 5.4 describes the context of our case study, our data, and the operationalization of the methodology. Section 5.5 shows the results and discusses their interpretation. Section 5.6 discusses the limitations of the methodology and future research avenues.

5.2. RELATED WORK

Discrete choice models (DCMs) relying on full-fledged Natural Language Processing (NLP) methods to make use of additional text data are not yet used in the literature (van Cranenburgh et al., 2021). A few papers indicate that both NLP methods and additional text data can capture subtleties that were overlooked in the literature before. A recent paper by Pereira (2019), for instance, shows how NLP methods can encode subtle yet important nuances in travel behaviour modelling using DCMs. For example, "student" and "employed" categorical characteristics are rather similar when it comes to departure time choice but dissimilar when it comes to car ownership. In the traditional variable encoding, these are one unit distance from each other. However, word embeddings (i.e. words represented with vectors of real numbers) allow us to encode this subtle difference in choice models. Pereira (2019) does not rely on additional text data but uses the words that are already part of most traditional data (i.e. attributes and personal characteristics). Studies that used free text data[2] in DCMs include Glerum et al. (2014) who used semi-open questions about different transport modes to include perceptions, and Baburajan et al. (2020) who used open-ended questions to measure attitudes towards shared mobility services.

Studying morality through natural language has been vast and growing in the past decades as increasing computing power allows for higher quality and quantity of text mining and NLP techniques. Most studies in this field rely on the Moral Foundations Theory (MFT, Graham et al., 2009). MFT originates from moral psychology and postulates that people have five innate moral foundations: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation. According to the theory, these foundations are cross-cultural; they can be found in everyone, only their extent differs across people. Measuring this extent has two main methods. First is a closed-ended questionnaire, the Moral Foundations Questionnaire (MFQ), which asks respondents to what extent different things (e.g. whether or not someone suffers emotionally) affect their moral judgement in a situation. Second is the Moral Foundations Dictionary (MFD), which contains words related to each foundation and direction (i.e. a word can belong to either virtue or vice in the same moral foundation). The first version of MFD was extended several times (Araque et al., 2020; Frimer et al., 2019; Hopp et al., 2021). Operationalization of MFD ranges from word counting methods to sophisticated NLP algorithms. There are two main tools to extract moral foundations from text: MFD (or one of its extended versions, e.g. Kaur and Sasahara, 2016; Mutlu and Tütüncüler, 2020; van den Broek-Altenburg et al., 2020) and manual (expert) annotation (e.g. Hoover et al., 2020). Furthermore, complex NLP models can be trained using MFD, annotated data, or both, to classify a piece of text into one of the moral foundations (Araque et al., 2020;

[2] By free text, we refer to a piece of text that is not the result of a closed-ended question. Free text can be either the response to an open-ended question or something that a person expresses on their own initiative, for example, social media posts.

Hoover et al., 2020). When it comes to interpretation, MFQ is straightforward; scoring high on a moral foundation means a higher emphasis on the given foundation when making a moral judgement. This is not necessarily true for language use. In a political context, interestingly, it was found that although conservatives adhere to loyalty more than liberals, loyalty appears more in liberals' moral rhetoric graham2009liberals. Although this effect was small, Frimer (2020) found it to be robust. This means that moral rhetoric does not necessarily represent the intrinsic values of a person, and one must be careful with interpreting outcomes.

5.3. METHODOLOGY

In this paper, we propose a method of using moral images as inputs in Discrete Choice Models. Moral images are the quantified moral dimensions of text data, as a text created by humans has the purpose of projecting an image. This could be honest because one might want to project their actual values. However, it is also possible that one purposefully talks or writes in a specific way to be perceived as endorsing different values. Therefore, moral images do not necessarily reflect the 'true' values of the text's creator, but they do reflect the values the piece of text projects.

In order to quantify morality in text data, we need 1, moral text data and 2, an NLP method called feature vector representation. Moral text data is data on different dimensions of morality, such as care, fairness or loyalty. The moral dimensions could be based on Moral Foundations Theory (Graham et al., 2009), Schwartz Values (Schwartz, 1992), or Morality-as-Cooperation (Curry et al., 2019), to name a few. Moral Foundations Theory has a large body of literature relating it to text analysis and has a dictionary that was updated several times; thus, without claiming that other definitions of morality are incorrect or less useful, we use the moral domains of MFT in this paper. Feature vector representation means that all words in a text are represented with a vector of real numbers. This can be done in several ways, from more simple such as bag-of-words method[3] to state-of-the-art Transformers methods (Vaswani et al., 2017). In order to create a moral image for any piece of text, first, we create feature vectors for all moral domains based on the moral text data. Then we do the same for the piece of text at hand and measure the similarity between the text's and each moral domain's vector. To see how similar a text is to each moral domain, we compute the cosine similarity[4]between their feature vectors. This way, a piece of text's moral image determines how similar the text is to each of the moral domains. Note, that similarity is not equal to endorsing a particular value. For example, the sentences *"Compassion for those who are suffering is the most crucial virtue"* and *"Compassion for those who are suffering is not a crucial virtue"* both score highest on care virtue, despite they mean the opposite in terms of endorsement[5]. This is

[3]The bag-of-words representation carries information on the words and their number of occurrence in a piece of text. It disregards the grammar and word order.

[4]Cosine similarity is a measure of similarity between two sequences of numbers viewed as vectors. For two vectors, A, B and the angle between them θ, the cosine similarity is calculated with the following formula: $\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$, where A_i is the ith element of vector A.

[5]Interestingly, in some cases the model does reflect endorsement differences, for example *"I am proud of my*

important when building the models in section 5.4 and interpreting the results in section 5.5. The similarity score can range from -1 to 1. 1 means perfect similarity, 0 means no relation, and -1 means a perfect opposite relation between two vectors. See Figure 1 for an illustration and the detailed description below.

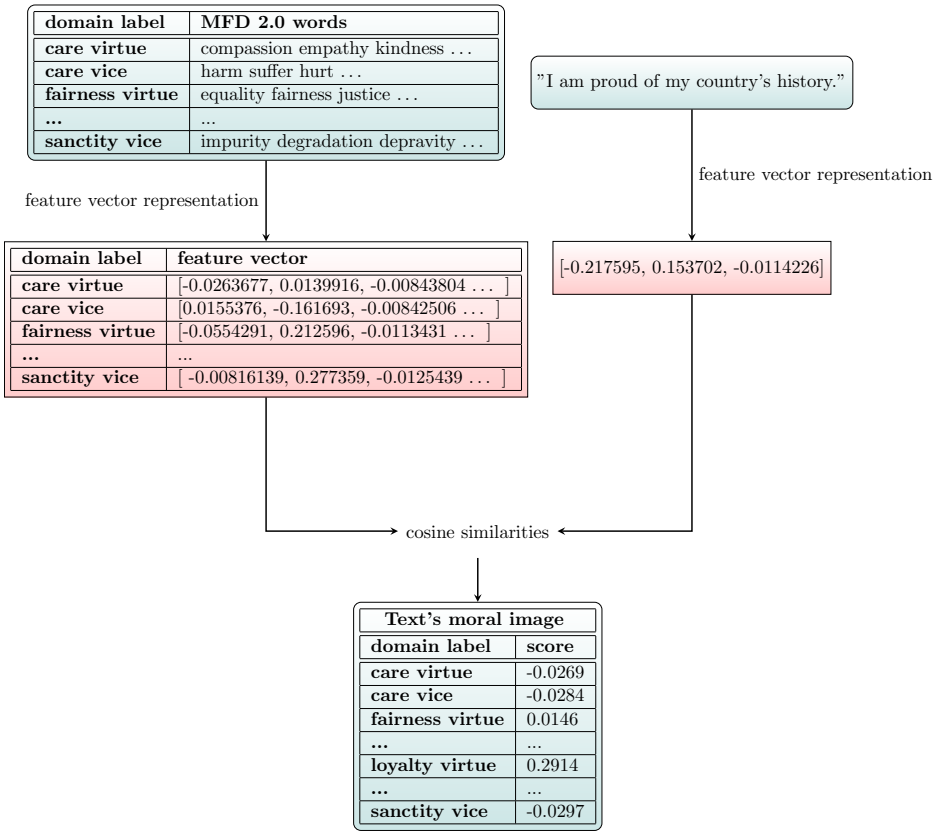


Figure 1: *The process of creating moral images for a piece of text. Inputs and output are coloured in blue, and the intermediate steps of the process are coloured in red. The calculation methods are on the corresponding arrows. The example sentence is from the Moral Foundations Questionnaire (MFQ), and we find that the domain of "loyalty virtue" has the highest score. The sentence corresponds to the loyalty foundation according to the creators of MFQ too.*

In order to utilize behavioural data (i.e. choice data), we use the Discrete Choice Model family. According to DCMs, the probability of individual n choosing alternative i can be generally expressed as:

$$P_{ni} = Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall j \neq i) \tag{5.1}$$

country's history" scores highest on loyalty virtue, while the sentence *"I am not proud of my country's history"* scores highest on loyalty vice.

where V_{ni} is the observed part of the latent continuous variable representing the motivation of decision maker n to choose alternative i. ε_{ni} is the random error term, or the unobserved part of the latent motivation.

V_{ni} can be generally characterized as follows.

$$V_{ni} = f(X_{ni}, S^m_{ni}) \tag{5.2}$$

where X_{ni} are the attributes of alternative i for individual n, depending on the choice situation, and S^m_{ni} are the scores of the moral domains (i.e. the output of the NLP model). The specification of $f(X_{ni}, S^m_{ni})$ depends on the choice task at hand. For instance, one may want to include the moral images of the decision-makers or the moral images of different product descriptions, or both at the same time.

5.4. CASE STUDY: VOTING IN THE EUROPEAN PARLIAMENT

To empirically test and illustrate how moral images can lead to subtle behavioural insights, we use a case study in the field of politics, namely voting in the European Parliament. This section describes our case study, the operationalization of the above methodology, and the results.

One of the most critical decisions is arguably political decision making: elected representatives decide about policies that potentially have a significant effect on many people's lives. These decisions also often have a moral component: protecting fundamental human rights in foreign countries, helping the poor, investing for the sake of future generations or preserving the environment. In our case study, we examine voting behaviour in the European Parliament (EP) on subjects, such as "Search and rescue in the Mediterranean" or "The impact of Covid-19 on youth and on sport". In the EP, there are 705 members (MEPs), whom the citizens of the European Union elect in their home countries. Most MEPs belong to one national party in their home country, and these national parties usually join in 7 EP party groups. There are independent representatives too. Although the EP started as a consultative body, it gained power, and by now, it can enact legislation, amend the budget, or censure the Commission. Most voting procedures are not recorded, but roll-call votes are required on final legislation votes and whenever a political group or 30 MEPs request it. In roll-call voting, the vote of each member is recorded.

5.4.1. RESEARCH BACKGROUND OF POLITICAL VOTING BEHAVIOR

MFT has a history of explaining moral value differences across people, and a large amount of literature focuses on political, ideological differences. In the American political system, which is primarily dominated by two main ideologies, liberalism and conservatism, it has been observed that there is a systematic difference between the two groups in terms of moral foundations. According to the initial studies into the subject, it was found that liberals score higher on the so-called *individualizing* foundations, namely care and fairness, while conservatives score lower on these two and higher on the *binding* foun-

dations; loyalty, authority, and sanctity (Graham et al., 2009; Haidt and Graham, 2007). This general hypothesis was corroborated by context-dependent studies, such as political text on stem-cell research (Clifford and Jerit, 2013) or abortion (Sagi and Dehghani, 2014), but also refuted in environmental contexts (Frimer et al., 2015) where liberals used language heavier in sanctity.

In the past few decades, the voting behavior of MEPs has been the subject of several political studies. Hix (2002) hypothesized that MEPs are driven by three main factors: personal preferences, national party discipline and EP party group discipline. The three are often correlated; people with similar beliefs join together in national parties, then national parties with similar agendas join in the EP as party groups. However, there are exceptions in some cases; national parties and EP party groups might disagree, individuals might defect one or both of their parties. These occasions allow for studying which motivations are more important in different situations. Hix (2002) found that the main driving force behind MEPs' voting behavior is their national party position; measured with distances between MEPs' EP party group and national party, based on the left-right location and EU-integration location. These were calculated based on a questionnaire where MEPs placed themselves and their parties on this political spectrum. MEP's individual distance from their EP party group was not significant. The high impact of national party discipline was supported by several studies and extended with additional insights on its reasons (Faas, 2002, 2003; Hix, 2004; Klüver and Spoon, 2015; Lindstädt et al., 2011).

Text data from the documents under votes proved to be valuable assets in several papers in the literature concerned with modelling roll-call vote outcomes, although they did not examine the context of the European Parliament. The goal of these studies is either better prediction (Gerrish and Blei, 2010; Korn and Newman, 2020; Kraft et al., 2016) or understanding preferences (Kim et al., 2018; Lauderdale and Clark, 2014). These latter studies estimate the underlying number of latent dimensions rather than imposing it a priori. This way, they provide insights into how different topics (characterized by sets of words) affect voting behaviour.

5.4.2. OPERATIONALIZATION OF MORAL IMAGE METHODOLOGY

Our goal is to test how moral images in discrete choice models can give nuanced behavioural insights in the context of MEPs voting behaviour. The above literature gives valuable insights based on political science. We test whether similar conclusions can be drawn from a different approach in the field of discrete choice modeling. For moral text data, we use the MFD 2.0 lexicon (Frimer et al., 2019). In a comparison study among the extended versions of MFD, MFD 2.0 was found to be the best in terms of similarity between human-annotated texts and dictionary labels (Mutlu and Tütüncüler, 2020). MFD 2.0 is a dictionary of the five moral foundations with corresponding 'virtue' and 'vice' words in English, thus resulting in ten moral domains in total. In order to collect choice data (i.e. roll-call voting data), we use the website of the European Parliament. To collect text data from MEPs on their political views, we use their Twitter accounts, which are used as communication channels for political purposes. We collected 328 MEPs' latest tweets (up to 100) in 2021 April. This data includes short text pieces (up to 140 char-

acters) in 26 different languages. From the European Parliament website, we collected document text data on 24 different voting subjects, such as "Reducing inequalities with a special focus on in-work poverty" or "The EU Strategy for Gender Equality" (see Appendix 5.A for the complete list). Besides the text data, we also collected choice data (i.e. the roll-call votes'), containing whether each MEP voted 'in favour, 'against', or 'abstain'.

We operationalize our proposed methodology (section 5.3) the following way. We use MFD 2.0 for moral text, thus we create moral images based on 10 domains. To create feature vector representations, we use a Transformer-based SBERT[6] (Reimers and Gurevych, 2019) model. SBERT is a cutting edge NLP method that allows the words to have a spot in a so-called semantic space, where for instance, "cat" is closer to "dog" than to "car". Furthermore, SBERT is able to understand the context of words, meaning that the word "right" has a different vector when the context is human rights and when it is right-wing politics. Its practical advantages are multilingual ability and high speed. We tested this method by creating images for the sentences of MFQ[7]. We found that in 27 moral images out of 30, the highest score belonged to the actual foundation the sentence represented (see, for example, a loyalty sentence in figure 1). For our case study, we create moral images for all tweets, which are then averaged by MEPs to get their individual moral images. Then MEPs individual moral images are averaged within parties (both national parties and EP party groups) to get party-specific moral images. We also create moral images for the documents under vote. We use roll-call votes and party defection (casting a different vote than the party majority) as choice data. After a descriptive analysis, we first model EP party defection based on moral image scores and distances, then we model voting outcome based on document text.

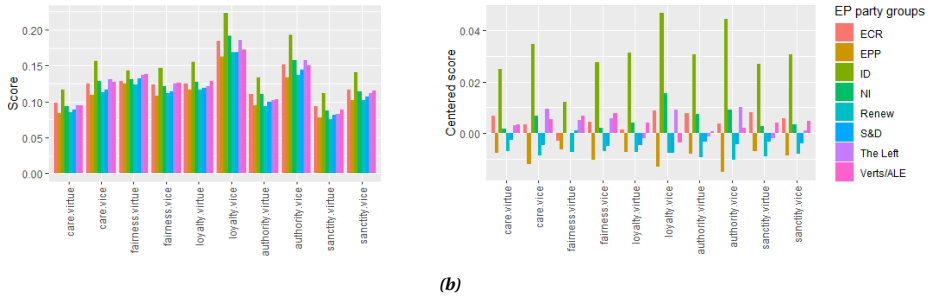
DESCRIPTIVE ANALYSIS

It was found in several studies that conservatives and liberals endorse different moral foundations to a different extent. Liberals put more emphasis on individualizing foundations, while conservatives to binding foundations. Similar systematic moral language use (which does not necessarily reflect one's endorsement of values) differences were also observed (Clifford and Jerit, 2013), but in some cases such differences were refuted (Frimer et al., 2015). The American liberal-conservative division in European context is often substituted with left-right division, however in a non-bipolar system partisan differences cannot always be explained by this distinction (Patkós, 2022). For instance, Kivikangas et al. (2017) empirically found that in the Finnish political landscape "liberalism-conservatism" cannot interchangeably used with "left-right" in terms of political spectrum division. In a language use examination Proksch and Slapin (2010) also found that EP debate speeches poorly reflect partisan divisions over left–right politics.

In our case study, we first examine whether moral image differences can be found in the European political spectrum by plotting the scores of EP party groups based on their members' tweets. To do this, we create moral images for EP party groups by averaging their members' moral images. The member's moral images are the average moral images

[6]We use the model 'paraphrase-multilingual-mpnet-base-v2' (Reimers and Gurevych, 2020).
[7]We used MFQ30, which can be found at https://moralfoundations.org/questionnaires/.

of all their tweets. Figure 2 plots the scores of each party group on the ten moral domains.



(a) *(b)*

Figure 2: *Moral images of EP parties: Figure (a) shows the average scores of the EP party groups (and independents as "NI") on the 10 moral domains. Figure (b) shows the centered average scores.*

From Figure 2a we can see that the overall pattern is more or less the same for the EP party groups. We find no prominent differences in moral language use nor along the left-right or other political spectrum divisions. The moral image scores seem to be the same for all parties. This finding is aligned with the line of literature that refutes that the liberalism-conservatism division can be substituted with left-right in the European context (Kivikangas et al., 2017; Patkós, 2022). This can be the result of general political discourse: politicians' topics potentially have a general level of similarity to the MFD 2.0 lexicon. Interestingly, this general similarity is higher for the vice-domain in each foundation, except for fairness. It is also intuitive, as political discourse is often about problems that need solving, and parties might criticize or frame each other negatively (e.g. Turk, 2019). The one obvious outlier on figure 2a is the ID party group (positioned on the far right). It scores higher on all domains than the other party groups and also shows a somewhat different pattern; their fairness virtue score is relatively low. To see the more subtle differences between party groups, we also plot the centered moral images (figure 2b). Figure 2b shows that the lowest-scorers on each domain are Renew and EPP, two large parties in the center of the left-right spectrum. This can be interpreted as more radical parties, compared to ones in the center, tend to moralize more to build on people's (negative) emotions instead of their rational mind (e.g. Salmela and Von Scheve, 2017; Turk, 2019).

EP PARTY DEFECTION

It is established that MEPs most often vote in line with the majority of their EP party group (e.g. Hix, 2002; Klüver and Spoon, 2015; Lindstädt et al., 2011). The reason for this is twofold: people gravitate towards parties they agree with, and there is party discipline. Therefore in cases when MEPs defect their party group, we can assume they have a strong reason for it. We explore whether MEPs' and voting documents' moral images have explanatory power on defecting one's EP party group. To do so, we estimate binary logit models where the outcome variable is defecting the EP party group (or not). Binary logit means that equation 5.1 takes the form of

$$P_{defection,i} = \frac{\exp(V_{defection,i})}{1+\exp(V_{defection,i})} \tag{5.3}$$

where $V_{defection,i}$ is the latent continuous variable representing a MEP's motivation to defect, and based on equation 5.2 it is characterized using the following explanatory variables: alternative- and party group-specific constants (X_{ni} from equation 5.2), moral images and moral image distances (S_{ni}^{m} from equation 5.2).

In our defection analysis we explore two avenues: score-based models, and distance-based models. First we examine how moral scores of individuals, parties and documents under vote relate to party defection (see the models' explanatory variables in table 1).

Table 1: *Score-based models of defection and their included attributes. See corresponding estimated values in Table 3.*

	Model 1	**Model 2.A**	**Model 3.A**	**Model 4.A**
ASC and EP party group specific constants	✓	✓	✓	✓
Individual moral image scores		✓		✓
National party's moral image scores			✓	✓
Document's moral image scores				✓

Table 2: *Distance-based models of defection and their included attributes. See corresponding estimated values in Table 4.*

	Model 1	**Model 2.B**	**Model 3.B**	**Model 4.B**
ASC and EP party group specific constants	✓	✓	✓	✓
Individual distances from EP party group		✓		✓
National party's distance from EP party group			✓	✓
Documents' moral image distance from individuals				✓

For the score-based defection analysis, the fully specified model's latent motivation to defect is characterized as follows:

$$V_{defection,i} = ASC + \sum_{q \in Q} \beta_q x_{q,i} + \sum_{m \in M} \big(\beta_{m,ind} S_{i,m,ind} + \beta_{m,party} S_{i,m,party} + \\ + \beta_{m,doc} S_{m,doc} \tag{5.4}$$

where ASC is the alternative specific constant for defection, Q is a set of party groups, and $x_{q,i}$ is a binary variable, taking the value of 1 when individual i is in party group q, and 0, when they are not. M is the set of moral domains. $S_{i,m,ind}$ is the individual score of individual i on moral domain m. $S_{i,m,party}$ is the average score of the national party of individual i on domain m. Sm,doc is the score of the document under vote on domain m.

For the distance-based defection analysis, the fully specified model's latent motivation to defect is characterized as follows:

$$V_{defection,i} = ASC + \sum_{q \in Q} \beta_q x_{q,i} + \sum_{m \in M} \big(\beta_{m,ind} D_{i,m,ind} + \\ + \beta_{m,party} D_{i,m,party} + \beta_{m,doc} \cdot I_{against} \cdot D_{m,doc} \quad (5.5)$$

where $I_{against}$ is binary indicator of whether the EP party group of individual i prefers "against"[8]. $D_{i,m,ind}$ and $D_{i,m,party}$ are the dimensions of moral image distances. Moral image distances are calculated domain-wise; for example, for "care virtue", the individual distance of MEP i from their EP party group is:

$$D_{i,carevirtue,ind} = |S_{care.virtue,MEP_i} - S_{care.virtue,EPpartygroup_i}| \quad (5.6)$$

The distance between the national party and EP party group of MEP i is similarly:

$$D_{i,care.virtue,party} = |S_{care.virtue,nationalparty_i} - S_{care.virtue,EPpartygroup_i}| \quad (5.7)$$

And finally, the distance between the document and MEP i is similarly:

$$D_{i,care.virtue,party} = |S_{care.virtue,MEP_i} - S_{care.virtue,doc}| \quad (5.8)$$

where $S_{care.virtue}$ is the score corresponding to the "care virtue" domain in the given moral image. Similarly, $S_{m,doc}$ is the score of the document under vote on moral domain m.

The latent motivation for non-defection (i.e. voting with the party group) is normalized to be 0.

$$V_{non-defection,i} = 0 \quad (5.9)$$

For both avenues of defection analysis (moral image score- and distance-based) we first estimate a baseline model, where voting defection is modelled only with EP party groups as explanatory variables. Then we add individual-, party-specific- and document-specific scores and distances in three steps (see table 1 and 2 in Results). The key indicators we compare are model fit and the number of significant parameters.

VOTING OUTCOME

Moving away from party politics, we examine whether legislative texts' moral images have explanatory power in voting modeling. In this case, the outcome variable is "for", "against", or "abstain". The explanatory variables are alternative specific constants and moral image scores of the documents under vote. The modeling has two stages; first, we estimate a model with only alternative specific constants (this model serves as a benchmark), then add the moral images. We test whether there is a significant improvement in model fit and whether the a priori imposed moral domains are significant for explaining voting results. The model's systematic components (i.e. latent continuous variables representing MEPs' motivation to vote for, against or abstention) are specified as follows.

[8] The reason for this relates to interpretability. Individual distance from a document is not expected to relate to defection, however individual moral image distance is intuitively expected to relate to voting "against" a proposed document. See section 5.5.1 for more details on the interpretation of this model.

$$V_{for} = ASC_{for} + \sum_{m \in M} \beta_{for,m} S_{m,doc} \tag{5.10}$$

$$V_{against} = ASC_{against} + \sum_{m \in M} \beta_{against,m} S_{m,doc} \tag{5.11}$$

$$V_{abstain} = 0 \tag{5.12}$$

where M represents the ten moral domains, and $S_{m,doc}$ is the document's score on moral domain m. Abstention is normalized to be zero.

EXPECTATIONS

To formulate our expectations regarding our data analysis, we rely on three main findings from the literature. First, moral foundations can capture political and ideological differences. Second, ideological distances between MEPs' parties (national and EP) have more explanatory power than MEPs' individual distance from their EP party when it comes to defecting from their EP party. Third, text analysis of documents has explanatory power in modelling voting outcomes.

Expectation 1 We expect that model 3.B in table 4 has higher explanatory power than model 2.B (in table 4). Several studies consistently find that political distance between parties has a significant effect on voting defection. We test whether distance based on moral images has explanatory power when modelling voting defection from the EP party group. Defecting one's EP party group is naturally assumed to be related to the ideological distance one has from the rest of the party. Interestingly Hix (2002) found that instead of the individual distance, the national party's distance had more explanatory power when modelling defection. As ideological distance is potentially reflected in one's moral rhetoric, we expect that ideological distance measured with moral images based on natural text shows a similar pattern; the national parties' distance from the EP party group (Model 3.B in table 4) has more explanatory power than the individual distance (Model 2.B in table 4).

Expectation 2 We expect moral images to have significant explanatory power in modelling voting outcomes. Modelling voting outcomes based on document text has not been a subject of EP related literature. However, in cases of voting in the US congress and supreme court, text was a good predictor of outcome, and several topic-related preferences were uncovered. For text analysis, these studies used deep learning[9] methods or topic modelling approach. We test whether the moral images extracted from the text also have explanatory power. We model voting outcomes with moral dimensions (which correspond to MFT) and expect (some of) them to be significant (model based on equations 5.10-5.12).

[9]The goal of such deep learning models is accurate classification, and no interpretation or reasoning is provided in these black-box approaches.

5.5. RESULTS

In this section we present the results; addressing expectation 1 and 2 in subsections 5.5.1, 5.5.2 accordingly. Subsection 5.5.3 summarizes the behavioural findings and limitations of the case study and discusses possible directions for further investigations.

5.5.1. EP PARTY DEFECTION

First, we examine how moral scores of individuals, parties and documents under vote relate to party defection (see the models' explanatory variables in table 1). Then, in order to address expectation 1, we examine whether distances between national parties', individual MEPs', EP party groups' and documents moral language use have explanatory power when modelling EP party defection (see the models' explanatory variables in table 2).

Table 3 presents the score-based models of defection. For a baseline model, we estimate a binary logit model, where explanatory variables are the alternative specific constant (ASC) for defection and EP party groups (model 1 in table 3, based on equation 5.4, assuming $\beta_{m,ind} = \beta_{m,party} = \beta_{m,doc} = 0 \quad \forall m$). The ASC represents the average tendency to vote against one's party group in the benchmark party group. The benchmark party group is the EPP, the largest one in the EP, positioned in the centre-right. In our data, compared to EPP, ID, Renew, The Left and ECR are more likely to defect, while S&D and The Greens are less likely to defect, assuming everything else is constant. The most likely-to-defect party group is also the one that has the highest scores in their moral image (i.e. ID party group).

Table 3: *Score-based models of defection based on moral images (see subsection 5.4.2 for details): (1) is the baseline model, in (2.A) individual scores in (3.A) national party scores of moral images are the explanatory variables. Model (4.A) contains both of the previous moral images and moral images of the documents under vote. Moral image domains that relate to higher than average defection (significant on at least 5% and have positive sign) are highlighted in red, those that relate to higher than average party group cohesion (significant on at least 5% and have negative sign) are highlighted in blue. *,**, and *** represents significance on 5%, 1% and 0.1% accordingly.*

	(1)	(2.A)	(3.A)	(4.A)
ASC_{defect}	-2.3099***	-2.3077***	-2.30852***	-2.372***
ID	1.3806***	1.4460***	1.58171***	1.629***
S&D	-0.9308***	-0.9895***	-0.98902***	-1.043***
Renew	0.5299***	0.4210**	0.29554*	0.255
The Left	0.4765**	0.4420*	0.32553	0.334
ECR	1.0613***	1.1064***	1.19879***	1.200***
The Greens	-3.8642***	-3.9524***	-3.93563***	-4.021***
care.virtue		-5.5780		-4.671
care.vice		20.3101*		-16.285
fairness.virtue		-1.7170		-1.261
fairness.vice		14.2867*		8.614
loyalty.virtue		5.7068		0.385
loyalty.vice		-5.5858		-5.405
authority.virtue		-5.5437		1.710
authority.vice		20.7299**		13.459
sanctity.virtue		-2.7415		-3.984
sanctity.vice		-6.5576		1.412
$care.virtue_{party}$			-4.12785	-2.210
$care.vice_{party}$			-26.89903	-23.962
$fairness.virtue_{party}$			1.48901	0.409
$fairness.vice_{party}$			21.86468*	19.903
$loyalty.virtue_{party}$			13.33322*	12.231
$loyalty.vice_{party}$			-0.08253	-1.439
$authority.virtue_{party}$			-28.78930*	-25.967*
$authority.vice_{party}$			24.30376*	26.826*
$sanctity.virtue_{party}$			8.49223	7.575
$sanctity.vice_{party}$			-19.79860	-22.502
$care.virtue_{doc}$				8.441
$care.vice_{doc}$				21.258***
$fairness.virtue_{doc}$				13.921***
$fairness.vice_{doc}$				-9.866
$loyalty.virtue_{doc}$				-10.991***
$loyalty.vice_{doc}$				14.536***
$authority.virtue_{doc}$				-21.783***
$authority.vice_{doc}$				-28.148***
$sanctity.virtue_{doc}$				14.941*
$sanctity.vice_{doc}$				7.714
Log-Likelihood	-2038.1	-2027.5	-2026.0	-1974.5
LL(0)	-4346.0	-4346.0	-4346.0	-4346.0
BIC	4137.4	4203.7	4200.5	4272.5
Rho^2	0.53	0.53	0.53	0.55
Number of observations	6270	6270	6270	6270
Number of estimated parameters	7	17	17	37

Including individual moral image scores (model 2.B in table 3, based on equation 5.4,

assuming $\beta_{m,party} = \beta_{m,doc} = 0 \quad \forall m$) and national party average scores (model 3.A in table 3, based on equation 5.4, assuming $\beta_{m,ind} = \beta_{m,doc} = 0 \quad \forall m$) performs similarly in terms of model fit: they show moderate improvement in log-likelihood compared to the baseline model, and the BIC is higher. The significant moral parameters are mostly positive, except for authority.virtue$_{party}$. This means, that for example, someone who scores high on care vice or fairness vice, is more likely to vote against their party group. Those whose national party scores high on authority virtue, are more likely to vote with their party group.

Including both individual and national party scores, along with the documents' moral images (model 4.B in table 4, based on equation 5.4) results in a significantly better model fit in terms of log-likelihood, but in terms of BIC[10], it only outperforms model 2.B. Seven out of ten moral domains are significant from the documents' scores. A positive sign means that subjects that score high on a given domain are more likely to co-occur with a higher than average defection rate. A negative sign correspondingly means that subjects that score high on a given domain are more likely to co-occur with a higher than average level of party group cohesion. Having seven out of ten moral domains significant, we can say that a subject that is heavily loaded with morality (on almost any domain) will be more likely to result in either higher than average defection or, oppositely, cohesion rate. This is intuitive as a morally salient topic can be an incentive to stand up against party groups if one's own beliefs differ. However, critical moral questions are also likely to be where party groups strongly agree. Model 4.B in table 4 shows in red the moral domains more likely to be involved in intra-party-group controversy (i.e. care vice, fairness virtue, loyalty vice, sanctity virtue), and in blue that have the most consensus within parties (i.e. loyalty virtue, authority virtue, authority vice). Table 3 also shows that statistically significant individualizing foundations (i.e. care and fairness) consistently (as individual, national party or document scores) relate to higher than average defection. This is intuitive as those who value or express individualistic foundations verbally are less likely to be driven by group loyalty in moral questions. However, binding foundations (loyalty, authority and sanctity) give a mixed picture: they sometimes relate to cohesion, sometimes to defection. This can be the result of individual MEPs having two parties and their group loyalty being compromised when those do not agree.

[10] Note that BIC measures model fit while penalizing a high number of parameters. The models of table 3 and 4 are not targeted to find the best model for this particular case but to illustrate how moral image scores and their distances can be used in choice models. Thus, we did not merge scores and distances to find better BIC.

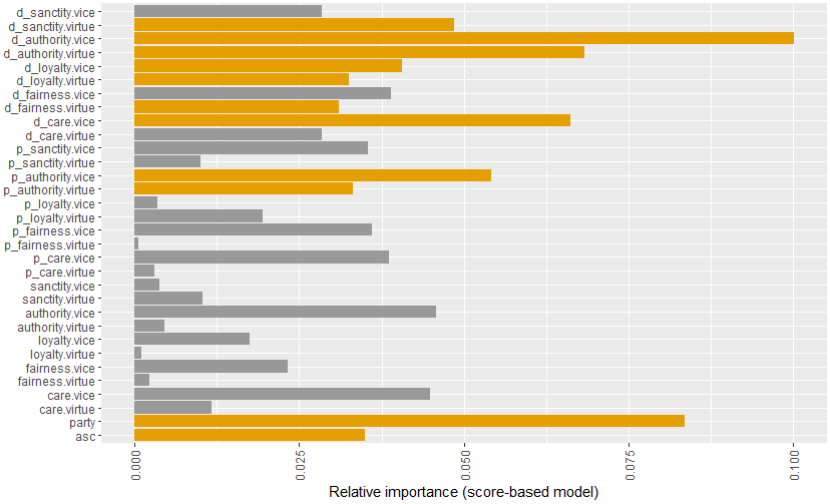


Figure 3: *Relative importance of moral images in the score-based full model (model 4.A of table 3). Significant parameters are signalled with yellow color.*

Figure 3 shows the relative importance of moral image dimensions when voting against one's EP party group. We can see that the significant parameters (from document scores and national parties' authority scores) account for approximately the same fraction of the latent motivation to defect as the alternative specific constant, or the EP party group affiliation.

EXPECTATION 1

To address expectation 1, we also model defection based on moral image distances. Results are presented in table 4.

Table 4: *Distance-based models of defection based on moral images (see subsection 5.4.2 for details): (1) is the baseline model, (2.B) the individual distance-based model, (3.B) the national party distance-based model and (4.B) both of the previous distances and individual MEPs' distance from document under vote are the explanatory variables. Moral image domains that relate to higher than average defection (significant on at least 5% and have positive sign) are highlighted in red, those that relate to higher than average party group cohesion (significant on at least 5% and have negative sign) are highlighted in blue. *,**, and *** represents significance on 5%, 1% and 0.1% accordingly.*

	(1)	(2.B)	(3.B)	(4.B)
ASC_{defect}	-2.310***	-2.269***	-2.483***	-2.45661***
ID	1.381***	1.387***	1.565***	1.36797***
S&D	-0.931***	-0.942***	-0.814***	-0.86512***
Renew	0.530***	0.594***	0.624***	0.71723***
The Left	0.477**	0.564***	0.761***	0.66248***
ECR	1.061***	1.020***	0.914***	0.94051***
The Greens	-3.864***	-3.833***	-3.756***	-3.79553***
care.virtue		-3.823		-5.57490
care.vice		6.367		4.88266
fairness.virtue		-13.466**		-16.42872***
fairness.vice		6.154		6.24089
loyalty.virtue		-3.212		2.66383
loyalty.vice		2.915		0.11359
authority.virtue		-4.026		-9.02763
authority.vice		3.974		5.97266
sanctity.virtue		14.005		6.55534
sanctity.vice		-17.419		-7.91254
$care.virtue_{party}$			-15.174	-13.32217
$care.vice_{party}$			4.333	14.14726
$fairness.virtue_{party}$			-27.529***	-29.14921***
$fairness.vice_{party}$			18.769*	18.74683*
$loyalty.virtue_{party}$			-20.372**	-23.00789**
$loyalty.vice_{party}$			6.252	4.07297
$authority.virtue_{party}$			50.430**	53.07479***
$authority.vice_{party}$			4.877	0.03208
$sanctity.virtue_{party}$			32.550*	33.88142*
$sanctity.vice_{party}$			-51.132**	-52.95726**
$care.virtue_{doc}$				20.4351**
$care.vice_{doc}$				-54.11249***
$fairness.virtue_{doc}$				1.70708
$fairness.vice_{doc}$				5.00505
$loyalty.virtue_{doc}$				11.12370**
$loyalty.vice_{doc}$				-19.09322***
$authority.virtue_{doc}$				-5.19724
$authority.vice_{doc}$				35.6683***
$sanctity.virtue_{doc}$				-27.63144**
$sanctity.vice_{doc}$				27.43908*
Log-Likelihood	-2038.1	-2026.6	-2007.6	-1936.6
LL(0)	-4346.0	-4346.0	-4346.0	-4346.0
BIC	4137.4	4201.9	4163.8	4196.8
Rho^2	0.53	0.53	0.54	0.55
Number of observations	6270	6270	6270	6270
Number of estimated parameters	7	17	17	37

Model 1 in table 4 is the baseline model without moral distances, exactly the same as

model 1 in table 3.

Model 2.B in table 4 shows the estimates of our first moral image distance model: the baseline model extended with the individual moral image distances between MEPs and their corresponding party groups, following equation 5.5, assuming $\beta_{m,party} = \beta_{m,doc} = 0 \quad \forall m$. We can see an improvement in the model fit; the log-likelihood ratio test shows a significant difference ($p = 0.0109$) from the baseline model. From the moral dimensions, only fairness virtue is significant, with a negative sign. Next, we estimated the defection model based on moral image distances between the national parties and EP party groups of MEPs (model 3.B of 4) based on equation 5.5, assuming $\beta_{m,ind} = \beta_{m,doc} = 0 \quad \forall m$. The model shows an even better model fit (p-value of log-likelihood ratio test against the baseline model is 0.0000), and six moral domains are significant out of ten. Finally, we estimated the model including the moral image distances between individual MEPs the documents under vote (model 4.B of 4) following equation 5.5. This model significantly outperforms the previous ones in terms of model fit, however, the benchmark model (1) has the lowest BIC. From the additional ten document-distance parameters, six are significant.

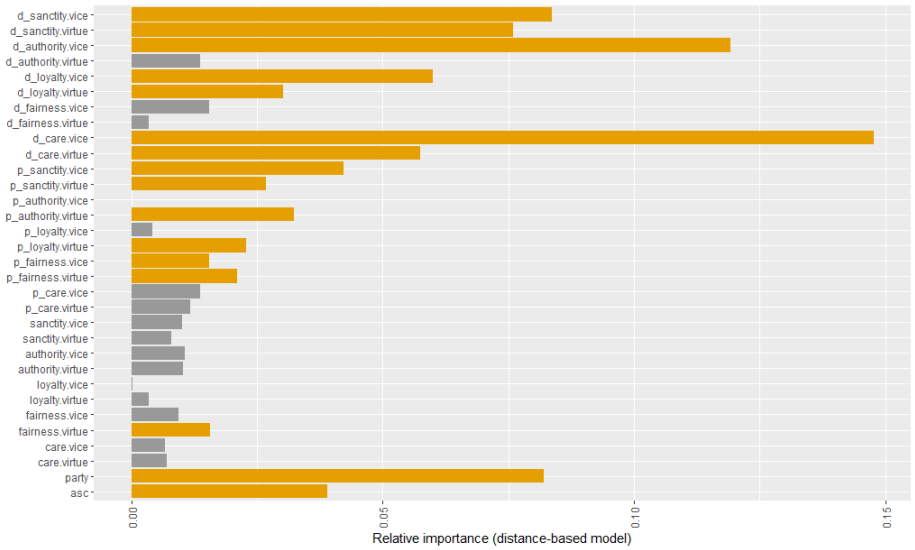


Figure 4: *Relative importance of moral images in the distance-based full model (model 4.B of table 4). Significant parameters are signalled with yellow color.*

Figure 4 shows that the relative importance of moral image distance dimensions is approximately the same party affiliations' and often considerably higher than the alternative specific constant.

Based on the distance-based models of EP party group defection (table 4), we can partly confirm expectation 1. Our results show that more dimensions of the moral-image-distance are significant, and the model fit is also better when distances are based on party difference (model 3.B of table 4) instead of individual difference (model 2.B of table 4). These results could indicate that the subtle ideological differences are captured

through language use. Therefore the distance between parties had higher explanatory power than individual distances, similarly to the results found in the literature (e.g. Hix, 2002, 2004; Klüver and Spoon, 2015). Examining the individual parameters, however, interpretation differs for dimensions with positive and negative weights.

If two moral images have a high distance, that can be attributed to either of two things: the two texts (or people or groups) are covering different subjects (for instance, one talks about decreasing the gender pay gap, and the other about protecting vulnerable animals) or they have different arguments about the same subject (one can frame the same policy as promoting gender equality or as destroying traditional family structures). Thus, high distance is expected to relate to defection. However, our results show that high distance can be related to stronger than average cohesion (positive weights, highlighted in blue in table 4) as well as to defection (negative weights, highlighted in red in table 4).

When a high distance relates to party group cohesion, political forces and other motivations may play a role, thus resulting in a seemingly counterintuitive pattern. We see, for instance, that the fairness virtue dimension has a significant negative (in blue) coefficient in model 2.B of table 4; thus relates to higher than average EP party group cohesion. This indicates that the more distant someone is from their party group in the fairness virtue dimension, the more likely they vote with their EP party group (with everything else assumed to be constant). One reason this can happen is that despite valuing fairness-related subjects to a different degree, the MEP votes in line with the party group. Then they might want to explain their views toward their constituents. MEPs can have multiple goals that affect voting behaviour, including political ambition. If an MEP intends to climb the legislature's internal hierarchy, they have a solid incentive to vote with their party groups (Meserve et al., 2009). However, if, for instance, a party group communicates a certain level of fairness-related issues that the MEP finds too low, they might want to reassure their constituents about their values and intentionally tweet more about fairness. This might happen when the MEP scores higher than their party group. The opposite can also happen; if an MEP finds the party group's communication about a value excessive, they may purposely ignore it on their social media and focus on other issues more relevant for their constituents. In this case, the distance is high because the MEP's scores are lower than the party group's. The phenomenon of politicians voting with their party but communicating something different was found by Schwarz et al. (2017) too. In their case study on the Swiss parliament, they find that text analysis reveals more considerable intra-party differences than roll-calls; thus, underlying preferences do not necessarily echo through the choices made by representatives.

In the national party distance model (model 3.B of table 4), we also find negative coefficients for three moral domains. These can be interpreted slightly differently than in the individual distance model above. The leadership of a party group can exert pressure to ensure national delegations vote inline (Hix et al., 2006). However, national party members might want to appeal to their constituents in their home country and express their different values on their social media, despite voting with the party group due to political pressure. It is also possible that to make their decision acceptable to their followers, they present their voting choice in a frame that resonates more with their follow-

ers, which can be very different from the party group's framing.

In the national party distance model (model 3.B of table 4), there are positive (in red) coefficients, too; these mean that the farther the MEP's national party scores from their EP party group, the more likely that they will defect. As MEPs most often vote with their national parties, it is intuitive that when the moral image distance is high between a national party and a party group, the delegation will likely defect. It is also possible that when there is tension between a national party and the party group (which can manifest in a high defection rate), their language use will diverge so that they distance themselves from the other. This can happen either by taking opposite stances on specific issues or by discussing different topics online. Tatalovich and Wendell (2018) presents a few examples of how morality policies are typically framed in argumentation, and Clifford and Jerit (2013) empirically shows how stem-cell research is framed relying on the foundation of care or sanctity, depending on whether the argument is "for" or "against". Slapin and Proksch (2010) looked into the relationship between giving parliamentary speeches and defecting EP party group. They found that those voting against the EP party (often being disciplined by their national party) are more likely to take the floor in parliamentary debates. The reason for this was found to be MEPs demanding speaking time to explain their defection and show their support to their national party and voters on public record. This can be a potential incentive for posting on social media too.

In model 4.B of table 4 we interacted the document distance (from individual MEPs) with the EP party group's majority voting "against". The reason for this is the following. Scoring very different from a document intuitively means that we expect the individual to vote against it or just to vote with their party group. So defection has a different interpretation when the party group preference is "for" and when it is "against". Our results show that three domains (care vice, loyalty vice and sanctity virtue) follow the intuitive pattern: those who score different from the document are more likely to vote with their party "against". However, four domains (care virtue, loyalty virtue, authority vice, and sanctity vice) show the opposite pattern. Despite the party group's preference to vote "against", when the distance between the document and MEP is high in these four domains, MEPs are more likely to defect; either by voting "for" or "abstain". This potentially signals that these values play a role in a way which does not echo through words. For instance, if someone scores very low on care virtue, and a document comes to vote which scores very high on care virtue, and the party group discipline is voting "against", then an individual voting "for" could mean they have values that they do not express on Twitter, likely because of their political agenda.

The domains that are significant when included as EP and national party distance variable, and when included as individual distance from the document too, show a consistent behaviour in the following sense. Loyalty virtue and sanctity vice display a counterintuitive pattern in both cases. Large distance between parties with respect to these domains results in cohesion, and large distance from the document along with party discipline "against" still results in "for" or "abstain". Thus issues related to these foundations are most likely to stir political pressure from the EP party group's side, or issues related to these foundations are most likely to stir moral motivations that do not echo

through words. On the other hand, sanctity virtue behaves the intuitive way in both cases: large distance between parties relates to defection, and when the distance from the document is high, those whose party group discipline is 'against' are significantly less likely to defect. Thus sanctity virtue related issues seem to be where actions and words are most aligned.

5.5.2. MODELLING VOTING OUTCOME

EXPECTATION 2

To address expectation 2, we model voting outcome, meaning whether MEPs voted in favour, against or abstention on a subject. Table 5 shows two multinomial logits with the vote as the dependent variable. Model 1.C only uses ASCs and shows that voting "for" is the most likely choice in our sample and voting "against" is also more likely than abstaining. Next, in model 2.C, we include the moral images of documents and use alternative-specific weights for them, following equations 5.10 - 5.12. The model fit improves significantly, and the additional 20 parameters are justified based on the BIC. Seven moral domains relate to significantly lower or higher than average "for" rate, and eight domains relate to significantly lower or higher than average "against" rate.

Moral images of documents seem to have explanatory power in the voting model, too, similarly to the score-based defection model (model 4.A in table 3). The results of table 5 indicate that the moral images of documents have explanatory power in modelling voting behaviour as a trinary choice; thus, expectation 2 is met. For example, high fairness virtue score in a proposal is more likely to result in abstention than on average. We see from model 4.A of table 3 that fairness virtue is also related to a higher than average defection rate. This indicates that fairness virtue is a domain that may stir defection when present in a document under vote, and it materializes in voting abstention. Authority (both virtue and vice), on the other hand, relates to significantly higher "for" and "against" votes, thus resulting in fewer abstentions. This finding is in line with the distinction of individualizing/binding foundations. Fairness is an individualizing foundation; thus, defecting one's group when a fairness-related issue is at hand is intuitive. Authority is a binding foundation; thus, the individualistic moral motivations to potentially defect play a less significant role.

The model of table 5 has the potential to predict outcomes of future votes, while retaining interpretability. To gain even more detailed insights, and potentially improve prediction, it could be a fertile ground for research to include themes (besides values), such as environment protection or gender equality, and model interactions between themes and values.

5.5.3. SUMMARY AND LIMITATIONS

Overall, our results indicate that moral images of MEPs and documents under vote have explanatory power in modeling voting behavior. Expectation 1 is partly met and expectation 2 is met. Furthermore, we can gain subtle insights about voting behavior by interpreting the modelling results, such as

- high moral image scores in individualizing foundations (i.e. care and fairness) of

Table 5: *Model of voting on documents (subsection 5.4.2 for details), dependent variable has three possible values: "for", "against" or "abstain". *,**, and *** represents significance on 5%, 1% and 0.1% accordingly.*

	(1.C)	(2.C)
ASC_{for}	1.6250***	1.743***
$ASC_{against}$	0.5480***	0.478***
$\beta_{for} care.virtue$		-25.684***
$\beta_{for} care.vice$		-32.280***
$\beta_{for} fairness.virtue$		-8.378**
$\beta_{for} fairness.vice$		8.259
$\beta_{for} loyalty.virtue$		13.662***
$\beta_{for} loyalty.vice$		-10.608**
$\beta_{for} authority.virtue$		30.026***
$\beta_{for} authority.vice$		25.748***
$\beta_{for} sanctity.virtue$		-9.285
$\beta_{for} sanctity.vice$		3.154
$\beta_{against} care.virtue$		-27.526***
$\beta_{against} care.vice$		22.930**
$\beta_{against} fairness.virtue$		-22.623***
$\beta_{against} fairness.vice$		-13.274*
$\beta_{against} loyalty.virtue$		0.553
$\beta_{against} loyalty.vice$		-9.907*
$\beta_{against} authority.virtue$		43.219***
$\beta_{against} authority.vice$		19.534**
$\beta_{against} sanctity.virtue$		-24.376***
$\beta_{against} sanctity.vice$		11.606
Log-Likelihood	-5498.0	-4971.4
LL(0)	-6888.3	-6888.3
BIC	11013.6	10135.2
Rho^2	0.20	0.28
Number of observations	6270	6270
Number of estimated parameters	2	22

individual MEPs, national parties or documents under vote relate to higher than average party group defection,

- binding foundations (loyalty, authority and sanctity) can relate to both cohesion and defection,
- individualizing foundations in documents more often result in abstentions,
- issues related to loyalty virtue and sanctity vice are most likely to stir political pressure from the EP party group's side, or issues related to these foundations are most likely to stir moral motivations that do not echo through words,
- sanctity virtue is the moral image domain where actions and words behave consistently in an intuitive way.

These interpretations must be taken with caution. There are limitations in the case study due to possible selection bias. In our data, those not present at the voting are not included. However, not showing up could also be a strategy similar to abstention, revealing even less information on an MEP's preferences. Furthermore, we only have data on the tweeting MEPs. MEPs who do not tweet may adopt a different voting strategy than those who do. Lastly, roll-call votes, as they are only part of the legislative decisions, were also argued to cause selection bias (Carrubba et al., 2006); however, Hix et al. (2018) found this effect of being negligible in the EP.

FUTURE RESEARCH DIRECTIONS IN THE VOTING BEHAVIOURAL CONTEXT

The above section described several possible reasons for particular signs of our estimated parameters. This paper does not attempt to disentangle the possible effects further. However, there are several ways to go deeper into modelling and answer a wide range of possible research questions regarding moral policymaking. Including, but not limited to:

1. Why do MEPs vote with their national parties as opposed to their EP party group? This can have several practical reasons. For instance, Hix (2004) found that country-specific institutions which reinforce the control national parties can exert over their members increase MEPs' defection of their EP party groups. Faas (2002, 2003) found that MEPs whose reelection is more dependent on their national parties are more likely to defect. These are national parties that have a centralized candidate selection method, invest more in monitoring their members or are in government in their home countries. Lindstädt et al. (2011) finds that proximity to elections in the home country shifts MEPs' votes towards their national parties' when principles of the national party and EP party group conflict. Examining moral images along these empirically observed effects may shed light on when defection is more likely to be strategic and when is it conviction.
2. How do observations on moral voting behaviour differ across topics? For example, Klüver and Spoon (2015) found that the more salient an issue is to a national party, the stronger the effect of the ideological distance between the national party and its EP party group on MEP defection. In order to gain insight on, for instance,

how gender equality related topics differ from the rest, one can model moral image distances or document scores on divided data sets or include topic-specific categorical variables.

3. How do observations on moral voting behaviour differ across political groups? Different national parties and EP party groups may differ in their relations to moral domains. Individualizing foundations might be more robust predictors in modelling behaviour in progressive parties than in more conservative parties. Such difference, although not in the voting context, was observed in behavioural economics games (Clark et al., 2017).

4. How do behavioural findings change through time? Many studies examined how political rhetoric (e.g. Slapin and Proksch, 2008) or voting behavior changes over time (e.g. with election cycles, Lindstädt et al., 2011). Using moral images in discrete choice models can also shed light on the changing relationship between rhetoric and vote: for instance, is the domain of care virtue (strongly related to health/healthcare) differently related to voting behaviour before and after COVID-19?

Wendell and Tatalovich (2021) argues that some policies are more value-laden than others. There are mixed and pure morality policies. Moral image extraction is suitable for both mixed and pure morality policies, as, through similarity scores, it is expected to reflect a mixed nature compared to pure morality policies. In the political science direction, other case studies could use moral images, for instance, examining the voting behaviour of the general public. For this, social media feed or other collected text data, such as a values essay or opinion description about specific topics, could be used.

5.6. DISCUSSION

In this paper, we proposed a method for enriching discrete choice models with moral images extracted from text, thus connecting the two ways morality can manifest itself: words and actions. We used state-of-the-art Natural Language Processing methods and a well-established moral psychological taxonomy of values, Moral Foundations Theory. We showed in a case study of voting in the European Parliament what subtle insights such moral image models could provide and discussed other potential applications. Note, however, that in any potential applications, one must be careful with interpretation; there is a complex relationship between moral language, judgement and behaviour, and causal directions are not straightforward. People can have various incentives to hide or obfuscate their true moral judgement when speaking or making decisions, including insecurity, fear of social disapproval, or intention to convince others about something. It is also possible, that in some morally salient situations, people act based on intuition, and create a rational narrative after taking action (Haidt, 2001).

The method proposed in this study allows the researcher to study to what extent are words and actions aligned, that can give insights into how strategic behaviour plays a

role in various situations. Moral images can help to identify latent behavioural constructs in more complex discrete choice models such as latent class models or latent variable models. Latent class models are often used to identify classes with moral motivations. Using questionnaires or product attributes was instrumental in finding consumption or behavioural patterns across different classes of people. For instance, Zha et al. (2020) identifies environmentally responsible classes when buying electric appliances, or Langen (2011) identifies groups with different attitudes towards fair trade, organic production and donations when buying coffee. Moral images may help to identify that in moral situations, groups that display different behavioural patterns can be identified partly based on their language use.

Furthermore, it has been theorized and empirically found that latent variables have a significant effect on what decision-making rule is applied (Hess and Stathopoulos, 2013). In moral contexts, this could be even more relevant for two reasons. First, utility maximization is often replaced by moral heuristics (Gigerenzer, 2010; Sunstein, 2005), which give a wide range of possible decision rules. Second, deep-seated moral values can affect several observable factors, for instance, political affiliation, the way one talks (thus their moral image) and their choices in moral situations. Thus, the combination of alternative decision-making rules and moral images can be instrumental in latent variable latent class modelling.

Latent motivations to take one action or another in moral questions potentially have a more complex form than the standard linear additive specification. Different attributes in a choice task might interact with, for instance, how strongly one talks about fairness. Different attributes in a choice task might interact with, for instance, how strongly one talks about fairness. Less obvious interaction effects could also be discovered with, for example, machine learning-assisted methods (see, e.g. Hillel et al., 2019). Such methods can improve the utility specification and thereby improve predictions based on language use while retaining the interpretability of the discrete choice model. Such enriched moral DCMs have the potential to advance persuasion techniques, which have relevance in various applications and research fields, such as marketing, psychology, political science, or policy design.

Future ways for methodological research in moral images could involve comparison with text classification methods (i.e. the probability of a text belonging to a specific foundation) where the moral image would be a composite of probabilities and not similarity scores. Future research on the implications of moral image analysis could be cross-cultural or cross-contextual comparisons: some moral domains may prove to be robust in explaining actions, while some are not. They may vary across cultures, times, or the decision-making situation. Such knowledge can give valuable insights on communication strategies and behavioural phenomena through easily obtainable text data.

APPENDIX

5.A. SUBJECTS OF ROLL-CALL VOTES

1. "2019-2020 Reports on Bosnia and Herzegovina"
2. "2019-2020 Reports on Kosovo"

3. "A WTO-compatible EU carbon border adjustment mechanism"
4. "Artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice"
5. "Children rights in occasion of the 30th anniversary of the Convention of the Rights of the Child"
6. "Decent and affordable housing for all"
7. "EU Association Agreement with Ukraine"
8. "Human Rights and Democracy in the World and the EU policy on the matter - annual report 2019"
9. "Human rights and political situation in Cuba"
10. "Human rights situation in Kazakhstan"
11. "Meeting the Global Covid-19 challenge: effects of waiver of the WTO TRIPS agreement on Covid-19 vaccines, treatment, equipment and increasing production and manufacturing capacity in developing countries"
12. "Prisoners of war in the aftermath of the most recent conflict between Armenia and Azerbaijan"
13. "Promoting gender equality in science, technology, engineering and mathematics (STEM) education and careers"
14. "Reducing inequalities with a special focus on in-work poverty"
15. "Regulatory fitness, subsidiarity and proportionality - report on Better Law-Making 2017, 2018 and 2019"
16. "Search and rescue in the Mediterranean"
17. "Strengthening the single market: the future of free movement of services"
18. "Systematic repression in Belarus and its consequences for European security following abductions from an EU civilian plane intercepted by Belarusian authorities"
19. "The EU Strategy for Gender Equality"
20. "The adequate protection of personal data by the United Kingdom"
21. "The gender perspective in the COVID-19 crisis and post-crisis period"
22. "The impact of Covid-19 on youth and on sport"
23. "The right to disconnect"
24. "EU accession to the Istanbul Convention and other measures to combat gender-based violence"

REFERENCES

Araque, O., Gatti, L., & Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-based systems*, *191*, 105184.

Baburajan, V., e Silva, J. d. A., & Pereira, F. C. (2020). Open-ended versus closed-ended responses: A comparison study using topic modeling and factor analysis. *IEEE Transactions on Intelligent Transportation Systems*, *22*(4), 2123–2132.

Baburajan, V., e Silva, J. d. A., & Pereira, F. C. (2022). Open vs closed-ended questions in attitudinal surveys–comparing, combining, and interpreting using natural language processing. *Transportation research part C: emerging technologies*, *137*, 103589.

Boyd, R., Wilson, S., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1).

Carrubba, C. J., Gabel, M., Murrah, L., Clough, R., Montgomery, E., & Schambach, R. (2006). Off the record: Unrecorded legislative votes, selection bias and roll-call vote analysis. *British Journal of Political Science*, *36*(4), 691–704.

Clark, C. B., Swails, J. A., Pontinen, H. M., Bowerman, S. E., Kriz, K. A., & Hendricks, P. S. (2017). A behavioral economic assessment of individualizing versus binding moral foundations. *Personality and Individual Differences*, *112*, 49–54.

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics*, *75*(3), 659–671.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69.

Faas, T. (2002). *Why do meps defect?: An analysis of party group cohesion in the 5th european parliament*. ECSA-Austria.

Faas, T. (2003). To defect or not to defect? national, institutional and party group pressures on meps and their consequences for party group cohesion in the european parliament. *European Journal of Political Research*, *42*(6), 841–866.

Frimer, J. A. (2020). Do liberals and conservatives use different moral languages? two replications and six extensions of graham, haidt, and nosek's (2009) moral text analysis. *Journal of Research in Personality*, *84*, 103906.

Frimer, J. A., Boghrati, R., Haidt, J., Graham, J., & Dehgani, M. (2019). Moral foundations dictionary for linguistic analyses 2.0. *Unpublished manuscript*.

Frimer, J. A., Tell, C. E., & Haidt, J. (2015). Liberals condemn sacrilege too: The harmless desecration of cerro torre. *Social Psychological and Personality Science*, *6*(8), 878–886.

Gerrish, S., & Blei, D. (2010). The ideal point topic model: Predicting legislative roll calls from text. *Proceedings of the Computational Social Science and the Wisdom of Crowds Workshop. Neural Information Processing Symposium*.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science*, *2*(3), 528–554.

Glerum, A., Atasoy, B., & Bierlaire, M. (2014). Using semi-open questions to integrate perceptions in choice models. *Journal of choice modelling*, *10*, 11–33.

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.

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

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.

Hess, S., & Stathopoulos, A. (2013). A mixed random utility—random regret model linking the choice of decision rule to latent character traits. *Journal of choice modelling*, *9*, 27–38.

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). Weak teachers: Assisted specification of discrete choice models using ensemble learning. *8th Symposium of the European association for research in transportation, Budapest*.

Hix, S. (2002). Parliamentary behavior with two principals: Preferences, parties, and voting in the european parliament. *American Journal of Political Science*, 688–698.

Hix, S. (2004). Electoral institutions and legislative behavior: Explaining voting defection in the european parliament. *World politics*, *56*(2), 194–223.

Hix, S., Noury, A., & Roland, G. (2006). Dimensions of politics in the european parliament. *American Journal of Political Science*, *50*(2), 494–520.

Hix, S., Noury, A., & Roland, G. (2018). Is there a selection bias in roll call votes? evidence from the european parliament. *Public Choice*, *176*(1), 211–228.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, *11*(8), 1057–1071.

Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2021). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, *53*(1), 232–246.

Kaur, R., & Sasahara, K. (2016). Quantifying moral foundations from various topics on twitter conversations. *2016 IEEE International Conference on Big Data (Big Data)*, 2505–2512.

Kim, I. S., Londregan, J., & Ratkovic, M. (2018). Estimating spatial preferences from votes and text. *Political Analysis*, *26*(2), 210–229.

Kivikangas, J. M., Lönnqvist, J.-E., & Ravaja, N. (2017). Relationship of moral foundations to political liberalism-conservatism and left-right orientation in a finnish representative sample. *Social Psychology*.

Klüver, H., & Spoon, J.-J. (2015). Bringing salience back in: Explaining voting defection in the european parliament. *Party Politics*, *21*(4), 553–564.

Korn, J. W., & Newman, M. A. (2020). A deep learning model to predict congressional roll call votes from legislative texts. *Machine Learning and Applications: An International Journal (MLAIJ) Vol*, *7*.

Kraft, P., Jain, H., & Rush, A. M. (2016). An embedding model for predicting roll-call votes. *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2066–2070.

Langen, N. (2011). Are ethical consumption and charitable giving substitutes or not? insights into consumers' coffee choice. *Food Quality and preference*, *22*(5), 412–421.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, *58*(3), 754–771.

Lindstädt, R., Slapin, J. B., & Vander Wielen, R. J. (2011). Balancing competing demands: Position taking and election proximity in the european parliament. *Legislative Studies Quarterly*, *36*(1), 37–70.

Meserve, S. A., Pemstein, D., & Bernhard, W. T. (2009). Political ambition and legislative behavior in the european parliament. *The Journal of Politics*, *71*(3), 1015–1032.

Mutlu, E., & Tütüncüler, E. (2020). Moral rhetoric of politicians in social media: How republicans and democrats differ in moral values.

Patkós, V. (2022). Measuring partisan polarization with partisan differences in satisfaction with the government: The introduction of a new comparative approach. *Quality & Quantity*, 1–19.

Pereira, F. C. (2019). Rethinking travel behavior modeling representations through embeddings. *arXiv preprint arXiv:1909.00154*.

Proksch, S.-O., & Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, *40*(3), 587–611.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social science computer review*, *32*(2), 132–144.

Salmela, M., & Von Scheve, C. (2017). Emotional roots of right-wing political populism. *Social Science Information*, *56*(4), 567–595.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in experimental social psychology* (pp. 1–65). Elsevier.

Schwarz, D., Traber, D., & Benoit, K. (2017). Estimating intra-party preferences: Comparing speeches to votes. *Political Science Research and Methods*, *5*(2), 379–396.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705–722.

Slapin, J. B., & Proksch, S.-O. (2010). Look who's talking: Parliamentary debate in the european union. *European Union Politics*, *11*(3), 333–357.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and brain sciences*, *28*(4), 531–541.

Tatalovich, R., & Wendell, D. G. (2018). Expanding the scope and content of morality policy research: Lessons from moral foundations theory. *Policy Sciences*, *51*(4), 565–579.

Turk, Ž. (2019). Subsidiarity and the moral foundations of populism. *European View*, *18*(1), 71–79.

van den Broek-Altenburg, E., Gramling, R., Gothard, K., Kroesen, M., & Chorus, C. G. (2020). Exploring heterogeneity in moral terminology used by patients in palliative care consultations. *The Patient*, *13*(1), 139–140.

van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2021). Choice modelling in the age of machine learning-discussion paper. *Journal of Choice Modelling*, 100340.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 5998–6008.

Vij, A., & Walker, J. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, *90*, 192–217.

Wendell, D. G., & Tatalovich, R. (2021). Classifying public policies with moral foundations theory. *Policy Sciences*, *54*(1), 155–182.

Zha, D., Yang, G., Wang, W., Wang, Q., & Zhou, D. (2020). Appliance energy labels and consumer heterogeneity: A latent class approach based on a discrete choice experiment in china. *Energy Economics*, *90*, 104839.

6

CONCLUSION

This is a concluding chapter discussing scientific reflection and practical implications of the research findings of this thesis in detail. The main research aim of this thesis is *to develop and evaluate the potential of new discrete choice modelling methods to identify latent morality, thus increasing the behavioural realism of discrete choice models (DCMs) in moral decision-making*. To do this, I used two approaches. The first one focuses on parameter identifiability in novel models that have the potential to capture behaviours characterizing moral decision-making when only choice data is available (Part I). The second one makes use of additional psychometric data to identify latent behavioural constructs related to morality (Part II). In section 6.1 and 6.2, I address these two approaches accordingly by first summarizing the research outcomes and then discussing the scientific and practical implications along with future research directions. Section 6.3 takes a broader perspective to reflect on the future of morality research in discrete choice models.

6.1. PART I: IDENTIFIABILITY WHEN ONLY CHOICE DATA IS AVAILABLE

Mainstream discrete choice models' great advantage is that parameter identifiability has a well-established line of literature, and the behavioural interpretations are straightforward. Alternative discrete choice models (1.1.1) often have the potential to capture moral behavioural phenomena, such as contemplation (Chapter 2) or obfuscation (Chapter 3), but their identifiability, and as a consequence, the interpretation of their estimated parameters, is often not straightforward. These models do not use additional data to choices, and as using only choices is standard in many fields, from health care to transportation, examining parameter identifiability can determine what behavioural inferences can be drawn from the estimates.

In the first study concerning parameter identifiability (Chapter 2) I examined Decision Field Theory's (DFT's) parameters to see whether the contemplation process can be

uniquely recovered from choice data alone. To do this, I used analytical derivations and Monte Carlo simulations. Results show that DFT's process parameters are not uniquely identifiable in two special cases, and the full model's process parameters are biased.

In the second study (Chapter 3), I examined whether a decision-maker trying to obfuscate their underlying preferences from an onlooker affects an analyst trying to recover their preferences. To do this, I used Monte Carlo analysis on synthetic data. Unsurprisingly, a modeller ignoring obfuscation results in biased parameter estimates. Taking the obfuscation into account allows for the joint identification of the preference- and the obfuscation parameter; however, as the obfuscation intention increases, the standard error of each estimate increases, and the modeller can be less confident about the estimates.

6.1.1. IMPLICATIONS AND FUTURE RESEARCH RECOMMENDATIONS

1. Conclusions should not be drawn on the mental process from choice data alone; collecting richer data is advisable. Results in Chapter 2 indicate that in some cases, time and memory cannot be interpreted as their names and the corresponding theory suggest. Specifically, someone who deliberates for a long time and has bad memory is indistinguishable from someone who has a perfect memory and decides fast. Suppose obtained estimates were to infer, for example, how much time one needs to be prompted to choose a desirable outcome. In that case, one could use this information to incentivize either a long deliberation process or a quick decision. In such cases, if the nudge is designed based on an unidentifiable model's estimates, it could lead to ineffective or counterproductive policies. To avoid this, policymakers and practitioners should not draw conclusions on the mental process from choice data alone. Collecting richer data, for instance, decision-making time (Hancock et al., 2018) or attention via eye-tracking (Noguchi and Stewart, 2014), is recommended to identify the deliberation processes.

2. DFT and other process models when relying on choice data alone, should be compared to machine learning methods rather than standard DCMs. Establishing parameter identification is a must if an analyst aims to draw behavioural conclusions from the estimated parameters. The results of Chapter 2 indicate that the unidentifiable and indistinguishable specifications of DFT should not be used to draw behavioural conclusions. In some cases, the unidentifiable DFT models might provide better model fit and predictions than a traditional multinomial logit; however, as the additional parameters cannot be interpreted, the models' performance should be measured against advanced machine learning classification methods, such as deep learning, which tend to outperform DCMs in model fit and prediction significantly.

3. Unidentifiable structures should be flagged in software. Discrete choice modelling software, such as Biogeme (Bierlaire, 2020), Apollo (Hess and Palma, 2019) or mixl (Molloy et al., 2021) make the estimation of various choice models easier for practitioners. Estimating an unidentifiable model often results in infinite standard errors, which immediately signal that the model is unidentifiable. However, it is not

always that easy to spot identification problems. Including warnings in the case of unidentifiable models could help practitioners avoid such pitfalls and misguided conclusions.

4. DFT's and other process models' identifiability should be further researched. The full DFT model's identifiability should be further examined, with various data generating processes, to establish under what conditions capture the parameters what they aim to capture. The same should be done for other process models that have the potential to capture contemplation processes; for instance selective integration (Tsetsos et al., 2012) or drift-diffusion (Krajbich and Rangel, 2011). To examine DFT's identifiability, Chapter 2 derived under what conditions the model is equivalent to probit models. Thus, established methods from the field of choice modelling allowed for examining identification in special cases when the conditions are met. This method has the potential for other process models too. For an overview of the various advantageous properties of such models (one of them being DFT), see Busemeyer et al. (2019). Building bridges between psychological models and traditional discrete choice models is an important endeavour to increase behavioural realism, mainly through the effect of time and sudden prompts that trigger attention to a particular attribute or trigger emotions. Implementing psychological models in discrete choice modelling contexts has been a subject of recent literature (e.g., Hancock et al., 2021). Establishing identifiability in adapted models by deriving equivalence and applying methods from discrete choice modelling or by using synthetic data simulation is crucial before drawing conclusions from the parameters.

5. Obfuscation should be further researched. The analytical identifiability analysis of the obfuscation model breaks down to that of the multinomial logit model. Thus, theoretically, the model is identifiable. However, addressing empirical identifiability with Monte Carlo simulations provided scientifically relevant insights: if someone obfuscates their preferences from an observer increasingly, the analyst (not equal to the observer) will be less confident in their estimates. The rate of decreasing certainty differs if the analyst examines a single-choice obfuscator or a sequential obfuscator (a variation of the original model proposed in Chapter 3). As the size of the standard errors largely depend on the size of the data, it could be established that for various levels of obfuscation intention, how much more data is required to achieve the same precision about the estimates than without the obfuscation intention. It is also worth noting that the estimates depend on the decision-maker's interaction with another agent. Whether someone obfuscates, in what way, and to what extent all affect the estimates, even is the analyst perfectly aware of them and takes them into account. The results of chapter 3 are promising; however, further steps need to be taken for the model to be used in a broad range of use cases. Practical implementation of the obfuscation model is yet to establish best practices. Chapter 3 relaxes one behavioural assumption to increase the model's realism. For future investigations, it is essential to establish what possible states of the world (i.e., potential preference weights that the decision-maker assumes their onlooker assumes about the decision-maker) can or should be as-

sumed by the analyst. In current practice, it is a set of discrete values that rapidly increase the computation time required by the model as analysts wish to include more and more possible states. For instance, instead of assuming[1] a decision-maker's preference for travel time is either 1 or 2, the analyst may assume it is 1, 1.25, 1.5, 1.75, or 2 to be more realistic. However, the computation time of the model will be significantly larger, especially if there are several attributes. Another unrealistic expectation from an analyst is to know the parameter boundaries (i.e., knowing the preference weight is between 1 and 2). In reality, an analyst is unlikely to know this; thus, examining how giving wide ranges for possible betas affects the estimation is crucial for operationalizing the model. Formulating the model with continuous betas (instead of discrete states-of-the-world) could help computation time problems and behavioural realism simultaneously. Including the entropy term also requires further steps, as it can be used with an infinite amount of logarithm base, it can be scaled infinite ways and use several heuristics (such as including the entropy term as a dummy). A systematic study on how different decisions about modelling affect the estimations could help establish good practices for the obfuscation model. Furthermore, as obfuscation can be linked to compromise alternatives (Chorus et al., 2021), comparative studies with models such as the contextual concavity model or the compromise variable model (Chorus and Bierlaire, 2013; Kivetz et al., 2004) are also recommended to establish which models are best for particular applications.

6.2. PART II: IDENTIFYING MORAL MOTIVATIONS USING ADDITIONAL DATA TO CHOICES

Discrete Choice Theory can not only be extended through structural or modelling innovations to identify moral motivations. Enriching mainstream models (which are identifiable) with psychometric data can also lead to choice models that give valuable insights into choice behaviour in light of moral psychology. In Part II of this thesis, psychometric data is collected on morality: one measured with a standard survey and one extracted from text data using Natural Language Processing. Part II uses standard discrete choice models (i.e., multinomial and mixed logit) to identify behavioural constructs in situations involving morality, such as social routing (i.e., sacrificing individual travel time gain to decrease congestion and benefit all road users) or voting in the European Parliament about various subjects concerning, for example, gender equality or a refugee crisis.

In the first study (Chapter 4), I examine the effects of moral incentives and moral personality on decision-making in a social routing problem. To do this, I used psychometric data from a well-established standard questionnaire in the field of moral psychology (corresponding to the Moral Foundations Theory) and estimated standard discrete choice models. Results indicate that moral personality significantly affects the willingness to participate in different social routing schemes, thus affecting the effectiveness

[1] The assumption about the possible states of the world is given to the model by the analyst, based on what the decision-maker believes about their observer (i.e., what the potential observer believes about the decision-maker).

and distribution of burdens and benefits of these schemes.

In the second study (Chapter 5), I examined how decision-makers' free text[2] relates to their moral choice behaviour. To do this, I extract moral values from free text using Natural Language Processing and the well-established dictionary on moral values (again, corresponding to Moral Foundations Theory) and estimate discrete choice models on voting in the European Parliament. Results indicate that moral rhetoric has significant explanatory power in such models, and outcomes can be interpreted in light of political science literature.

6.2.1. IMPLICATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

1. Morality data should be added to advanced DCMs to identify moral motivations and latent behavioural constructs. DCMs with highly sophisticated behavioural assumptions, such as latent class models, or latent variables determining which decision rules to apply (Hess and Stathopoulos, 2013), are particularly likely to benefit from additional morality data. For instance, those who use more fairness-related verbal utterances might be more likely to apply a tit-for-tat heuristic. Such DCMs potentially achieve higher behavioural realism, and can be used to verify or refute moral psychological theories. This thesis used linear additive multinomial and mixed logit models to model moral choice behaviour, and draw conclusions relying on moral psychology and political science literature. Insights from choice models assisted with machine learning that can explore a wide range of specifications including interactions (Hillel et al., 2019), may also be used to generate new moral psychological theories.

2. Morality-enhanced DCMs should be used to gain information on the systematic relationships between judgement, action and language use. A straightforward relationship between *language and judgement* would be that people articulate their moral judgements by using more of the foundations that are more important to them and less those which are not. However, this is often not the case. People may want to obfuscate or hide their true moral judgements because of insecurity, shame, guilt, or to comply with group norms. Language and judgement disparity might be even more relevant in the case of politicians or those who wish to persuade others. These agents use their words to create narratives in a campaign, explain themselves to or convince a possibly large set of people: fellow politicians, constituents or others they wish to influence.
   The relationship between *judgement and behaviour* is the basis of discrete choice theory. Most models assume that people's preferences echo through the decisions they make. Therefore, their preferences can be inferred by observing their decisions under various circumstances. However, similar to their language, people can try to hide their true intentions in their actions, too (Chorus et al., 2021). Furthermore, several studies (e.g., Haidt, 2001; Skitka and Mullen, 2002) argued that in moral situations, people tend to make decisions first based on their emotions or

[2] By free text, I refer to text that was generated by the decision-makers on their own will without any constraints (i.e., not answers to particular questions).

intuitions, then their rational minds create a corresponding moral judgement or reasoning; thus, preference does not really create the choice, but the choice creates the preference.
Finally, the case study of chapter 5, examined the relationship between *language and behaviour*. Again, there could be a straightforward relationship between words and actions: one explaining their judgement which underpins their decision. However, similarly to the above point, one may decide based on emotions and then create the explanation post hoc. One may also use their words strategically; knowing their decision, for example, politicians may intentionally create a moral narrative to it. Modelling language use, general moral judgement and actions together in discrete choice models can give valuable insights on the interplay among language, judgement and behaviour; when are they aligned, when are they not, what systematic relations can be found through time, cultures, or contexts. Furthermore they can be valuable tools for studying theories, such as emotional decision followed by rationalization by Haidt (2001).

3. Policy design should be enhanced with morality data. Knowing how decision-making relates to moral values or expressions used in one's natural language helps design effective persuasion techniques. Thus, the use of moral scales in discrete choice experiments and models is recommended to increase the models' behavioural accuracy and design more efficient policies or incentives. Moral scales are often measured in the population; for instance, yourmorals.org measures the moral values of MFT within the United States. This allows practitioners to see which values are endorsed and to what extent in particular age groups or counties. Thus, they can design moral incentives for specific target groups, such as the collective good based social routing scheme in Chapter 4 to appeal to fairness-oriented commuters.
Such usage of moral scales also sheds light on how different incentives or policies distribute the burden in society. As we can see in Chapter 4 for instance, in a sacrifice-based scheme, those who score high on care will carry all the burden for those who score lower on care. Whether this is desirable is a matter of social justice; taking ethical considerations into account when designing policies is increasingly more important in several fields, for instance, in transportation (e.g., Lucas et al., 2016). Understanding morality and how it relates to individuals' decision-making and outcomes is essential when designing new technologies that rely on moral incentives.

4. Combining data-driven and theory-driven modelling techniques is a promising future research avenue. The results of Chapter 5 indicate that using free text data and Natural Language Processing is a promising research avenue in discrete choice modelling. This method, which models choices with the help of words, uses the powers of an unidentifiable model (NLP) in identifiable models firmly rooted in behavioural theory (DCM).
Predicting different outcomes that relate to morality, such as judicial decisions

(Aletras et al., 2016) or voting on legislative bills (Korn and Newman, 2020; Kraft et al., 2016) relying solely on text data is the subject of various studies in the field of machine learning. The main goal of these studies is accurate prediction and pattern extraction, mainly through the topical content of the text (e.g., Aletras et al., 2016). Compared to these methods, which solely rely on text data, and efficient (in terms of prediction) 'black-box' classification methods, such as deep learning, the method presented in Chapter 5 is more theory-driven.
Instead of focusing on whether a law is passed or not (or, in other contexts: what are the market shares of different products or travel modes), modelling is often concerned with individual behaviour. Voting outcomes are then modeled with econometric methods (Hix et al., 2006), or location-based voting models (Godbout and Høyland, 2011). These are usually rooted strongly in theory and often rely on additional data on the individuals, which are collected through questionnaires, for instance, on where individuals are located on the left-right spectrum or the EU-integration spectrum (e.g., Hix et al., 2006). These, contrary to the previous approach that predicts the final outcome, can give insight into what drives a person's individual choice behaviour. Chapter 4 is positioned in this line of literature. The approach presented in Chapter 5 lies between these two: it relies on free-text data and natural language processing, modelling the choice at the individual level. In other words, it uses a data-driven method to create input in a theory-driven method to draw moral behavioural conclusions. The case study uses both personal level data (e.g., tweets of individuals) and choice-alternative level data (e.g., the text of a bill) to create moral features in a choice situation, which are then used in discrete choice models allowing for theory-driven behavioural interpretations. This method thus gives the advantage of not relying on questionnaires that are often problematic, costly or biased due to the closed-ended nature of questionnaires. Although their predictive accuracy is lower than that of the more data-driven methods, the outcomes are behaviourally interpretable.

6.3. OUTLOOK

These studies aimed to extend DCMs, partly because DCMs have a welfare analysis framework, making them appealing in the economic appraisal of different policies. Many moral aspects of a decision can be evaluated with traditional methods; for instance, DCMs can be used to measure one's willingness to pay for fair trade or local product labels (Howard and Allen, 2008; Rotaris and Danielis, 2011). However, extending the economic appraisal framework to the moral domains is not straightforward in either of the studies presented in this thesis. When models are not consistent with random utility maximization, welfare implications cannot be derived (e.g., Hess et al., 2018). For evaluating obfuscation or moral values (which are all technically present in discrete choice models the way, for instance, travel time is), several philosophical questions arise, which need to be addressed if one wants to put a value on obfuscation, fairness or loyalty. Similar questions arise when one wants to evaluate another model capturing morality, the taboo trade-off model (Chorus et al., 2018) and potentially many others where the basic structure is standard logit, but additional variables (constructed based on theory) cap-

ture moral values or behavioural effects. The most critical problems with evaluating, for example, fairness are 1, the different perceptions one has about fairness and 2, the unstable nature of preferences. Regarding the first problem, free text data can be a first step toward the solution: allowing individuals to express how they perceive a task can help quantify how relevant fairness is compared to other foundations. Regarding the second problem, examining the relations between values, words and actions and observing their systematic relations through time can help establish new ways of evaluating morality.

REFERENCES

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, *2*, e93.

Bierlaire, M. (2020). A short introduction to pandasbiogeme (technical report transpor 200605). *Transport and Mobility Laboratory, ENAC, EPFL: Lausanne, Switzerland*.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in cognitive sciences*, *23*(3), 251–263.

Chorus, C. G., & Bierlaire, M. (2013). An empirical comparison of travel choice models that capture preferences for compromise alternatives. *Transportation*, *40*(3), 549–562.

Chorus, C. G., Pudāne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: A discrete choice model and empirical analysis. *Journal of choice modelling*, *27*, 37–49.

Chorus, C. G., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. *Mathematical Social Sciences*, *109*, 28–44.

Godbout, J.-F., & Høyland, B. (2011). Legislative voting in the canadian parliament. *Canadian Journal of Political Science/Revue canadienne de science politique*, *44*(2), 367–388.

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

Hancock, T. O., Hess, S., & Choudhury, C. (2018). Incorporating response time in a decision field theory model. *The Transportation Research Board (TRB) 97th Annual Meeting*.

Hancock, T. O., Hess, S., Marley, A. A., & Choudhury, C. F. (2021). An accumulation of preference: Two alternative dynamic models for understanding transport choices. *Transportation Research Part B: Methodological*, *149*, 250–282.

Hess, S., Daly, A., & Batley, R. (2018). Revisiting consistency with random utility maximisation: Theory and implications for practical work. *Theory and Decision*, *84*(2), 181–204.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Hess, S., & Stathopoulos, A. (2013). A mixed random utility—random regret model linking the choice of decision rule to latent character traits. *Journal of choice modelling*, *9*, 27–38.

Hillel, T., Bierlaire, M., Elshafie, M., & Jin, Y. (2019). Weak teachers: Assisted specification of discrete choice models using ensemble learning. *8th Symposium of the European association for research in transportation, Budapest*.

Hix, S., Noury, A., & Roland, G. (2006). Dimensions of politics in the european parliament. *American Journal of Political Science*, *50*(2), 494–520.

Howard, P. H., & Allen, P. (2008). Consumer willingness to pay for domestic 'fair trade': Evidence from the united states. *Renewable Agriculture and Food Systems*, *23*(3), 235–242.

Kivetz, R., Netzer, O., & Srinivasan, V. (2004). Alternative models for capturing the compromise effect. *Journal of marketing research*, *41*(3), 237–257.

Korn, J. W., & Newman, M. A. (2020). A deep learning model to predict congressional roll call votes from legislative texts. *Machine Learning and Applications: An International Journal (MLAIJ) Vol*, *7*.

Kraft, P., Jain, H., & Rush, A. M. (2016). An embedding model for predicting roll-call votes. *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2066–2070.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Lucas, K., Van Wee, B., & Maat, K. (2016). A method to evaluate equitable accessibility: Combining ethical theories and accessibility-based approaches. *Transportation*, *43*(3), 473–490.

Molloy, J., Becker, F., Schmid, B., & Axhausen, K. W. (2021). Mixl: An open-source r package for estimating complex choice models on large datasets. *Journal of choice modelling*, *39*, 100284.

Noguchi, T., & Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, *132*(1), 44–56.

Rotaris, L., & Danielis, R. (2011). Willingness to pay for fair trade coffee: A conjoint analysis experiment with italian consumers. *Journal of Agricultural & Food Industrial Organization*, *9*(1).

Skitka, L. J., & Mullen, E. (2002). The dark side of moral conviction. *Analyses of Social Issues and Public Policy*, *2*(1), 35–41.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences*, *109*(24), 9659–9664.

ACKNOWLEDGEMENTS

First and foremost I would like express my gratitude to my promotor, Caspar Chorus and my daily supervisor, Sander van Cranenburgh. Caspar, I am grateful to you for giving me the opportunity to be part of this project, and for your all-encompassing guidance throughout the past five years. Sander, our brainstorming conversations, your constant guidance and feedback were invaluable, I sincerely thank you for that.

I would like to express my gratitude to the BEHAVE-group; Andreia, Tanzhe, Tom, Nicolas and Aemiro. Thank you for your friendship, the fun conversations, and our collaboration on the joint project, which resulted a chapter in this thesis. I could not have undertaken this journey without Paul Koster, who guided me in my first years of academia and suggested this PhD program to me, for which I am truly grateful. I am also thankful for the mentorship of Anae Sobhani, who advised me during the first year of my PhD. I want to give my deepest appreciation to the committee members, who accepted the invitation to review this thesis. I also want to thank the academic community, anonymous reviewers who gave me assessment on my work and helped me improving along the way. Special thanks to the members of TRAIL and TLO, for their support and regular scientific inspiration and feedback.

I also want to thank my parents who always supported me since childhood. I would like to thank my sister, Enikő, for her listening ears, encouragment and helpful advice. My daughter Nara, who brought tremendous happiness to my life and made the final steps of PhD more enjoyable. And last but not least, I want to thank my husband, Zoltán, for always being there for me.