Essays in Decision Theory and Information Design

Alexis Ghersengorin

UNIVERSITE PARIS I PANTHÉON SORBONNE
UFR d'Economie
Laboratoire de rattachement : PJSE

THÈSE
Pour l'obtention du titre de Docteur en Economie
Présentée et soutenue publiquement
le 7 décembre 2022 par

Alexis GHERSENGORIN

Essays in Decision Theory and Information Design

Sous la direction de M. Jean-Marc TALLON
Professeur, Directeur de Recherche CNRS
Sous la co-direction de M. Franz Dietrich
Professeur à PSE, Directeur de Recherche CNRS

Membre du Jury
M. Francis Bloch, Examinateur, Université Paris 1 Panthéon-Sorbonne
M. Eduardo Perez-Richet, Président, Sciences-Po Paris
Mme. Paola Manzini, Rapporteuse, University of Bristol
M. Pietro Ortoleva, Rapporteur, Princeton University

Essays in Decision Theory and Information Design

Doctoral Thesis in Economics

Alexis Ghersengorin

Paris School of Economics — Université Paris 1 Panthéon-Sorbonne

Remerciements

La thèse est certes un périple quelque peu solitaire mais qui ne saurait être mené à son terme sans toutes les personnes qui nous accompagnent une partie du voyage, celles qui nous indiquent parfois un chemin à suivre ou simplement celles qui apportent leur soutien.

Je commencerai par remercier mon directeur de thèse, Jean-Marc Tallon, qui a su trouver le bon équilibre entre me laisser vagabonder au travers de mes idées souvent farfelues et m'aiguiller sur des sentiers plus balisés lorsque je risquais de m'égarer. Si son esprit (très) critique m'a poussé à interroger avec modestie mes recherches et à calmer mon ardeur parfois naïve de jeune chercheur, j'ai toujours trouvé en lui une oreille attentive lors des moments de désenchantement et de doute. Je remercie également Franz Dietrich, devenu au cours de ma thèse mon co-directeur et avec qui les discussions ont profondément étendu ma compréhension des fondements philosophiques de la théorie du choix; son inexhaustible enthousiasme a plusieurs fois su raviver ma curiosité et mon envie. Plusieurs autres professeurs m'ont apporté leur aide au cours de ces années, parmi lesquels tout particulièrement: Frederic Koessler, qui m'a initié au monde des jeux de communication et du design de l'information; et Eduardo Perez-Richet, qui a suivi mes développements dans ce domaine et m'a gentiment accueilli à Sciences-Po pour mes quatrième et cinquième années. Je suis reconnaissant à Francis Bloch, Paola Manzini et Pietro Ortoleva pour avoir généreusement accepté d'être membres de mon jury de thèse et pour leurs commentaires sur un premier manuscrit.

Je ne saurais exprimer avec assez de vigueur toute ma gratitude pour mes co-auteurs, cette thèse leur appartient également. Au premier rang de ceux-ci se trouve Niels, fidèle compagnon de route, avec qui je partage nombre de passions intellectuelles, pour le meilleur et pour le pire. Si notre collaboration avec Simon fut

certes tumultueuse, elle fut également source d'une remise en question permanente de chacune de nos idées et notre projet commun s'en est trouvé grandi. A Victor et Daniel, qui m'ont offert l'hospitalité des combles et auprès desquels j'ai trouvé des interlocuteurs pour parler autant d'idées saugrenues que de jazz: le premier pour sa (quasi) obsession pour la persuasion bayésienne et les théorèmes loufoques; le second pour son amour des énigmes et des blagues douteuses.

A Thomas, mon compagnon de doute avec qui nos chemins ne cessent de se croiser. A Yagan et Eustache, pour amener nos questionnements d'économistes de Montrouge jusqu'aux pieds des parois. A Etienne et Irène, colocataires de la première heure du bureau R6-58. A Philippe, Giacomo et Laure avec qui je partage un goût quelque peu passé de mode pour la théorie. A tous.tes les doctorant.es des combles de Sciences-Po pour leur accueil chaleureux.

Rien de cela n'aurait été possible sans l'aide de Véronique Guillotin puis Roxana Ban qui ont su écouter, guider et réprimander avec bienveillance la génération de thésard.es de PSE à laquelle j'appartiens.

Enfin, une thèse finit malgré nous par imprégner notre vie au-delà des frontières floues de la recherche. Ma reconnaissance va donc envers les personnes qui m'ont accompagné durant ces années et ont enduré mes élucubrations, mes joies et mes tristesses. A Coline, pour sa patience, son enthousiasme et son dévoument. A ma famille, qui ne comprend pas toujours ce que je fais, mais au fond, qu'importe. A mes colocataires, Arthur, Marion, Greg, Liv et Mae, qui sont bien plus que ça. A tous.tes mes ami.es, qui sauront se reconnaître.

Contents

Introduction **6**

1 Revealed Deliberate Preference Change **16**

1.1 Introduction 17

1.2 Deliberate Preference Change 22

1.2.1 Preliminaries 22

1.2.2 Revealed Relevant Attributes 23

1.2.3 Principle of Sufficient Reason 25

1.2.4 Principle of Deliberation 27

1.2.5 The Representation 28

1.2.6 Transitive Attribute Ordering 33

1.2.7 Identification of the Revealed Relevant Attributes 33

1.3 Motivated Preference Change 35

1.4 An Application 37

2 Note on the Identification of Deliberate Preference Change **41**

3 Grabbing the Forbidden Fruit: Restriction Sensitive Choice **49**

3.1 Introduction 50

3.2 The Model 55

3.2.1 Preliminaries 55

3.2.2 Restriction Sensitive Choice . . . 56
3.2.3 Interpretations . . . 58
3.3 Identification and Welfare . . . 60
3.3.1 Revealed Reaction to Restriction . . . 60
3.3.2 Identification of Restriction Sensitive Choice . . . 60
3.3.3 Welfare . . . 64
3.4 Characterization . . . 65
3.5 Measuring Freedom . . . 71
3.6 Applications . . . 75
3.6.1 Conspiracy Theories . . . 75
3.6.2 Integration Policy Backlash . . . 78
3.6.3 Optimal Delegation and Reactance . . . 82

4 Price Discrimination with Redistributive Concerns 86
4.1 Introduction . . . 87
4.2 Model . . . 91
4.3 Three-Value Case . . . 95
4.4 Optimal Segmentations . . . 104
4.4.1 General Properties . . . 104
4.4.2 Strongly Redistributive Social Preferences . . . 105
4.4.3 Optimal Segmentations and Informational Rents . . . 108

Appendices 112

Appendix A Proofs of Chapter 1 113
A.1 Proof of Proposition 1 . . . 113
A.2 Proof of Theorem 1 . . . 114
A.3 Proof of Proposition 2 . . . 116
A.4 Proof of Proposition 3 . . . 116

A.5 Proof of Proposition 4 117
A.6 Proof of Proposition 6 119
A.7 Proof of Theorem 2 120
A.8 Proof of Theorem 3 120

Appendix B Proof of Theorem 4 **124**

Appendix C Proofs of Chapter 3 **132**
C.1 Proofs of Section 3.3 132
C.2 Proof of Theorem 5 134
C.3 Proof of Theorem 6 148
C.4 Proofs of Section 3.6 149

Appendix D Proofs of Chapter 4 **153**
D.1 Proof of Section 4.4.1 153
D.2 Proofs of Section 4.4.2. 155
D.3 Proofs of Section 4.4.3. 157

Bibliography **159**

Résumé en français **172**

Introduction

As one can deduce from the nebulous title of this dissertation, trying to find a common thread to the various issues addressed therein would certainly be, if not vain, at least artificial. The chapters, however, follow the chronological order of my questioning and the evolution of my interests regarding economics. This leaves a chance to unify the slightest bit this writing, which is the goal I aim to pursue in this introduction. I first give a brief history of the study of individual decision-making in economics. A particular focus is given to the questions that led to the formulation of revealed preference theory (hereafter RPT). As we shall see, my first projects (Chapters 1, 2 and 3) are related to the two main objectives pursued by RPT: (1) finding an observable counterpart to each theoretical concept used in models of individual choice and (2) making these models falsifiable. Chapters 1 and 2 were motivated by the following epistemological interrogation: can a reasonable model of preference change be falsifiable? Chapter 3 also lies within the realm of RPT but tackles a more practical issue. It was initiated by a concern to integrate considerations about freedom in economic modeling of individual choice. As I found my doubts regarding foundations of microeconomic theory less and less justified, my interest towards more concrete economic questions grew, leading to my third project (Chapter 4), which stands apart as it does not lie within the field of decision theory but falls into a domain that recently flourished in economic theory: information design. It studies the distributive effect of price discrimination based

on individual characteristics.

From utility maximization to revealed preference theory. Economics used to arouse mixed feelings in me; for, a social science that uses mathematics as its main tool and language to study human decisions and interactions was a source of both fascination and doubt. Hence, the beginning of my PhD was devoted to better understand the foundations of modern economic analysis. At the core of these foundations lies the modeling of individual decision-making as maximizing a utility function over any possible set of feasible options (for instance, in order to choose a consumption basket given a budget constraint). If information is incomplete, this utility is associated with a probability such that the individual now seeks to maximize the expectation over the utility derived from each possible outcome. The study of individual decisions became an independent sub-field of economics over the years, commonly referred to as decision (or choice) theory.

Under the influence of philosophical positivism (Carnap, 1923) and Popper (1934)'s notion of falsifiability, economists wanted: (1) to relate their theoretical concepts, such as utility and probability, to observable behaviors; and (2) to make their theories falsifiable.[1] To that end, choice theory first replaced utility (or pleasure) with the notion of preferences. Although preferences are more tangible than utility (as it required only to rank options, not to give a quantified measure of pleasure), they are not directly observable. RPT, as inaugurated by Samuelson (1938), completed the positivist program of choice theory by giving a "behavioural understanding of preferences, whereby [it] equates preference with actual or hypothetical choice behaviour" (Thoma, 2021). Thus, revealed preference theorists enunciate consistency conditions on choice behaviors (*axioms*) that are necessary and sufficient to represent an agent's choice *as if* they were maximizing a prefer-

[1] For the positivist legacy in economics, see Clarke (2016); Guala (2019); Gilboa, Postlewaite, Samuelson, and Schmeidler (2019).

ence with some specific characteristics (e.g., transitivity). Therefore, RPT kills two birds with one stone: (1) it merely derives preferences from observable choices and (2) provides a framework to formally state falsifiability conditions for models of individual decisions.

The first of these two objectives fulfilled by RPT, i.e., the behavioural foundation of preferences, is often referred to as "behaviourism" in the philosophical literature and has been subject to a widespread critic, giving birth to the antagonistic position: mentalism.[2] This debate is structured around the question whether economists should explicitly appeal to conative and cognitive mental states to rationalize an agent's behavior or only found their analysis on observables. Without entering into the details of this controversy, let us simply say that, in line with some recent developments (Guala, 2019; Thoma, 2021), we do not share the view that embracing RPT entails to hold the radical behaviourist posture commonly criticized.[3] As a support of this claim, it seems that recent works in RPT frequently give possible psychological interpretations of their choice models.

Under the impulse of experimental and behavioral economics, RPT recently experienced a vivid expansion. Given the increasing evidence of violations of the canonical model of choice (i.e., the maximization of a complete and transitive ordering over options), new sophisticated choice procedures accommodating these violations were given a choice axiomatic foundation.[4] Though this is not always the case, these models are often supported by references to psychological explanations. The first three chapters of this thesis fall within this strand of the literature. As a consequence, to me, the main contribution of adopting the framework of RPT in these works has to do with the second objective fulfilled by RPT: the falsifiability of

[2]For references and a summary of the debate between mentalism and behaviourism in economics, see, among others, Clarke (2016); Dietrich and List (2016a); Guala (2019); Thoma (2021).

[3]I also emphasize that this does not necessarily imply that I fully accept the position of mentalism. I rather consider my work as being somewhat in-between.

[4]Many references of this literature can be found in Chapters 1 and 3.

individual choice models. This epistemological concern is what initially motivated the project of Chapters 1 and 2. Namely, economic models incorporating preference changes have generally been criticized for their lack of empirical content: in short, any social phenomenon could be explained by assuming *ad hoc* changes of preferences.[5] These chapters aim to bridge this gap by proposing a falsifiable model of preference change.

We shall add that the revealed preference method serves another goal, as it is noticeable in most recent works: it allows to determine whether different psychological explanations or theories can be behaviourally distinguished. At least, it offers a language to discuss the behavioural implications of different theoretical concepts inherited from psychology or philosophy. Chapter 3 gives an illustration of this: we propose a model that rationalizes possible reactions to restrictions on the set of opportunities of an agent and argue that this model is compatible with several psychological explanations.

Before turning to the more detailed introduction of each chapter, we wish to highlight another common goal of the axiomatic approach in choice theory. Although it was not explicitly among the initial objectives of RPT, the axiomatic approach (which, beyond RPT, is the main method in decision theory) is often used to make normative analysis. This incidentally explains why the canonical model of RPT is sometimes considered as the main formulation of rational choice theory. In this regard, the axioms "can help the economist (or decision theorist) to convince the people she addresses that they would indeed like to follow her recommendation, or can call attention to weaknesses of a model" (Gilboa et al., 2019). In line with this research agenda, we argue that our model of preference change in Chapters 1 and 2, not only is falsifiable, but is founded on two normative principles that are translated into axioms.

[5] See Grune-Yanoff and Hansson (2009) for a review of the reasons advanced against the integration of preference change in social sciences.

Falsifiability and preference change (Chapters 1 and 2). As I pointed earlier, central among the reasons why economists have long been reluctant to invoke preference change in their analysis is the lack of falsifiable foundations. Historically, economists have contented themselves with explaining changes of behavior, either by changes in the individuals' constraints (e.g., their budget constraints), or by the processing of new information, that is, individuals update their knowledge (or beliefs) about their environment and adapt their behavior consequently. Nevertheless, a wide range of phenomena seem better explained by preference changes because they involve values such as fairness, conservatism, etc. For instance, the expansion of abortion rights in western societies (or their current reassessment in some countries) is more plausibly due to the diffusion (or the questioning) of values such as women's rights than to changing beliefs on some underlying state of the world.

Related to the difficulty of obtaining a falsifiable model of preference change is the lack of apparent normative foundations (e.g., compared to Bayesian updating for beliefs). The challenge is therefore twofold: finding an empirically testable and normatively compelling model of preference change. Chapter 1, jointly written with Niels Boissonnet and Simon Gleyze, is a humble attempt to take up this challenge. This entails to narrow down the class of preference changes we want to describe. For that purpose, we propose a model of deliberate preference change that is identifiable from the observation of successive choices and characterized by two falsifiable normative principles: the *principle of sufficient reasons* and the *principle of deliberation*.

Our setting is the following: there is a finite set of options, each of which is defined by observable attributes, and an outside analyst observes a time-indexed sequence of choices made by a decision maker (DM) over these options (at each period, the DM makes multiple choices, which allows the analyst to retrieve

a complete preference over the options). By observing choices and the object's attributes we can deduce which of them are *relevant* to the agent's choice at each period. Intuitively, if two options that only differ on one attribute are not ranked indifferently, this attribute must be taken into account by the DM for their choice. We can thus keep track of the sequence of *relevant attributes* at each period.

Our first axiom, the principle of sufficient reasons, states that the DM should rank similarly, within and between periods, options that have the same relevant attributes. From this, we obtain that the DM changes their preferences if and only if it can be justified by an attribute that is made relevant or irrelevant. For instance, if an employer becomes aware that their hiring decision is based on the attribute "gender", they might make this attribute irrelevant in the future to stop being discriminatory.[6] In other words, preferences are determined by the set of attributes that are relevant for the DM's choices. This is in line with our view of the attributes as being a material translation of the notion of *reasons* used in philosophy.[7]

The principle of deliberation states that the DM should not make mistakes (from their perspective) when changing preferences; that is, they cannot change their mind twice regarding an attribute if no additional event occurred meanwhile. Otherwise, this would indicate that they fail to deliberate and lack internal consistency.

Theses two axioms are shown to be equivalent to a procedure that we name *deliberate preference change*: (i) preferences at each period are rationalized by the relevant attributes together with a time-independent ordering on the alternatives' attributes; (ii) preference changes are induced by the *awareness* of new attributes and a deliberation about which set of relevant attributes should be adopted for

[6] *Implicit* discrimination would also imply that the attribute "gender" is *relevant*. Therefore, an attribute can be relevant even if the DM does not consciously use this attribute.

[7] See Dietrich and List (2013) for a more detailed discussion on founding the concept of preferences on reasons and a review of the related philosophical literature.

the future periods—this is rationalized by the maximization of an ordering on preferences themselves, the *meta-preference*.

Our interpretation is the following: whenever the DM becomes *aware* of an attribute—through education, social interactions, medias or introspection—they can decide to make it relevant or irrelevant for the next period, inducing a preference change. These changes are made deliberately and therefore are consistent across time; this may result from the DM's moral values, motivated reasoning, social objectives, norms, etc.

Chapter 2, jointly written with Niels Boissonnet, simply tackles an indeterminacy problem that is left aside in the first chapter. It characterizes a more general version of deliberate preference change where other possible sequences of relevant attributes are allowed.

Choices responsive to restrictions (Chapter 3). When being denied the availability of some opportunities, it has been observed in different contexts that people's desire may be redirected toward the unavailable alternatives or their substitutes. This phenomenon is commonly referred to as the *forbidden fruit effect*, an allusion to the famous biblical episode of the Genesis (Levesque, 2018). Chapter 3, written in collaboration with Niels Boissonnet, proposes a model of choice, together with its characterizing conditions on choice behavior, that rationalizes the forbidden fruit effect.

The two prominent explanations of the forbidden fruit effect in psychology have been *reactance theory* and *commodity theory*, both of which are consistent with our model.[8] Reactance relates people's reaction to restriction or prohibition to their attitudes toward freedom of behavior. When a restriction is the source of a threat on their freedom of behavior, they experience an unpleasant emotional

[8] See Rosenberg and Siegel (2018) for a review on psychological reactance theory; Lynn (1991) for a review on commodity theory.

state that motivate them to *restore* this lost freedom, what psychologists have called reactance (Brehm, 1966). The explanation of the forbidden fruit effect by commodity theory has a more 'hedonistic' flavour (Brock, 1968). According to this theory, the more a commodity is perceived as unavailable or requiring much effort to be obtained, the more it will be valued, thereby increasing the desire of agents for this option.

We use the typical framework of RPT; namely, we observe the choices made by a DM in each possible menu formed from a finite set of options. In this setting, the forbidden fruit effect manifests through changes in the choice following the removal of an apparently irrelevant option: e.g., z is chosen in $\{x, y, z\}$, but once y is removed, x is chosen in $\{x, z\}$. The deprivation of y steers the DM's desire toward a potential substitute (x). Formally, studying reactions to restrictions amounts to investigating violations of the "Independence of Irrelevant Alternatives" (IIA) (Chernoff, 1954; Sen, 1971, property α) triggered by the *removal* of opportunities.

In addition to the axiomatic characterization of our choice model, we show how its ingredients can be identified from the observed reactions to restrictions. We explore what welfare judgment can be drawn from choices and demonstrate the difference between our criterion and the prominent ones in the literature. We also derive a measure of the freedom of choice offered by the different possible sets of opportunities faced by an agent whose final choices are responsive to restrictions.

We finally study three applications of our model. We first show that it can accommodate the emergence of conspiracy theories and the backlash of integration policy targeted toward minorities—two phenomena that have often been associated with reactance. We next study a principal's problem who delegates decision to a misaligned but better-informed agent whose choices follow our procedure. We find that the effect on the agent's welfare is ambiguous.

The distributive effects of price discrimination (Chapter 4). Chapter 4, a joint project with Daniel Barreto and Victor Augias, stands apart in this dissertation as it is the only chapter that does not lie in the field of decision theory. This is an applied information design work (see Bergemann and Morris, 2019). Its aim is to provide a normative analysis of the distributive effects of personalized pricing, a practice that has attracted a growing interest from regulators given the increasing amount of consumption data collected on the internet. Indeed, consumers are continuously leaving traces of their identities on the internet, be it through social media activity, search-engine utilization, online-purchasing and so on. This consumer data is highly valued by the actors of the digital economy.[9] A practice of particular regulatory interest is that of charging personalized prices to different consumers based on their estimated willingness to pay for products.

As shown by Bergemann, Brooks, and Morris (2015), not only can personalized pricing be used to increase economic surplus—by implementing prices that allow every consumer to buy—, but it can also be performed in a way that ensures that all the created surplus accrues to consumers. However, an important aspect of personalized pricing that remains overlooked by the literature is its distributive effect: since different consumers pay different prices, this practice defines how surplus is distributed *among* consumers, raising questions about how it can benefit poorer consumers relative to richer ones.

Our aim is to study how personalized pricing impacts different consumers and how it should be performed under the objective of increasing consumer welfare while prioritizing poorer consumers. The latter is captured through Pareto weights that are greater for poorer consumers. The relative richness of consumers is identified from their willingness to pay under the simple assumption that individuals

[9]This can be illustrated by the rapid ascension of the French digital analytics unicorn Contentsquare, which raised over $1.1 billion in investment funding between May 2021 and June 2022 and whose services allow firms to tailor decisions such as pricing and advertising specifically to different consumers.

who are ready to pay higher prices are on average richer. Our results draw qualitative characteristics of the price discrimination policies that achieve this goal. Importantly, our analysis also shows that the prioritization of poorer consumers can be inconsistent with the maximization of total consumer surplus: raising the surplus of poorer consumers may only be possible while granting additional profits to the producer and sacrificing surplus made by the wealthy consumers. We characterize the markets for which this is the case and give a procedure to construct optimal segmentations given a strong redistributive objective. For the remaining markets, we show that the optimal segmentation is surprisingly simple: it generates one segment with a discount price and one segment with the same price that would be charged under no segmentation.

Chapter 1

Revealed Deliberate Preference Change

Joint with Niels Boissonnet (Bielefeld University) and Simon Gleyze (Uber).

1.1 Introduction

Understanding how individuals change their behavior is critical for social sciences. Economists traditionally argue that decision makers (DMs) are Bayesian; that is, they adapt their behavior by updating their beliefs about the environment. Although this mechanism has proved powerful and normatively appealing, a wide range of phenomena seem better described with preference change because they involve values such as fairness, conservatism, etc. For instance, Barrera, Guriev, Henry, and Zhuravskaya (2020) show experimentally that exposure to fake news about the European refugee crisis increases voting intentions toward far-right politicians, even though voters' beliefs may not change in case of fact-checking. Their explanation is that by raising voters' awareness of the migration issue, politicians may alter preferences. Another example is the expansion of abortion rights in western societies—along with its economical and political implications—that is more plausibly due to the diffusion of new values such as women's rights than to changing beliefs on some underlying state of the world.

Modeling preference changes raises two challenges: first, the lack of normative foundations compared to Bayesian updating; second the lack of testability of the model. To fill these gaps, we propose and axiomatize two testable normative principles: a *principle of sufficient reason* and a *principle of deliberation*. To express these normative principles, we use the attribute-based approach. Our primitive is the observation of successive preferences, as well as the attributes of each alternative. This allows us to define the attributes that are *relevant* to DM's choice at each period and, thereby, to reveal DM's reasoning behind preference changes. In doing so, we make progress toward a testable and normatively founded model of preference change.

The principle of sufficient reason states that DM changes her preferences if and

only if it can be justified by an attribute of the alternative that is made relevant or irrelevant. For instance, if an employer becomes aware that her hiring decision is based on the attribute "gender", she might make this attribute irrelevant in the future to stop being discriminatory.[1] Formally, this translates into an identification axiom called Restricted Reversals, which guarantees that preference reversals can be explained by changes of relevant attributes alone (proposition 1).

The principle of deliberation states that DM should not make mistakes (from her perspective) when changing preferences; that is, she cannot change her mind twice regarding an attribute if no additional event occurred meanwhile. Otherwise, this would indicate that she fails to deliberate and lacks internal consistency. Formally, this translates into an Acyclicity axiom, which guarantees that if an attribute is made relevant and then irrelevant it must be explained by *other* attributes becoming (ir)relevant meanwhile.

Our main representation theorem states that Restricted Reversals and Acyclicity hold if and only if (i) preferences are represented by an ordering on the alternatives' attributes—we call this the *attribute ordering*—, and (ii) preference changes are explained by the maximization of an ordering on preferences themselves—we call this the *meta-preference* (theorem 1).

Preference changes take the following form: whenever DM becomes *aware* of an attribute—through education, social interactions, medias or introspection—she can decide to make it relevant or irrelevant for the next period, inducing a preference change. The succession of such changes is consistent with the maximization of a meta-preference relation, capturing DM's moral values, motivated reasoning, social objectives, norms, etc. Therefore, the reasoning behind preference changes is revealed through the meta-preference relation and the sequence of awareness. Such a sequence represents DM's constraint regarding which preferences are

[1] *Implicit* discrimination would also imply that the attribute "gender" is *relevant*. Therefore, an attribute can be relevant even if DM does not consciously use this attribute.

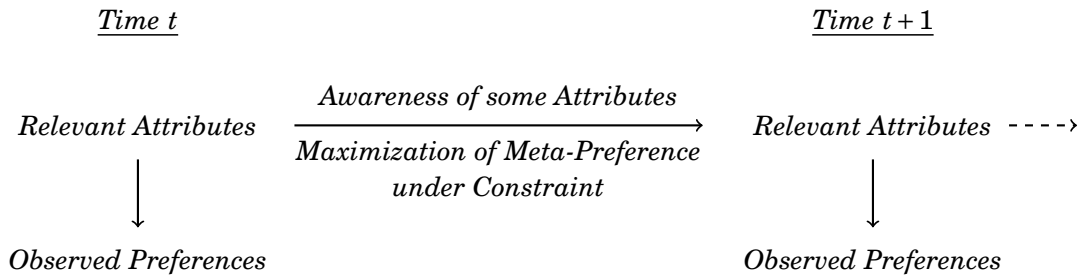

Figure 1.1: *The Dynamics of Deliberate Preference Change.*

reachable at each period. The existence of such a constraint follows from the principle of deliberation and the observation of multiple preference reversals. Indeed, should DM be unconstrained in the maximization of her meta-preference she would directly reach her most preferred set of relevant attributes and never change preferences again. Note that the attribute ordering remains stable and that only the set of relevant attributes changes; this implies that if DM deems relevant the same set of attributes from one period to another, she must make exactly the same choices.[2] See Figure 1.1 for a representation of the model. Models of chosen preferences are receiving renewed attention since Bernheim, Braghieri, Martínez-Marquina, and Zuckerman (2021), and the present paper is the first to investigate its revealed preference implications.

The attribute ordering and the meta-preference are essentially unique.[3] Furthermore, if two sequences of awareness represent DM's constraint on meta-choices, so does their intersection (proposition 2). We, however, stress that the sequence of relevant attributes is not uniquely identified in general. Hence, we investigate specific conditions that make this sequence set-identified or point-identified (proposition 6 and theorem 2).

[2] We discuss why it would be problematic that DM changes her "taste" towards the attributes in section 1.2.5.

[3] That is, if two distinct attribute orderings (resp. meta-preferences) rationalize the preference changes, any completion of their intersection do so.

We then investigate a particular type of meta-preference (i.e., a particular type of reasoning) in which DM chooses the preferences that maximize her underlying utility (theorem 3). This captures *motivated preference change* in which DM's evaluation of the attributes is guided by her own-interest alone. We show that motivated preference change provides new insights on the formation of political preferences. For instance, if two voters with identical preferences become aware of the same attributes in a different order, they can end up endorsing antagonistic views. Whether a voter becomes aware that a politician is corrupted before or after learning his political affiliation can lead to very different outcomes: in the latter case, the voter might ignore this attribute because it undermines the view of her preferred candidate. This type of path-dependent motivated reasoning is specific to our model and provides empirically testable implications.

Our contribution is threefold: first, we show that models incorporating preference changes can have empirical content and normative foundations. Second, our model suggests that choice reversals need not be irrational, and may reflect DM aligning her choice behavior with her values. Any deliberate preference change must break (or create) indifference with respect to other pairs of alternatives that share the same attribute, which indicates that this attribute becomes relevant (resp. irrelevant). This is a necessary condition for preference change to be induced by a coherent reasoning from DM. Finally, we illustrate the explanatory power of our model through an application.[4]

Related Literature. The idea of representing objects by their attributes goes back to Lancaster (1966). Moreover, we draw on an important literature on reason-based theories of choice, most notably Simonson (1989), Shafir, Simonson, and Tversky (1993), Tversky and Simonson (1993), and Dietrich and List (2013, 2016b).

[4]All proofs until section 1.2.6 are in the Appendix. The remaining proofs can be found in the Supplement (Boissonnet, Ghersengorin, and Gleyze, 2022b).

Boissonnet (2019) and Dietrich and List (2011) also use an attribute-based approach to model non-informational preference change. Our paper should be seen as the first counterpart of these models within the revealed preference theory.

There is an important literature on "changing tastes" understood as time inconsistency. Strotz (1955) is the first to uncover the problem of consistent planning and to investigate how should individuals with non-exponential discounting make dynamically consistent choices. Gul and Pesendorfer (2001, 2005) and Dekel, Lipman, and Rustichini (2009) provide behavioral foundations of preferences for commitment, namely choosing a smaller choice set for one's future self to avoid temptation. The main differences with our paper is that they consider deviations between expected behavior and actual behavior which are typically *not deliberate* (inconsistent) from the point of view of past selves. Instead, we look at preference changes that are deliberate but completely myopic, meaning that DM is unaware that she may change preferences in the future. The closest paper in this literature to our own is Nehring (2006) who studies the revealed preference implications of second-order preferences as a self-control mechanism. The main differences with our paper are that he considers preferences over menus whereas we deal with preferences over alternatives, he does not introduce attributes, and the second order preferences act exclusively as a self-control mechanism whereas our meta-preference relation is completely general.

Our work relates to the literature on conflicting motivations—or justifiable choices—as we also obtain a representation with several (more precisely two) orderings. See among other contributions Kalai, Rubinstein, and Spiegler (2002), Heller (2012), De Clippel and Eliaz (2012), Cherepanov, Feddersen, and Sandroni (2013a), Dietrich and List (2016b) and Ridout (2021). Despite this similarity, these works focus on static choice data that violates the usual rational requirements—namely the Weak Axiom of Revealed Preferences (WARP) or the Independence

of Irrelevant Alternatives Axiom (IIA)—whereas in our work, the choice data consists in an ordered sequence of choices on the same collection of menus of options. We explore two distinct situations, one in which within-period choices are represented by not necessarily transitive binary relations, one in which within-period choices satisfy WARP. We focus on the irregularities in choices that arise between periods, hence the reversals can happen on the same menus. Furthermore, the time structure is used to rationalize the successive changes as being guided by a meta-maximization.

In the applied theory literature, the closest paper is Bernheim et al. (2021). Their model and ours share two important ideas. First, they argue that DM can choose "worldviews" which determine her valuation of future consumption streams. This is related to our concept of relevant attributes. Second, in their model DM is constrained by her "mindset flexibility" when changing worldviews. This echoes our constraint on awareness. For the purpose of falsification, our model makes some simplifications: in their model DM anticipates her preference change, and they allow for convex combinations of worldviews. Despite the differences in modelling assumptions, their paper is complementary with ours as we focus on the identification and falsification of deliberate preference changes. Other models of chosen preferences include Becker and Mulligan (1997), Akerlof and Kranton (2000), Palacios-Huerta and Santos (2004).

1.2 Deliberate Preference Change

1.2.1 Preliminaries

There is finite set X of **alternatives**, that are defined by their **attributes**. Formally, there are K attributes and an alternative is a vector $\mathbf{x} = (x^1, \ldots, x^K)$ in the

vector space $\mathbb{R}^K$ whose k^{th}-coordinate describes the value x^k of the attribute k.[5] For any subset $M \subseteq \{1,\dots,K\}$, denote $\mathbf{x}^M = (x^k)_{k\in M}$ and $\mathbf{x}^{-M} = (x^k)_{k\notin M}$.[6] We require that for any attribute k there exist $\mathbf{x}$ and $\mathbf{y}$ such that $x^k \neq y^k$, as otherwise this attribute could be removed.

The analyst observes (i) the value of each attribute for all alternatives, and (ii) choices over options for T periods of time. The latter are represented by a sequence of complete orders $(\succsim_t)_{t=1,\dots,T}$, where $\succ_t$ and $\sim_t$ denote the asymmetric and symmetric parts, respectively. For the first part of the analysis, we do not require each $\succsim_t$ to be transitive. We investigate the implications of transitivity within periods —that is, DM's choices satisfy WARP— in section 1.2.6.

Example 1: Labor Market Discrimination. *An employer wants to hire a worker. Her decision is based on the resume of each candidate that provides information on three attributes: (1) "education", (2) "experience", and (3) "gender" (1 for female and 2 for male). Therefore, a female college-educated worker entering the labor market is represented by* $\mathbf{x} = (4,0,1)$*, while a male non-educated worker with ten years of experience is represented by* $\mathbf{y} = (0,10,2)$.

1.2.2 Revealed Relevant Attributes

The attribute-based approach allows us to identify which attributes drive DM's choice behavior. These "relevant attributes" are easy to identify when the choice set X is sufficiently rich: the attribute k is revealed relevant at t if there is a pair of alternatives $\mathbf{x}$ and $\mathbf{y}$ that only differ on the k^{th}-dimension ($\mathbf{x}^{-k} = \mathbf{y}^{-k}$) and such that $\mathbf{x} \not\succ_t \mathbf{y}$. In this case, we are sure that DM uses attribute k in her decision

[5] Attributes can either code different categories (e.g colors, sex, etc.), indicate whether a property is possessed by the alternative (e.g whether a job applicant is a foreigner or not), or measure the intensity of a property (e.g years of experience of a candidate).

[6] If $M = \{k\}$ is a singleton, we simply write $\mathbf{x}^{-k}$ instead of $\mathbf{x}^{-\{k\}}$.

making. This richness assumption—that we can always find two alternatives that differ only on one dimension—would be too restrictive, however. We illustrate the construction of the *revealed relevant* attributes using our running example and then provide a formal definition.

Example 1 (continued): *Suppose that* $\mathbf{z} \succ_t \mathbf{x}$ *for the two candidates* $\mathbf{x} = (4,0,1)$, $\mathbf{z} = (4,2,2)$. *The idea is to identify a set of attributes* $M \subset \{1,2,3\}$ *that has to be relevant to explain this strict preference. From* $\mathbf{z} \succ_t \mathbf{x}$, *we can conclude that* $M = \{2,3\}$ *is revealed relevant because (i) the alternatives differ on* M *and are identical outside of* M, *and (ii) there is no pair of alternatives that differ on a strict subset of* M *and are ranked strictly. The second point captures conservatism in our definition of revealed relevant attributes: if we cannot disentangle which attributes drive DM's behavior exactly, we keep all attributes in* M. *The following definition formalizes points (i) and (ii).*

Definition 1 (Revealed Relevant Attributes). *A set* M *of attributes is* ***revealed relevant*** *at period* t *if:*

(i) there exists $\mathbf{x}, \mathbf{y} \in X$ *with* $\mathbf{x}^{-M} = \mathbf{y}^{-M}$ *and* $x^k \neq y^k$ *for every* $k \in M$, *such that* $\mathbf{x} \not\succ_t \mathbf{y}$;

(ii) for every $M' \subsetneq M$ *and every* $\mathbf{w}, \mathbf{z} \in X$ *with* $\mathbf{w}^{-M'} = \mathbf{z}^{-M'}$, $\mathbf{w} \sim_t \mathbf{z}$.

Remark: if two attributes are systematically revealed relevant together, they might be coded into a single attribute (for instance if colors have been coded into different binary attributes).

Let P_t denote the collection of sets of revealed relevant attributes at period t. We denote $\mathbf{m}_t \in \{0,1\}^K$ the **vector of revealed relevant attributes** such that $m_t^k = 1$ if $k \in \bigcup_{M \in P_t} M$ and $m_t^k = 0$ otherwise.[7]

[7] Our definition of revealed relevant attributes is analogous to the definition of a non-null state in expected utility theory (taking the attributes as states and the alternatives as acts).

We emphasize that an attribute can be revealed relevant, yet DM might be unaware that it causes her behavior. For instance, it is well known that implicit discrimination can have a strong impact on job performance (Bertrand, Chugh, and Mullainathan, 2005; Glover, Pallais, and Pariente, 2017; Bertrand and Duflo, 2017).

1.2.3 Principle of Sufficient Reason

We impose the following principle of sufficient reason: DM changes preferences if and only if the revealed relevant attributes change. The interpretation is that DM does not "wake up" with different preferences but must be able to justify her new preferences by making some attributes relevant or irrelevant. We view this as a normative principle: unjustified changes would not be normatively compelling.

Formally, the axiom states that if two alternatives $\mathbf{x}$ and $\mathbf{x}'$ have the same relevant attributes between periods t and t'—namely, if $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ where $\circ$ denotes the element-wise (Hadamard) product—DM should rank consistenly $\mathbf{x}$ against the other alternatives in period t and $\mathbf{x}'$ against the other alternatives in period t'.

Restricted Reversals. *Preferences $(\succsim_t)_t$ satisfy Restricted Reversals if for any t, t', and for any $\mathbf{x}, \mathbf{x}', \mathbf{y}, \mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'}$,*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_{t'} \mathbf{y}'.$$

Remark: although we do not impose restrictions on the values of the attributes, the value 0 *has a specific role. If no attribute can take the value* 0*, this axiom can simply be stated as: if $\mathbf{m}_t = \mathbf{m}_{t'}$, then $\succsim_t = \succsim_{t'}$. This would imply that DM changes her evaluation towards every option each time an attribute is made (ir)relevant. We*

stress that, although it makes sense to remove 0 *for some attributes (e.g a category attribute coding the color of an item), it is not necessarily the case for attributes measuring the intensity of a property (e.g years of experience) or binary attributes that indicate whether a property is possessed or not (e.g whether a job applicant is a foreigner or not). This axiom therefore imposes that some alternatives be ranked similarly in different periods, even though the revealed relevant attributes might change.*

Example 1 (continued). *Consider four candidates* $\mathbf{x} = (6,2,1), \mathbf{x}' = (0,2,1)$, $\mathbf{y} = (5,0,2)$ *and* $\mathbf{y}' = (0,0,1)$. *Suppose that the only strict rankings of* $\succsim_1$ *are* $\mathbf{x} \succ_1 \mathbf{x}' \succ_1 \mathbf{y}'$ *whereas the only strict ranking of* $\succsim_2$ *is* $\mathbf{x}' \succ_2 \mathbf{y}'$. *It is verified that the vectors of revealed relevant attributes are* $\mathbf{m}_1 = (1,1,0)$ *and* $\mathbf{m}_2 = (0,1,0)$ *respectively. Observe that* $\mathbf{x}' \circ \mathbf{m}_1 = \mathbf{x} \circ \mathbf{m}_2$, *hence* $\mathbf{x}$ *and* $\mathbf{x}'$ *have the same relevant attributes at periods 1 and 2. Similarly,* $\mathbf{y}' \circ \mathbf{m}_1 = \mathbf{y} \circ \mathbf{m}_2$. *Therefore, this sequence of choices violate Restricted Reversals, given that* $\mathbf{x}' \succ_1 \mathbf{y}'$ *whereas* $\mathbf{x} \sim_2 \mathbf{y}$.

A consequence of this axiom is the existence of a bijection between vectors of revealed relevant attributes and preference relations. Namely, this axiom is necessary and sufficient to represent the sequence of preferences $(\succsim_t)_t$ by the sequence of revealed relevant attributes $(\mathbf{m}_t)_t$ together with a time-independent binary relation over vectors of attributes. Formally, for any period t, let $X(\mathbf{m}_t) = \{\mathbf{x} \circ \mathbf{m}_t : \mathbf{x} \in X\}$ be the set of alternatives "filtered" through the revealed relevant attributes $\mathbf{m}_t$, and denote $\bar{X} = \bigcup_t X(\mathbf{m}_t)$.

Proposition 1. *Preferences* $(\succsim_t)_t$ *satisfy Restricted Reversals if and only if there exists a complete binary relation* $\geqslant \subseteq \bar{X}^2$ *(called the* ***attribute ordering****), such that for any period* t *and any* $\mathbf{x}, \mathbf{y} \in X$*:*

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t. \tag{1.1}$$

The interpretation is that DM has a fundamental preference—called the *attribute ordering*—that, unlike her choices $(\succsim_t)_t$, does not change over time. This attribute ordering ranks vectors of attributes and does not depend on the relevant attributes.[8] The main consequence of Proposition 1 is that preference change can only be induced by changes in relevant attributes. Observe that the attribute ordering need not be transitive. We derive necessary and sufficient conditions for a transitive attribute ordering in Section 1.2.6.

1.2.4 Principle of Deliberation

The second normative principle that guides our analysis is a principle of deliberation: DM must evaluate all possible preferences at time t and consistently choose the best feasible one according to some criterion. This translates into an acyclicity axiom, which states that if DM changes her preference once, every future change should be due to the discovery of some new attributes—i.e that were not involved in the first change—and towards which DM has changed her attitude meanwhile.

ACYCLICITY. *Preferences $(\succsim_t)_t$ satisfy Acyclicity if for any t and any $t' > t+1$, if $\mathbf{m}_{t+1} \neq \mathbf{m}_{t'}$, then there exists k such that $m^k_{t'} \neq m^k_{t+1} = m^k_t$.*

Note that, as soon as several choice reversals are observed, the principle of deliberation implies the existence of a constraint on preference change. Indeed, would preference change be unconstrained, DM would directly reach her most preferred preference once and for all. We interpret this constraint as DM's awareness: she can change only the attributes she is aware of, that is, the ones she is able to question.

[8] In a slightly different framework, Dietrich and List (2013) provide an equivalence result between this separability condition (their axiom 2) and the existence of an attribute ordering.

Example 1 (continued). *Suppose that* $\mathbf{m}_1 = (0,0,1)$ *and* $\mathbf{m}_2 = (0,1,0)$*, namely the recruiter makes gender relevant but experience irrelevant at the second period. This could be because on the market men are more experienced, implying an unfair discrimination. Therefore, she must have been able to modify her relevant attributes (at least) on these two attributes. Acyclicity implies that she could never choose the following relevant attributes in the future:* $(0,0,0)$ *and* $(0,1,1)$ *as they were accessible between period 1 and period 2. Since she did not change the relevance of the education attribute, we conclude that she was not aware of this attribute at this point. Assuming for instance that education provides a fair criterion to rank the candidates, she could later on decide to remove again gender only if education is made relevant jointly, reaching* $\mathbf{m}_3 = (1,0,0)$*.*

1.2.5 The Representation

The constraint on preference change in the representation is formalized by a sequence of vectors $(\mathbf{a}_t)_{t=1}^{T-1}$, which represents DM's **awareness** between each period t and $t+1$. Namely, $\mathbf{a}_t \subseteq \{0,1\}^K$ for any t and codes as 1 attributes that DM can modify and as 0 the ones that she cannot modify between t and $t+1$. An awareness vector $\mathbf{a} \in \{0,1\}^K$ together with a vector of relevant attributes $\mathbf{m} \in \{0,1\}^K$ defines a set of **reachable attributes** for the next period $R(\mathbf{m},\mathbf{a})$:

$$R(\mathbf{m},\mathbf{a}) = \left\{\mathbf{m}' \in \{0,1\}^K : \text{for all } k, \mathbf{a}^k = 0 \text{ implies } \mathbf{m}'^k = \mathbf{m}^k\right\}.$$

To state our main result, define for any set A and any linear order $P \subset A^2$, $\max(A,P) = \{a \in A \mid aPb,\ \forall b \in A\}$.

Theorem 1 (Representation)**.** *Preferences* $(\succsim_t)_t$ *satisfy Restricted Reversals and Acyclicity if and only if there exists a complete binary relation* $\succcurlyeq \subseteq \bar{X}^2$*, a sequence*

of awareness $(\mathbf{a}_t)_t$ *(with* $\mathbf{a}_t \in \{0,1\}^K$*), and a linear order* $\rhd \subseteq \{0,1\}^K \times \{0,1\}^K$,[9] *such that, for any* t *and any* $\mathbf{x},\mathbf{y} \in X$,

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t, \tag{2.1}$$

$$\{\mathbf{m}_{t+1}\} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \rhd). \tag{1.2}$$

The principle of sufficient reason together with the principle of deliberation are necessary and sufficient for what we name a **deliberate preference change model**. If the tuple $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ satisfy the conditions in theorem 1, we say that it **rationalizes** $(\succsim_t)_t$. In this model, DM's behavior is represented by the maximization of two binary relations: (2.1) a preference relation on alternatives that together with the relevant attributes determine choices in each period and (2.2) a meta-preference relation on vectors of relevant attributes that determine the change of preference between periods. The revealed preference implication of our model is that when we observe choice reversals between alternatives $\mathbf{x}$ and $\mathbf{y}$, we should observe other choice reversals on alternatives that share attributes with $\mathbf{x}$ and $\mathbf{y}$. For instance if an employer stops discriminating at work this should impact her preferences in other contexts, such as her political preferences.

Let us emphasize that our model is complementary with Bayesianism to explain preference change. Even though evidence suggests that agents do not always follow Bayes' rule, we do not think that an exhaustive theory of social interactions could do without belief updating. Instead, we argue that preference change and belief updating can occur simultaneously. This thesis receives empirical support in experiments on fake news (see Barrera et al., 2020, cited in the introduction).

[9] It is observationally equivalent to construct a linear order or a complete preorder together with a tie-breaking rule for the meta-choice such that if $\mathbf{m}_t = \mathbf{m}$ and $\mathbf{m}_{t'} \neq \mathbf{m}$ for some $t' > t$, then $\mathbf{m}_\tau \neq \mathbf{m}$ for all $\tau > t'$.

What is falsified? The fact that attributes can only be made relevant or irrelevant—and that DM cannot change her "taste" (attribute ordering) towards an attribute due to the stability of the attribute ordering—might seem arbitrary at first. Beyond the normative appeal of this principle of sufficient reason, this is also important for the testability of the model, as otherwise almost any sequence of observed choices could be rationalized by changing DM's tastes. Furthermore, if the space of attributes is correctly specified from the beginning, there is no need to change DM's tastes. For instance, if the employer makes "gender" irrelevant to avoid discrimination, but makes it relevant again in the future due to an affirmative action policy, this policy should be thought of an attribute that is complementary with the attribute "gender". Therefore, it is not that DM changes her tastes toward the attribute "gender", but that the combination of "gender" and "affirmative action" is strictly preferred to "gender" alone. This suggests that the specification of the attributes is a crucial step that the researcher should discuss carefully, and commit to before observing choice data to avoid ex-post rationalization.

The latter point naturally leads to a worry that our axiomatization may not offer a genuine falsification of our model; for, a violation of the axioms can always be interpreted as resulting from an incomplete account of the potential attributes determining the choice of the DM. Said differently, one may (artificially) add attributes and redefine the options integrating these attributes until the axioms are satisfied (possibly by making them vacuous). Two points deserve to be highlighted regarding this critic. First, we view attributes as a translation of the notion of *reasons* used in philosophy, hence they should neither be determined artificially nor be individual specific.[10] In particular, the analyst should commit to the set of attributes, hence this should be done in situations in which she is reasonably confident that all the potentially relevant attributes are known (e.g., in the lab).

[10]See Dietrich and List (2013) for a more detailed discussion on founding the concept of preferences on reasons and the link with the philosophical literature.

Second, a similar criticism can be addressed to most revealed preference theory, as it was already noted by Sen (1973): if you add more and more contextual elements in the descriptions of options (the extreme case being that the same option is considered different from one menu to another), then you can never falsify WARP. While we recognize the importance of this point we also believe that this can be seen as an advantage. After all, it is common in choice theory to use violation of WARP to identify patterns of choice causing this violation. In this case, the model can be used to test whether a set of attributes appropriately account for DM's preference changes. The analyst, rather than committing to the attributes would commit to the model. Then, any violation of the axioms signals that another attribute must be found to rationalize the observed choice.

What can be further inferred if one of the two axioms is violated? First, a violation of Restricted Reversal indicates that preference changes do not arise from changes in DM's revealed relevant attributes. Indeed, it is a necessary and sufficient condition for the existence of a time-independent attribute ordering that rationalizes each period's preference together with a set of relevant attributes (proposition 1). Therefore, the analyst's knowledge of what determines DM's preference is incomplete: we may not observe all attributes, or the attribute ordering may change because DM discovers new consequences of an attribute for instance. Second, a violation of Acyclicity suggests that DM does not change her preferences *rationally*, meaning that no linear order can rationalize the sequence of meta choices. Canonical examples of non-deliberate preference changes are nudges, conformism or random utility. Alternatively, a violation of these axioms may suggest that the revealed relevant attributes are not the "truly" relevant attributes for DM, and her behavior could be rationalized by our model with a different sequence $(\mathbf{m}'_t)_t$.[11]

[11]Note that if one does not want to restrict attention to revealed relevant attributes, it is possible to write axioms on multiple "candidate" sequences of relevant attributes (this axiomatization of the

Uniqueness. Without further restrictions, only the attribute ordering is uniquely revealed and the meta-preference identified up to an arbitrary completion. Furthermore, any intersection of two rationalizing sequences of awareness can also rationalize preference changes.

Proposition 2 (Uniqueness)**.** *Let* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *and* $(\geqslant', \rhd', \mathbf{m}_t, \mathbf{a}'_t)$ *rationalize* $(\succsim_t)_t$*. Then, any completion of* $\geqslant \cap \geqslant'$ *and* $\rhd \cap \rhd'$*, together with* $(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ *also rationalize* $(\succsim_t)_t$*.*

Growing awareness. An implicit assumption or our deliberate preference change model is that DM does not remember the attributes she was aware of in previous periods. One may then naturally ask whan can be said about deliberate preference changes if $(\mathbf{a}_t)_t$ is required to be growing, i.e., if $(\mathbf{a}_t)_t$ is such that for any attribute k, $a_t^k = 1$ implies $a_{t+1}^k = 1$? The following axiom characterizes this situation.

Perfect Recall Acyclicity. *Preferences* $(\succsim_t)_t$ *satisfy Perfect Deliberation if any for any* t *there exists* k *such that for all* $t' < t$*,* $\mathbf{m}_{t'} \neq \mathbf{m}_t \implies m_t^k \neq m_{t'}^k = m_1^k$*.*

Proposition 3. *Preferences* $(\succsim_t)_t$ *satisfy Sufficient Reason and Perfect Deliberation if and only if there exists a deliberate preference change model* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *that rationalizes them and such that* $(\mathbf{a}_t)_t$ *is growing.*

An important consequence of imposing that $(\mathbf{a}_t)_t$ be growing is that preference changes will not exhibit path dependence: becoming aware of $(\mathbf{a}_t)_t$ sequentially or simultaneously ultimately lead to the same vector of revealed relevant attributes. As illustrated in the application, this contrast with the non-growing awareness setting.

more general model can be found in Boissonnet and Ghersengorin, 2022).

1.2.6 Transitive Attribute Ordering

Our main representation theorem does not guarantee that the attribute ordering is transitive and does not require that the observed preferences $(\succsim_t)_t$ are transitive. Indeed Restricted Reversals constraints choices only between pairs of periods which is not enough to guarantee transitivity. For instance, suppose that $\mathbf{x},\mathbf{y} \in X(\mathbf{m}_t)$, $\mathbf{y},\mathbf{z} \in X(\mathbf{m}_{t'})$ and $\mathbf{x},\mathbf{z} \in X(\mathbf{m}_{t''})$ but $\mathbf{z} \notin X(\mathbf{m}_t), \mathbf{x} \notin X(\mathbf{m}_{t'})$ and $\mathbf{y} \notin X(\mathbf{m}_{t''})$. It could be that $\mathbf{x} \succ_t \mathbf{y}, \mathbf{y} \succ_{t'} \mathbf{z}$ and $\mathbf{z} \succ_{t''} \mathbf{x}$ because Restricted Reversals does not constraint choices on triplets of periods. In fact, this problem is more general and may arise with any number of periods strictly greater than two.

Transitivity of preferences is sometimes viewed as a condition for rationality, hence it might be of interest to characterize transitivity of the attribute ordering. The following axiom extends Restricted Reversals to address this problem.

STRONG RESTRICTED REVERSALS. *For any* $\{t_1,\dots,t_n\}$ *and any* $\{\mathbf{x}_k,\mathbf{x}'_k\}_{k=1,\dots,n}$ *such that, for* $k=1,\dots,n-1$, $\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}}$ *and* $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1}$, *preferences* $(\succsim_t)_t$ *satisfy Strong Restricted Reversals if:*

$$\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k, \text{ for every } k=1,\dots,n-1 \implies \mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n.$$

Proposition 4. *Suppose that preferences* $(\succsim_t)_t$ *are transitive. Preferences satisfy Strong Restricted Reversals and Acyclicity if and only if there exists a deliberate preference change model* $(\geqslant, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *that rationalizes them with* $\geqslant$ *being a complete preorder.*

1.2.7 Identification of the Revealed Relevant Attributes

The relevant attributes are typically not identified without further restrictions on preferences. This is the case because when we observe an indifference, we cannot

always identify whether this is due to an attribute being irrelevant, or whether DM is indifferent towards this attribute in the attribute ordering. Denote $\mathscr{M}(\succsim_t) = \{\mathbf{m} : \exists \text{ a preorder } \geqslant \text{ s.t. } (\mathbf{m}, \geqslant) \text{ rationalizes } \succsim_t\}$ the set of relevant attributes that rationalize preferences at t using a transitive attribute ordering. To explore the structure of $\mathscr{M}(\succsim_t)$ we make the following richness assumption.

RICHNESS ASSUMPTION. *For all $\mathbf{x}, \mathbf{y} \in X$ that differ only on a subset M of n attributes, there is a sequence of alternatives $\mathbf{z}_1, \ldots, \mathbf{z}_n \in X$ such that $\mathbf{z}_1 = \mathbf{x}$, $\mathbf{z}_n = \mathbf{y}$ and $\mathbf{z}_i^{-k} = \mathbf{z}_{i+1}^{-k}$ for some $k \in M$, for all $i = 1, \ldots, n-1$.*[12]

We show that, under the richness assumption and the transitivity of the preferences $\succsim_t$, the set of vectors of relevant attributes $\mathbf{m}$ that can be used to rationalize preferences in the baseline model has a lattice structure. The most parsimonious vector is the vector of revealed relevant attributes $\mathbf{m}_t$,[13] but in principle other vectors could be used to rationalize DM's preferences.

Proposition 5. *Assume richness and suppose that preferences $\succsim_t$ are transitive. If Restricted Reversal is satisfied, $\mathscr{M}(\succsim_t)$ is a lattice ordered by $\geq$. Its minimum is the vector of revealed relevant attributes $\mathbf{m}_t$ and its maximum is $(1, \ldots, 1)$.*

This indeterminacy problem between irrelevant attributes and indifference can be solved if we impose that indifference are *only* caused by an attribute being irrelevant. In this case, an indifference $\mathbf{x} \sim_t \mathbf{y}$ has a clear interpretation in the sense that there is no attribute that motivates DM to choose $\mathbf{x}$ over $\mathbf{y}$. This is the content of the following axiom.

[12] For this assumption to be satisfied, it might be necessary to regroup certain attributes. For instance, splitting a category attribute (e.g color) into binary attributes will violate this assumption.

[13] If X is not rich, the vector of revealed relevant attributes $\mathbf{m}_t$ need not be the minimum of the lattice.

JUSTIFIED INDIFFERENCE. *Preferences* $(\succsim_t)_t$ *satisfy Justified Indifference if for any* t *and any alternatives* $\mathbf{x}, \mathbf{y} \in X$*,*

$$\mathbf{x} \sim_t \mathbf{y} \implies |\mathbf{x} - \mathbf{y}| \circ \mathbf{m}_t = (0, \ldots, 0).$$

When Justified Indifference is satisfied and if we restrict attention to strict attribute ordering, the relevant attributes are uniquely identified by the revealed relevant ones. Formally, let $\mathscr{M}^\star(\succsim_t) = \{\mathbf{m} : \exists$ a partial order $>$ s.t. $(\mathbf{m}, >)$ rationalizes $\succsim_t\}$ be the set of relevant attributes that rationalize preferences at t using a strict attribute ordering. When Justified Indifference is satisfied, we have $\mathscr{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.

Theorem 2. *Assume richness and suppose that preferences* $(\succsim_t)_t$ *are transitive. Preferences satisfy Strong Restricted Reversal, Acyclicity and Justified Indifference if and only if there exists a deliberate preference change model* $(>, \rhd, \mathbf{m}_t, \mathbf{a}_t)$ *that rationalizes* $(\succsim_t)_t$ *with* $>$ *being a partial order. Furhermore, for any period* t*,* $\mathscr{M}^\star(\succsim_t) = \{\mathbf{m}_t\}$.

1.3 Motivated Preference Change

Our main representation theorem shows that preference change can be represented by the maximization of a meta-preference. The representation, however, does not provide a straightforward interpretation of the meta-preference. It could be that DM is changing her behavior to make it more aligned with her values, or she may change preferences to serve her own-interests instead of purely disinterested motives—this is referred to as *motivated preference change*. In this section, we investigate the latter idea. We show that motivated preference change admits a tractable functional representation—this proves convenient for applications in the

next section.

First, we construct an extension of the attribute ordering which allows us to keep track of (i) preferences over perceived alternatives at period t, and (ii) preferences over perceived alternatives at period t if she were to change her preferences to make good alternatives even better.

Definition 2. *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$. *Denote* $\mathbf{a} \gg_t \mathbf{b}$ *if* $\mathbf{x} \circ \mathbf{m}_t = \mathbf{a}$ *for some* $\mathbf{x} \in X$ *and*

(i) $\mathbf{y} \circ \mathbf{m}_t = \mathbf{b}$ *for some* $\mathbf{y} \in X$ *and* $\mathbf{x} \succsim_t \mathbf{y}$*; or*

(ii) $\mathbf{y} \circ \mathbf{m} = \mathbf{b}$ *for some* $\mathbf{y} \in X$, $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, *and* $\mathbf{x} \succsim_t \mathbf{z}$ *for all* $\mathbf{z} \in X$.

The following axiom, which extends Strong Restricted Reversals, guarantees that DM makes attributes relevant if and only if these attributes are valued positively—that is, making these attributes (ir)relevant increases DM's utility.

Motivated Restricted Reversals. *Preferences* $(\succsim_t)_t$ *satisfy Motivated Restricted Reversals if for any* $\{t_1, \ldots, t_n\}$ *and any* $(\mathbf{a_k})_{k=1,\ldots,n} \in (\mathbb{R}^K)^n$ *such that* $\mathbf{a}_{k+1} \gg_{t_k} \mathbf{a}_k$ *for* $k = 1, \ldots, n-1$,

$$\mathbf{a}_1 \gg_{t_n} \mathbf{a}_n \implies \mathbf{a}_1 \ll_{t_n} \mathbf{a}_n.$$

The next axiom guarantees that there are no indifference between vectors of relevant attributes when changing preferences. Intuitively, the axiom states that if there is a tie between two vectors $\mathbf{m}$ and $\mathbf{m}'$ that yield identical utility, DM breaks the tie in favor of one vector by virtually increasing her utility for some alternative $x \in X$ so that $\mathbf{m}$ becomes strictly preferred to $\mathbf{m}'$.

Motivated Tie-Breaking. *Preferences* $(\succsim_t)_t$ *satisfy Motivated Tie-Breaking if for all* t, *all* $\mathbf{x} \in \max(X, \succsim_t)$, *and all* $\mathbf{y}, \mathbf{y}' \in X$ *such that there exists* $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t -$

$\mathbf{m}_{t-1}|$) *with* $\mathbf{y}' \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m} \circ \mathbf{m}_t$,

$$\mathbf{y}' \in \max(X, \succsim_t) \implies \mathbf{m} = \mathbf{m}_t.$$

These two axioms are necessary and sufficient for the motivated preference change representation.

Theorem 3 (Representation)**.** *Suppose that preferences* $(\succsim_t)_t$ *are transitive. Preferences* $(\succsim_t)_t$ *satisfy Motivated Restricted Reversals and Motivated Tie-Breaking if and only if there exists a sequence of awareness* $(\mathbf{a}_t)_t$ *and a function* $u : \mathbb{R}^K \longrightarrow \mathbb{R}$ *such that for all* t *and all* $\mathbf{x}, \mathbf{x}'$,

$$\mathbf{x} \succsim_t \mathbf{x}' \iff u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{x}' \circ \mathbf{m}_t),$$

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\operatorname{argmax}} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}).$$

As in the previous representation, DM chooses alternatives to maximize her attribute ordering, which can be represented by a utility function here. The main difference is that preference change must maximize DM's utility. Therefore, all attributes that are "negatively valued" will be made irrelevant as soon as possible, and all attributes that are "positively valued" will be made relevant as soon as possible.

1.4 An Application

An important feature of the model is path dependence—that is, the order in which DM becomes aware of certain attributes has a strong impact on the path of preference change. We illustrate this aspect in a voting context: ex-ante identical voters deliberately ignore what other voters think is relevant because this would

undermine their view of their preferred candidate.[14] Therefore, we show that our model can account for polarization of political preferences among ex-ante identical voters in a simple and intuitive way.

Polarization refers to disagreement on policy issues or distrust of the other party members among politicians and citizens (Iyengar, Lelkes, Levendusky, Malhotra, and Westwood, 2019). There is now widespread agreement concerning the growing importance of ideological divisions both among politicized and educated voters as well as non-politicized citizens (Abramowitz and Saunders, 2008). There is no agreement, however, on the causes of polarization.[15]

From a Bayesian perspective, it is surprising that polarization increases as rational agents whose posterior beliefs are common knowledge cannot agree to disagree, even if their posteriors are based on different observed information about the world (Aumann, 1976). Arguing that voters have different priors certainly explains polarization, but it only moves the goalpost: where do differences in prior come from? Instead, our model provides a foundation for the concept of "partisan social identity" introduced in the political science literature (Iyengar and Westwood, 2015). This theory captures the tendency of voters to classify opposing partisans as members of an outgroup and copartisans as members of an ingroup. We show that our model can account for the construction of such opposing groups, and how partisan cues can reinforce division.

We consider a very stylized model with motivated preference change. There are two voters i and j and two candidates: $\mathbf{x}^D = (x^1, x^2, x^3)$ and $\mathbf{x}^R = (\tilde{x}^1, \tilde{x}^2, \tilde{x}^3)$ with $\tilde{x}^1 < 0 < x^1$, $x^2 < 0 < \tilde{x}^2$, $x^3 < 0 < \tilde{x}^3$ and $\tilde{x}^2 - \tilde{x}^1 > x^1 - x^2$. The first attribute captures

[14] Note that Bayesian updating cannot induce this type of path dependence because it is order invariant (Cripps, 2018).

[15] Recent finding suggests that the emergence of the internet or rising economic inequality are less plausible causes than changes that are specific to the US—e.g., changing party composition, growing racial divisions, or the emergence of partisan cable news (Boxell, Gentzkow, and Shapiro, 2020).

the candidates' support for social policies (e.g. health care), the second attribute captures how conservative candidates are, and the third attribute represents corruption. Voters are ex-ante identical: they both value integrity and prefer candidates with strong convictions (represented by a high absolute value of the difference between the first and the second attributes). We can represent their preferences as follows:

$$u(\mathbf{x} \circ \mathbf{m}) = \left(x^1 m^1 - x^2 m^2\right)^2 - x^3 m^3.$$

They both initially start with the vector of relevant attributes $(0,0,0)$. Suppose that voter i attends a political debate with both candidates: $\mathbf{a}_1^i = (1,1,0)$. She will change her preferences and value more candidate $\mathbf{x}^R$ who has stronger convictions: the meta-maximization writes

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)) = \left(\tilde{x}^2 - \tilde{x}^1\right)^2 &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,1,0)) = \left(\tilde{x}^2\right)^2 \\ &> \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,0,0)) = \left(x^1\right)^2 \\ &> 0 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (0,0,0)).\end{aligned}$$

Later, voter i becomes aware that candidate $\mathbf{x}^R$ is corrupted: $\mathbf{a}_2^i = (0,0,1)$. She decides to ignore this information and keep this attribute irrelevant if:

$$\begin{aligned}\max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,1)) &= \max\left\{\left(\tilde{x}^2 - \tilde{x}^1\right)^2 - \tilde{x}^3, \left(x^1 - x^2\right)^2 - x^3\right\} \\ &< \left(\tilde{x}^2 - \tilde{x}^1\right)^2 = \max_{\mathbf{x}\in X} u(\mathbf{x} \circ (1,1,0)).\end{aligned}$$

i.e. whenever $\left(\tilde{x}^2 - \tilde{x}^1\right)^2 > \left(x^1 - x^2\right)^2 - x^3$. Namely, whenever candidate $\mathbf{x}^R$ has strong convictions that counterbalance her corruption. The intuition is that making "corruption" relevant would undermine her view of candidate $\mathbf{x}^R$. In the end, voter

i's most preferred candidate is $\mathbf{x}^R$.

Instead, voter j first becomes aware of a felony committed by candidate $\mathbf{x}^R$: $\mathbf{a}_1^j = (0,0,1)$. She will change her preferences to make it relevant: the meta-maximization writes

$$\max_{\mathbf{x}\in X} u(\mathbf{x}\circ(0,0,1)) = -x^3 > 0 = \max_{\mathbf{x}\in X} u(\mathbf{x}\circ(0,0,0)).$$

At this point voter j prefers the upstanding candidate $\mathbf{x}^D$.

Later, voter j attends a political debate with both candidates: $\mathbf{a}_2^j = (1,1,0)$. She will lean toward the candidate $\mathbf{x}^D$ even though he has less convictions than the candidate $\mathbf{x}^R$ whenever $(x^1 - x^2)^2 - x^3 > (\tilde{x}^2 - \tilde{x}^1)^2 - \tilde{x}^3$. Namely, whenever the convictions of $\mathbf{x}^R$ does not make up for his felonies. In the end, voter j's most preferred candidate is $\mathbf{x}^D$.

It is quite striking that two identical voters who become aware of the same attributes can become polarized. This arises due to the path dependence of preference change: past justifications can conflict with new justifications leading to rich dynamics.

Chapter 2

Note on the Identification of Deliberate Preference Change

Joint with Niels Boissonnet (Bielefeld University).

In a companion paper (Boissonnet, Ghersengorin, and Gleyze, 2022a, henceforth BGG), we propose a model of rational preference change that is axiomatically founded and falsifiable. For that purpose, we narrow down the scope of preference change our model captures and provide a normative foundation to what we call *deliberate preference change*. Namely, changes that are triggered deliberately by the decision maker (DM) following the awareness of new values, reasons or dimensions of the world. The primitives are the observation of a succession of preference orderings over a fixed set of options and the attributes that characterize these options. The axioms build on an object identified from these primitives: the sequence of *revealed relevant attributes*. DM's successive preference changes can be represented *as if* she were truly using the revealed relevant attributes, both to make her choice at each period and to change her preferences between periods. A violation of the axioms may thus suggest that the revealed relevant attributes are not the *actual* relevant attributes for DM, in which case her behavior may still be rationalized by our model but with a different sequence of relevant attributes. Therefore, BGG's analysis leaves aside an *indeterminacy problem*. In this note, we precise what is this problem and provide axioms that characterize a more general version of deliberate preference change.

Primitives. We restrict ourselves to the case where attributes are binary.[1] Hence, instead of defining alternatives as vectors, we define them by the set of attributes they possess. Formally, there is a finite set of attributes $\mathscr{M}$. Denote $\mathbf{M} = 2^{\mathscr{M}} \setminus \emptyset$ all the non-empty combination of attributes. An alternative is defined as an element of $\mathbf{M}$. Let $X \subseteq \mathbf{M}$ be the set of alternatives. We make a richness assumption that all combinations of attributes are instantiated by an alternative.

[1] A attribute is binary if it can take only two values that inform whether an object possesses this attribute or not. For instance "colour" cannot be an attribute, we would need to divide it into a set of binary attributes, one for each color. Similarly a continuous attribute would be divided into intervals, and there would be a binary attribute for each interval.

Perfect Instantiation (richness assumption): $X = \mathbf{M}$.

The analyst observes (i) the attributes of all alternatives and (ii) choices over options for T periods of time ($T \geq 2$). The latter are represented by a sequence of complete preorders $(\succsim_t)_{t=1,\dots,T}$, where $\succ_t$ and $\sim_t$ respectively denote the asymmetric and symmetric parts.

Deliberate Preference Change. From these primitives, we can identify the *revealed relevant attributes*; that is, the attributes an observer can be certain that DM uses to make her choice at period t.

Definition 3. *An attribute $m \in \mathscr{M}$ is **revealed relevant** at period t if $x \not\sim_t x \cup \{m\}$ for some alternative $x \in X$. $\underline{M}_t$ is the set of revealed relevant attributes at period t.*

In BGG we characterize a representation that we call **deliberate preference change**. In this model, DM's preference at each period is represented by a time-independent preorder on the set of attribute combinations $\geqslant \subseteq \mathbf{M}^2$—named the **attribute ordering**—together with the revealed relevant attributes. Namely, for any period t, any $x, y \in X$:

$$x \succsim_t y \iff x \cap \underline{M}_t \geqslant y \cap \underline{M}_t. \tag{2.1}$$

Preference changes take the following form: whenever DM becomes aware of an attribute—through education, social interactions, media or introspection—she can decide to make it relevant or irrelevant for the next period, inducing a preference change. Formally, she is aware of a subset of attributes $A_t \in \mathbf{M}$ (revealed through the representation) between t and $t+1$, which defines a set of **reachable relevant attributes**, together with the current set of revealed relevant attributes:

$$R(\underline{M}_t, A_t) \equiv \{M \in \mathbf{M} : M_t \setminus A_t \subseteq M \subseteq M_t \cup A_t\}.$$

The succession of such changes is consistent with the maximization of a meta-preference relation, that is, a linear order $\rhd \subseteq \mathbf{M}^2$, such that at each period t, $\underline{M}_{t+1}$ is the maximum among the set of reachable relevant attributes:

$$\underline{M}_{t+1} = \max(R(\underline{M}_t, A_t), \rhd). \tag{2.2}$$

Indeterminacy Problem. The **indeterminacy problem** relates to the fact that an indifference between x and $x \cup \{m\}$ for every x does not necessarily imply that DM considers m as *irrelevant*. It could be that she is fundamentally *indifferent* about this attribute but still considers it as relevant for her choices. Said differently, an indifference could be rationalized either by an indifference of the attribute ordering, or by making irrelevant the attributes that differ between the alternatives.

BGG leaves aside this indeterminacy problem by taking the stance that if an attribute is not revealed relevant, then this is *as if* it is irrelevant. But the set of revealed relevant attributes $\underline{M}_t$ is not the only candidate set of relevant attributes for period t. This is simply the most parsimonious one, in the sense that it is the intersection of every possible set of relevant attributes for which equation (2.1) is satisfied, and any superset of $\underline{M}_t$ is also a candidate. This is summarized by the following proposition—a direct implication of BGG (Proposition 4). Formally, for any $M \in \mathbf{M}$, if there exists a preorder $\succcurlyeq \in \mathbf{M}^2$ such that equation (2.1) is satisfied replacing $\underline{M}_t$ by M, we say that $(M, \succcurlyeq)$ **rationalizes** $\succsim_t$. We define $\mathscr{M}(\succsim_t) = \{M \in \mathbf{M} : \exists \text{ a preorder } \succcurlyeq \in \mathbf{M}^2 \text{ s.t. } (M, \succcurlyeq) \text{ rationalizes } \succsim_t\}$.

Proposition 6. *If $(\succsim_t)_t$ is represented by deliberate preference change, then for any t, $\mathscr{M}(\succsim_t)$ is a lattice ordered by $\subseteq$. Its minimum is $\underline{M}_t$ and its maximum is $\mathscr{M}$.*

Therefore, even though $(\succsim_t)_t$ fails to be represented by deliberate preference change, there can still be other sequences of relevant attributes $(M_t)_t$ and awareness $(A_t)_t$ such that equations (2.1) and (2.2) are satisfied. In this case, the

attributes in $M_t \setminus \underline{M}_t$ do not have have a direct impact on choice at t—as can be seen from Proposition 6—, but they impact how DM changes preferences through the meta-choice (2.2). We refer to attributes in $M_t \setminus \underline{M}_t$ as *background attributes*. We define this generalized version of deliberate preference change and then provide an axiomatization.

Definition 4. $(\succsim_t)_{t=1,\dots,T}$ *is represented by* ***general deliberate preference change*** *if there exists a preorder* $\geqslant \subseteq \mathbf{M}^2$*, a sequence of relevant attributes* $(M_t)_t \in \mathbf{M}^T$*, a sequence of awareness* $(A_t)_t \in \mathbf{M}^T$*, and a linear order* $\rhd \subseteq \mathbf{M}^2$*, such that, for any* t *and any* $x, y \in X$*,*

$$x \succsim_t y \iff x \cap M_t \geqslant y \cap M_t, \tag{2.1*}$$

$$M_{t+1} = \max(R(M_t, A_t), \rhd). \tag{2.2*}$$

Axioms. Instead of working directly on candidate sets of relevant attributes, it will prove useful to ask: what are the attributes that *must have changed* between two periods? Answering to this question typically does not identify a unique sequence of relevant attributes. Therefore, given the indeterminacy problem at each period, it is more appropriate to express "candidates" as sets of attributes that must have changed to explain DM's behavior between any two periods t and t'. Furthermore, the objective of our model is ultimately to understand preference change. If we were to work directly on candidate sets of relevant attributes, then the conditions would trivially coincide with the definition of the model. The drawback of course is that these are partially identified objects from preferences which are typically not unique.

We define an **explanation** $E = (E_{t,t'})_{t<t'}$ of the sequence $(\succsim_t)_t$ as a bi-sequence whose element $E_{t,t'}$ represents a change in relevant attributes between period t and t'. That is, $E_{t,t'} = M \triangle M'$ for some $M \in \mathscr{M}(\succsim_t)$ and for some $M' \in \mathscr{M}(\succsim_{t'})$.

Importantly, an explanation is compatible with *multiple* sequences of relevant attributes $(M_t)_t$, hence it is a non-trivial exercise to find conditions on explanations that characterize the model. Conversely, an explanation is typically not unique given a dataset hence these conditions cannot be interpreted as axioms on choice directly.

We first want a counterpart of Strong Restricted Reversal (see BGG) that ensures the existence of an attribute ordering. If an attribute m is a background attribute for all periods then the analyst can never discover the ranking of combinations of attributes that include this one. It follows that we should not impose any form of consistency in the ranking of alternatives that possess such an attribute. At the other extreme, the analyst directly observes the ranking of the attributes that are *revealed relevant* for all periods. Therefore we must impose consistency across periods of DM's choices with respect to these attributes; hence Strong Restricted Reversal is necessary. In-between, there are background attributes that are *sometimes* revealed relevant. Whenever these attributes are revealed relevant, the analyst observes DM's ranking on these attributes and therefore we must impose consistency on these preferences. For a given explanation E, let define for any t, t' with $t < t'$,

$$V^E_{t,t'} \equiv \{m \notin E_{t,t'} : m \text{ revealed relevant at } t'\},$$
$$V^E_{t',t} \equiv \{m \notin E_{t,t'} : m \text{ revealed relevant at } t\},$$

and, $V^E_{t,t} = \underline{M}_t$.

An attribute m is in $V^E_{t,t'}$ implies that m is necessarily relevant at t'. Because it is not in $E_{t,t'}$, it did not change between t and t'. This means that either m is also revealed relevant at t, or is a background attribute at t. In any case, it is

relevant for the choice at both periods. Therefore, choices that are made involving this attribute must be consistent between t and t'.

Strong Restricted Reversal*. *For any* $(t_1,\dots,t_n) \in \{1,\dots,T\}^n$ *and any* $(x_k,x'_k)_{k=1,\dots,n}$ *such that:*

$$x'_k \cap (V^E_{t_k,t_{k+1}} \cup \underline{M}_{t_k}) = x_{k+1} \cap (V^E_{t_{k+1},t_k} \cup \underline{M}_{t_{k+1}}) \quad \text{for } k = 1,\dots,n-1, \text{ and}$$
$$x'_n \cap (V^E_{t_n,t_1} \cup \underline{M}_{t_n}) = x_1 \cap (V^E_{t_1,t_n} \cup \underline{M}_{t_1}).$$

If $x_k \succsim_{t_k} x'_k$ *for* $k = 1,\dots,n-1$, then $x'_n \succsim_{t_n} x_n$.

We now impose some form of coherency on the analyst's explanation. The explanation should not exhibit "gaps" in the sense that any sequence of *local* changes of attributes between t and $t+1$, $t+1$ and $t+2$, ... until τ and $\tau+1$ should be consistent with the *global* explanation from t to $\tau+1$. For instance, if m becomes relevant between t and $t+1$ in the analyst's explanation, and then becomes irrelevant between $t+1$ and $t+2$, then m cannot be used to rationalize DM's behavior from t to $t+2$.

No Explanatory Gap. *For every* $t < t' < t''$: $E_{t,t'} \triangle E_{t',t''} = E_{t,t''}$.

Finally, an acyclicity condition captures the principle of deliberation that deliberate preference change imposes on the meta-choice between periods. If the analyst's explanation between t and $\tau > t+1$ is included in the explanation between t and $t+1$, this means that any attribute that changed between t and τ already changed between t and $t+1$. Therefore, for these changes to be consistent with a meta-maximization, no attribute must change between $t+1$ and τ.

ACYCLIC EXPLANATION. *For any t and $\tau > t+1$:*

$$E_{t,\tau} \subseteq E_{t,t+1} \Longrightarrow E_{t+1,\tau} = \emptyset.$$

Theorem 4. $(\succsim_t)_{t=1,\ldots,T}$ *can be represented by* ***general deliberate preference change*** *if and only if there exists an explanation that satisfies Strong Restricted Reversal*, No Explanatory Gap and Acyclic Explanation.*

Chapter 3

Grabbing the Forbidden Fruit: Restriction Sensitive Choice

Joint with Niels Boissonnet (Bielefeld University).

"Prohibitions create the desire they were intended to cure."

Lawrence Durell

3.1 Introduction

Restricting an individual's feasible opportunities may steer their desire toward the prohibited opportunities or their substitutes. This phenomenon is known as the *forbidden fruit effect*, a reference to the episode of the Genesis when God tells Adam and Eve that they are free to help themselves to any food in the Garden of Eden except the fruit from the tree of the knowledge of good and evil, which they finally eat (Levesque, 2018). The forbidden fruit effect has received empirical support in various contexts, such as the choice of environmentally harmful products, media choices, reluctance to follow a nudge policy, smoking decisions, alcohol intake, eating behaviors, etc.[1] Although many of such decisions may have important economic consequences, this has rarely been explored in economics.

The forbidden fruit effect generates choice behaviors that are incompatible with the canonical model of preference (or utility) maximization. According to the latter, an agent holds a fixed ranking over alternatives and chooses the option ranked the highest among any set of opportunities they might face. Accommodating the forbidden fruit effect entails relaxing this standard requirement and allowing for menu-dependent choices. Specifically, studying reactions to restrictions amounts

[1] For the choice of environmentally harmful products, see Mazis, Settle, and Leslie (1973), see also the "rolling coal" movement in the US (in reaction to regulations of cars gas emissions, some drivers modified their engine at significant costs in order to pollute more). See Arad and Rubinstein (2018) for the reaction to nudges. For smoking decisions, see Pechmann and Shih (1999). For alcohol consumption, see Hankin, Firestone, Sloan, Ager, Goodman, Sokol, and Martier (1993). For eating behaviors, see Jansen, Mulkens, and Jansen (2007); Jansen, Mulkens, Emond, and Jansen (2008). For media choices, see Bushman (2006); Sneegas and Plank (1998); Varava and Quick (2015); Gosselt, De Jong, and Van Hoof (2012).

to investigating violations of the "Independence of Irrelevant Alternatives" (IIA) (Chernoff, 1954; Sen, 1971, property α) triggered by the *removal* of opportunities.[2] Let us illustrate this with a field experiment studied by Mazis et al. (1973). In 1972, Miami-Dade county decided to forbid phosphate use for laundry. Despite its strong environmental rationales, this decision raised significant protests as well as unexpected reactions. For the sake of the "American freedom", some consumers, among whom some were not buying phosphate-based detergent prior to the law, started buying it in neighbouring counties, smuggling it at extra cost and stockpiling the (now) precious product for the 20 years to come.[3] Formally, denoting by x the phosphate detergent in a neighbouring county, y the same product in Miami and z a phosphate-free detergent in Miami, the following choice reversal happens: z is chosen from the set $\{x,y,z\}$ while x is chosen over z once y is removed, i.e., in the menu $\{x,z\}$.

In this paper, we study a class of choice procedures, named *restriction sensitive choice* (RSC), that account for the forbidden fruit effect (section 3.2.2). RSC can be seen as a four-stage process. First, the DM categorizes the set of options into *types* (e.g., horizontal differentiation). Second, options within types are ranked according to a *utility function* u, which represents the DM's intrinsic satisfaction, or material welfare (e.g., vertical differentiation). Third, within each type, the DM determines a *threshold* utility level, below which the options are evaluated by a *reaction function* v (which differs from u). Fourth, the choice is made by choosing among the top available elements from each type (according to u), where the top element is evaluated through v or u depending on whether it is above or

[2] Property α is a weakening of the *Weak Axiom of Revealed Preferences* (Samuelson, 1938), that is necessary and sufficient to explain a single-valued choice function by the maximization of a linear order (see Sen, 1971).

[3] As Mazis et al. (1973) showed, this astonishing effect on behavior was consistent with consumers' beliefs reversal: Miami consumers were, on average, more prone to praise phosphate detergent for its efficiency than their Tampa county neighbors.

below the threshold. To illustrate the model, consider three options x, y, z that are horizontally and vertically differentiated; namely, x is of the same *type* of product as y, but at a higher price; z is of another type. The decision maker (DM) has an intrinsic preference for z over the options of the other type. This is captured through the *utility function* u: $u(z) > u(y) > u(x)$. Therefore, z is chosen from the set $\{x, y, z\}$. However, when the access to options of the first type is restricted to the bad one (i.e., x), then the DM gets further motive for choosing an option of this type (i.e., choosing x over z), which generates a forbidden fruit effect. This is captured through the the threshold of the first type, which is between $u(y)$ and $u(x)$, and the *reaction function* v such that $v(x) > u(z)$. We interpret this as v combining welfare and the additional desire created by restrictions.

The two prominent explanations of the forbidden fruit effect in psychology have been *reactance theory* (Brehm, 1966) and *commodity theory* (Brock, 1968), both of them being consistent with RSCs (section 3.2.3).[4] Reactance relates people's reaction to restriction or prohibition to their attitudes toward freedom of behavior. When they feel that a specific freedom of behavior is threatened, they experience psychological reactance, a motivational state toward the *restoration* of this lost or threatened freedom. With this in mind, in our model, each type of options subjectively embodies a specific freedom.[5] The threshold delimits the minimal welfare requirement such that when only options below it are available, the DM perceives this as a threat on that particular freedom. The reaction function v therefore captures the propensity of the DM to restore a threatened freedom.

[4] See Rosenberg and Siegel (2018) for a review on psychological reactance theory; Lynn (1991) for a review on commodity theory.

[5] Importantly, types are not postulated *a priori* and objectively observed but subjectively perceived by the DM and thus revealed by the analysis. Psychologists emphasize that reactance reflects an attempt to restore the loss of concrete freedoms, that is, freedoms to choose diverse types of option. "Contrary to some interpretations (e.g. Dowd, 1975), the freedoms addressed by the theory are not "abstract considerations," but concrete behavioral realities. If a person knows that he or she can do X (or think X, or believe X, or feel X), then X is a specific, behavioral freedom for that person." (Brehm and Brehm, 2013, p.12)

Commodity theory predicts that the more a commodity is perceived as unavailable or requiring much effort to be obtained, the more it will be valued. According to this interpretation, a type gathers similar commodities and when only options below the threshold are available this makes salient the restriction on this type of options, thereby increasing their attractiveness.

We investigate the identification of our model (section 3.3). We define a notion of revealed reaction to restriction in the following way: when we observe a choice reversal such as $z = c\{x, y, z\}$ but $x = c\{x, z\}$, we say that x *reacts to the absence of* y (section 3.3.1). Our interpretation is that the removal of y creates an additional desire to choose x. We show that ingredients of an RSC are essentially pinned down by this revealed reaction relation (section 3.3.2). In particular, x reacts to the absence of y implies that x and y are of the same type and x is below the threshold; therefore, the types and the thresholds can be identified in this way. Furthermore, building on a uniqueness result for the utility and the reaction functions, we give a definition of *welfare improvement* for an RSC. We illustrate by means of examples how our welfare criterion differs both from the conservative one of Bernheim and Rangel (2007, 2008, 2009) and the preference identified from choice with limited attention by Masatlioglu, Nakajima, and Ozbay (2012), hence contributing to the literature on welfare analysis under nonstandard individual choice.[6]

Our identification results rely on the particular structure of RSCs. Therefore, this naturally leads to the question of the falsifiability of our model. We thus give an axiomatic characterization (section 3.4). To that purpose, we first suitably relax IIA by requiring a standard Expansion axiom. Then, building on our definition of revealed reaction to restriction and a companion one, we state four axioms that capture consistency conditions on choices responsive to restrictions. We show that these five axioms fully characterize RSC. Importantly, other frequently

[6] See Manzini and Mariotti (2012); Chambers and Hayashi (2012); Rubinstein and Salant (2012); Apesteguia and Ballester (2015); Grüne-Yanoff (2022).

observed phenomena generate similar choice patterns as the ones resulting from the forbidden fruit effect. In particular, the analysis of the attraction effect by Ok, Ortoleva, and Riella (2015) is also based on choice reversals. However, our axioms are incompatible with theirs as long as some choice reversals are observed. Furthermore, RSC is a specific case of *choice with limited attention* (Masatlioglu et al., 2012). Yet, the interpretation is different and therefore so are the welfare predictions. Finally, RSC is also a specific case of a large class of choice procedures, popularized by the seminal paper by Manzini and Mariotti (2007), that sequentially apply two rationals. In particular, RSC is a *transitive shortlist method*, according to which choices are made by sequentially applying a pair of transitive preferences (Horan, 2016). The reverse is however not true.

In section 3.5, we take the point of view of a DM who behaves according to an RSC and ask how would they evaluate the freedom offered by the different sets of opportunities they might face; thus contributing to the literature on freedom of choice (see Baujard, 2007, for a survey of the literature). Building on the series of papers by Pattanaik and Xu (1990, 1998, 2000), we axiomatize a criterion to rank menus, which simply counts the number of types from which sufficiently good options (i.e., above the threshold) are feasible. We argue that this ordering integrates considerations about similarities between options (see Pattanaik and Xu, 2000; Nehring and Puppe, 2002) and the role of the preferences of the agent (see Pattanaik and Xu, 1998), two aspects that have been studied separately in the literature.

We finally study three applications of our choice model (section 3.6). Two social phenomena have often been related to reactance and documented by the psychology literature, but they are not readily explained using existing (economic) models of choice. First, reactance is introduced as a possible determinant of the formation of conspiracy theories. To accommodate this phenomenon, we study

how reactance impacts the DM's belief when she has to choose a biased source of information. By removing an unchosen moderately biased source, the DM might reverse their choice and choose a more biased source in the opposite direction. This can represent why, if a DM feels that some information is not accessible or hidden, they might end-up holding extreme belief or adhere to conspiracy theories. Second, reactance provides an explanation of why repressive policies towards minorities may generate backlash, as suggested by empirical evidence. Additionally, it provides an argument for the evolutionary efficiency of reactance and its persistence in the long run. Finally, we introduce RSC in a principal-agent's setting. We study a typical delegation problem: a principal can constrain the decision set of an informed but biased agent, but cannot commit to contingent monetary transfers. In addition to the standard model (e.g. Alonso and Matouschek, 2008), the agent behaves according to an RSC. We find that this modifies the optimal delegation strategy. Either it forces the principal to restrict even more the set of allowed actions to prevent the agent from taking worse actions; or it forces the principal to allow the agent's preferred options. Hence the effect of reactance on the agent's material welfare is ambiguous. This depends on the principal's payoff and prior distribution over the states of the world.

3.2 The Model

3.2.1 Preliminaries

We work with a finite set of options X and denote by $\mathscr{X} = 2^X \setminus \emptyset$ the collection of non-empty subsets of X. Elements of $\mathscr{X}$ stand for the menus of options available to the DM. A **choice function** $c : \mathscr{X} \longrightarrow X$ associates to each menu the option chosen

by the DM in this menu.[7] Namely, for any menu A, $c(A) \in A$.[8]

We are interested in studying the effect of restrictions of the set opportunities on the DM's choices. Hence, our interpretation is that the menu is exogenously given to the DM, who must then choose an option within this menu. We however do not explicitly model an agent who actually restrains the DM's set of opportunities, except in some applications.

Let us finally stress that options are defined by objective features that can incorporate contextual properties — for instance, in our introductory example, we differentiated the phosphate laundry in a supermarket in Miami from the same product in a supermarket in a neighbouring county. Yet, we need not formalize these objective features, we only require the observer to be able to distinguish the different options. As it will become clear, some of these features may matter for the DM's subjective categorization of options and thus will be revealed through the choices.

3.2.2 Restriction Sensitive Choice

In this section, we state our model, then detail the choice procedure it induces and finally give the two main interpretations. We first introduce two definitions.

Definition 5. *The* **order induced** *by a function $f : X \longrightarrow \mathbb{R}$ is the complete and transitive binary relation $\succsim^f \subseteq X^2$ such that, for any $x, y \in X$, $x \succsim^f y \iff f(x) \geq f(y)$.*

Definition 6. *A function $f : X \longrightarrow \mathbb{R}$ is* **single-peaked** *with respect to the linear order $\succ \subseteq X^2$, if for any $x, y, z \in X$ such that $x \succ y \succ z$, $v(y) \geq \min\{v(x), v(z)\}$.*[9]

[7] We focus on choice functions for the sake of simplicity: dealing with choice correspondences would add another layer of complexity that, we think, is not necessarily relevant in the present context. Nonetheless, we conjecture that, with an appropriate weakening of Sens' property alpha, our results would extend to choice correspondences.

[8] For simplicity, if we enumerate a set $\{x_1, \ldots, x_k\}$, we write $c\{x_1, \ldots, x_k\}$ instead of $c(\{x_1, \ldots, x_k\})$.

[9] A *linear order* is a complete, transitive and antisymmetric binary relation.

Equipped with these two definitions, we now state the definition of our model, *restriction sensitive choice*.

Definition 7. *A choice function c is a* **restriction sensitive choice** *(RSC) if there exist a partition $\mathscr{T}$ of the options into* types*, a* threshold $\lambda_T \in \mathbb{R}$ *for each $T \in \mathscr{T}$, a* utility function $u : X \longrightarrow \mathbb{R}$ *that induces a linear order on each $T \in \mathscr{T}$ and a* reaction function $v : X \longrightarrow \mathbb{R}$*, such that:*

(i) for any menu A,

$$\{c(A)\} = \arg\max_{x \in d(A)} v(x), \tag{3.1}$$

where:

$$d(A) = \bigcup_{T \in \mathscr{T}} \arg\max_{x \in T \cap A} u(x);$$

(ii) for any $T \in \mathscr{T}$, $u(.) = v(.)$ on $\{x \in T \mid u(x) \geq \lambda_T\}$;

(iii) for any $T \in \mathscr{T}$, v is single-peaked with respect to the order induced by u on $\{x \in T \mid u(x) < \lambda_T\}$.

In this case, we say that $\langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$ is an **RS-structure** *that* **rationalizes** *the choice function c.*

According to RSC, options are partitioned into types. Choices are made sequentially: first, the DM retains the best available options from each type according to u, forming the set $d(A)$; then, the DM chooses among this set according to v. Points ii and iii specify how the two criteria u and v are related to each other. Namely, the options above the thresholds λ_T's are evaluated according to u in both stages of the maximization; while the options below are evaluated in the second stage by v, which can differ from u. We now give interpretations of the model. Comparisons to existing models are relegated to section 3.4.

3.2.3 Interpretations

We give two interpretations of RSC: one in terms of reactance, the other in terms of saliency of a prohibition and the additional attractiveness it generates.

Reactance: a freedom-based theory of choice. According to the psychology literature, reactance is a reaction of an individual to a threat to their freedom of behavior, that aims to *restore* the eliminated freedom. In our framework, potential threats to freedom are captured by restrictions of the opportunity set.

The DM categorizes the options into types, forming a partition. In the words of psychologists, a type represents a certain *behavioral freedom* (see Brehm and Brehm, 2013). Within each type, options are ranked and chosen according to an instrumental criterion, the utility function u. Our interpretation is that u represents the intrinsic satisfaction, or material welfare, of the agent. A clear example is when a type contains the same good but obtained or consumed through different channels. For instance, buying phosphate laundry in the supermarket next-door is less costly than getting it in a supermarket in a neighbouring county, but both options may be perceived as similar.

The thresholds represent the minimal welfare requirements for the DM's freedoms. Namely, as long as options with a satisfaction level at least as good as λ_T are available, no freedom concern is activated regarding the specific freedom embodied by T. This is captured by ii in the definition of an RSC, which implies that, in the second step maximization, the evaluation of those options are not distorted. On the contrary, if the best option available from T is below λ_T, the DM deems that they do not have access to a sufficiently good option regarding the freedom associated to T. They may then be prone to react by being even more willing to choose an option from T, although this option gives a lower satisfaction. This is how they "restore" the eliminated freedom. It is captured through the function v used in the second

maximization, which combines welfare and the propensity to react to a freedom limitation. Indeed, as it will become clear in section 3.3, to generate reactance, v must exceed u for options below λ_T.

Point iii imposes a specific shape of the reaction function v with respect to the utility function u. This reflects the DM's increasing willingness to react, up to a certain point, as the limitations on their freedom is tightened (see Rosenberg and Siegel, 2018, for evidence of this phenomenon). In RSC, it is captured through the fact that the less welfare the DM can obtain from a type, the more they are willing to react. Single-peakedness simply allows this to be true up to a certain point where welfare motives might weigh more in the trade-off between welfare and freedom, that is, there might be a point where the DM considers the welfare sacrifice to be too important.

Commodity theory: when salient prohibition increases desire. Commodity theory states that the value of objects for the individuals increase with their feeling that the objects are impossible of difficult to access (Brock, 1968).

According to this interpretation, a type gathers options that the DM considers as similar — e.g., providing similar consumption experience — but with different level of satisfaction, or material welfare (as captured by u). The threshold λ_T captures the level of satisfaction below which the restricted availability of options of type T becomes *salient*. In this case, the best alternative option available from T becomes a "forbidden fruit" (e.g., Levesque, 2018), and thus all the more attractive. This is captured by v which adds, on top of the welfare, an intrinsic pleasure of defying the restriction. Similarly to the interpretation in terms of reactance, point iii captures the idea that the additional desire may increase with the degree of the restriction, up to a certain point.

3.3 Identification and Welfare

3.3.1 Revealed Reaction to Restriction

We are interested in the effects of restrictions of the opportunity set on choice behavior. These effects are observed when the motivation created by the restriction — be it related to freedom or the intrinsic desire for forbidden objects — conflicts with other motives, such as welfare or material satisfaction.[10] Formally, they are revealed through choice reversals that are inconsistent with standard preference maximization; namely, through violations of the *Independence of Irrelevant Alternative* (IIA, or property α, Sen, 1971) triggered by the removal of options. In particular, we are interested in the DM's reaction to the deprivation of an "apparently irrelevant" option, hence the following definition.

Definition 8. *Let c be a choice function on $\mathscr{X}$ and $x, y \in X$. We say that x* ***reacts to the absence of*** *y, relative to c, if there exists z such that, $z = c\{x,y,z\}$, and $x = c\{x,z\}$. We denote it $x\mathbf{R}^c y$.*[11]

x reacts to the absence of y means that being deprived of the feasibility of y, the DM's motive to choose x is boosted. We show that the relation $\mathbf{R}^c$ allows to uniquely identify the ingredient of an RSC.

3.3.2 Identification of Restriction Sensitive Choice

The first proposition characterizes the relation $\mathbf{R}^c$ for an RSC.

[10] Following Brehm (1966), reactance is meaningful only when freedom conflicts with another motive. "Reactance is conceived to be a counterforce motivating the person to reassert or restore the threatened or eliminated freedom. It exists only in the context of other forces motivating the person to give up the freedom and comply with the threat or elimination." (Brehm and Brehm, 2013, p.37).

[11] Our results are the same if the relation $\mathbf{R}^c$ is not defined using a triplet, but any set, i.e.: $x\mathbf{R}^c y$ if there exists a set A such that $x = c(A \setminus \{y\}) \neq c(A) \neq y$.

Proposition 7. *Suppose c is rationalized by the RS-structure $\langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$. For any $x, y \in X$: $x\mathbf{R}^c y$, if and only if there exists $T \in \mathscr{T}$ and $z \notin T$, such that $x, y \in T$, $u(x) < u(y)$ and $v(x) > v(z) > v(y)$.*[12]

A direct consequence of this proposition is the following corollary.

Corollary 1. *Suppose c is rationalized by the RS-structure $\langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$. For any $x, y \in X$, if $x\mathbf{R}^c y$, then there exists $T \in \mathscr{T}$ such that $x \in T$ and $u(x) < \lambda_T$.*

These results corroborate our interpretations. Indeed, x reacts to the absence of y means that: x and y are considered by the DM as similar; y is intrinsically preferred to x ($u(x) < u(y)$); and being deprived of y boosts the DM's willingness to choose x ($v(x) > v(y)$). In terms of reactance, this restriction is perceived by the DM as a threat to their freedom. According to the second interpretation, removing x makes the restriction (more) salient and increases the desire to choose the alternative x.

Thanks to these results, we can define a specific kind of RS-structures that rationalize an RSC and whose elements are identified from this relation $\mathbf{R}^c$. For that purpose, let define for a choice function c, the set T_0^c of the options that are never involved in any reaction to restriction:

$$T_0^c = \{x \in X : \nexists\, y \text{ such that } x\mathbf{R}^c y \text{ or } y\mathbf{R}^c x\}.$$

Proposition 8. *Suppose c is an RSC. Then, there exists an RS-structure $\langle \mathscr{T}, \{\lambda_T\}, u, v\rangle$ that rationalizes c, such that:*

(i) $T_0^c \in \mathscr{T}$;

(ii) for any $T \in \mathscr{T}$, $\lambda_T = \min u\big(\{x \in T : \nexists\, y, x\mathbf{R}^c y\}\big)$.

[12] All proofs of this section can be found in Appendix C.1.

In this case, we say that $\langle \mathcal{T}, \{\lambda_T\}, u, v\rangle$ *is a* **minimal RS-structure**.

Minimal RS-structures are appropriate to study RSCs for several reasons. First, it shows that options in T_0^c do not matter: they can be removed from every type and gathered together.[13] Second, the thresholds really capture the idea that as long as options giving a satisfaction of at least λ_T are available, the DM does not react to any restriction of the available options from T. Vice versa, when only options with satisfaction levels below the threshold are available, the DM's is prone to react to this restriction. Finally, any type besides T_0^c is *relevant*, in the sense that it gathers options that are related through the DM's responsiveness to restrictions. Namely, for any x in this type, it is related to some y in the same type through the relation $\mathbf{R}^c$. The following corollary makes these statements explicit.

Corollary 2. *Suppose* c *is rationalized by the minimal RS-structure* $\langle \mathcal{T}, \{\lambda_T\}, u, v\rangle$ *and define for each* $T \in \mathcal{T}$, $x_T \equiv u^{-1}(\lambda_T)$. *Then, for any* $T \neq T_0^c$ *and* $x \in T$:

(i) if $u(x) < \lambda_T$, *then* $x \mathbf{R}^c x_T$;

(ii) if $u(x) \geq \lambda_T$, *then there exists* $y \in T$ *such that* $y \mathbf{R}^c x$ *and* $y \mathbf{R}^c x_T$.

Another important aspect is that the types and the thresholds of minimal RS-structures are essentially unique.

Corollary 3. *Suppose* c *is rationalized by the minimal RS-structures* $\langle \mathcal{T}, \{\lambda_T\}_T, u, v\rangle$ *and* $\langle \tilde{\mathcal{T}}, \{\tilde{\lambda}_T\}_T, \tilde{u}, \tilde{v}\rangle$. *Then* $\mathcal{T} = \tilde{\mathcal{T}}$ *and* $x_T = \tilde{x}_T$ *for each* T *(where* $x_T = u^{-1}(\lambda_T)$ *and* $\tilde{x}_T = u^{-1}(\tilde{\lambda}_T)$*).*

Remark. In light of proposition 8, it is worth noting that the absence of choice reversals as in definition 8 is to be interpreted as a lack of traceable reaction to restriction. This does not necessarily mean that the DM has no concern for

[13] The interpretation of T_0^c as a type must thus be qualified: these options do not particularly share common features, they simply do not fall in any category.

restrictions of their opportunity set. Rather, this means that these concerns (if any) are either too weak, or too aligned with their welfare to be identified as a force counterbalancing welfare.

We finally discuss the uniqueness of the functions u and v. Let c be rationalized by the RS-structure $\langle \mathcal{T}, \{\lambda_T\}_T, u, v\rangle$. What are the joint conditions on functions $\tilde{u}, \tilde{v}$ that ensures that $\langle \mathcal{T}, \{\lambda_T\}_T, \tilde{u}, \tilde{v}\rangle$ also represents c? One obvious sufficient condition is if there exists an increasing function $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\tilde{u} = f \circ u$ and $\tilde{v} = f \circ v$. Now let suppose that there exist two functions $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\tilde{u} = f \circ u$ and $\tilde{v} = g \circ v$. Denote the following set gathering all options that are above the threshold of their type:

$$F \equiv \{x \in X : u(x) \geq \lambda_{T(x)}\}.^{14} \tag{3.2}$$

One clear necessary condition is that f and g coincide and are increasing on $u(F)$, so that $f \circ u$ and $g \circ v$ satisfy points i and ii in definition 7. Because u is not directly used as a choice rule on options that are not in F, it is not necessary that f be increasing on $u(X)$. Yet, it represents choices within types, which implies that f is increasing on $u(T)$ for every T. Because v is ultimately the function through which choices are made, one might be tempted to say that g must be increasing on $v(X)$. This is however not exact because within types, the function v is never used to make choices. This problem does not arise as long as we impose one additional innocuous condition regarding the reaction function on certain pairs of options of similar types. This is captured by the following definition.

Definition 9. *An RS-structure $\langle \mathcal{T}, \{\lambda_T\}_T, u, v\rangle$ is an* **RS*-structure** *if, for any $T \in \mathcal{T}$ and any $x, y \in T$, such that $u(y) < u(x)$, $u(y) < \lambda_T$ and for all $z \notin T$, $v(y) > v(z) \implies v(x) > v(z)$, then $v(y) \geq v(x)$.*

The next proposition shows the existence of RS*-structure and that if we

restrict ourselves to RS*-structures, the conditions regarding the utility and the reaction functions stated above are not only sufficient, but also necessary.

Proposition 9. *Let c be an RSC.*

(i) There exists an RS*-structure $\langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$ that rationalizes it.

(ii) Furthermore, let $f, g : \mathbb{R} \longrightarrow \mathbb{R}$ be two real-mappings, $\langle \mathscr{T}, \{\lambda_T\}_T, f \circ u, g \circ v\rangle$ also rationalizes c if and only if f is increasing on $u(T)$ for every $T \in \mathscr{T}$, g is increasing on $v(X)$ and $f|_{u(F)} = g|_{u(F)}$.

3.3.3 Welfare

Our interpretation is that u captures the intrinsic preferences, or the welfare, of the DM. We however learn from proposition 9 that if c is rationalized by a minimal RS-structure with utility function u, having $u(x) > u(y)$ is not sufficient to conclude that x is welfare improving on y. It is sufficient if either x and y are in the same type, or both are above the thresholds (or a transitive closure of the latter ideas). Hence the following definition.[15]

Definition 10. *Suppose c is rationalized by a minimal RS-structure $\mathscr{S} = \langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$. Then, for any $x, y \in X$, x is* **welfare improving** *on y, denoted $x \succ^w_{\mathscr{S}} y$, if either:*

(i) $T(x) = T(y)$ and $u(x) > u(y)$; or

(ii) $T(x) \neq T(y)$, $u(x) \geq \lambda_{T(x)}$ and there exists $z \in T(y)$, such that $u(z) \geq \lambda_{T(y)}$ and $u(x) > u(z)$.

The following proposition, which easily follows from corollary 3 and proposition 9, ensures that minimal RS-structures uniquely identify the welfare improving relation.

[15] The focus on minimal RS-structure is justified by the fact that otherwise some options might be mis-categorized in a type while it is in T^c_0 and some conclusions about welfare could be wrongly drawn.

Proposition 10. *Suppose c is rationalized by the minimal RS-structures $\mathscr{S}$ and $\tilde{\mathscr{S}}$. Then $\succ^w_{\tilde{\mathscr{S}}} = \succ^w_{\mathscr{S}}$.*

Interestingly, by means of examples we show that neither the the model-free criterion $P^\star$ of Bernheim and Rangel (2009) nor the preference identified from *choice with limited attention* P^R (Masatlioglu et al., 2012) coincide with $\succ^w_c$.[16]

Example 1: $\succ^w_c \not\subset P^\star$. Let $z = c\{x,y,z\}$ and $x = c\{x,z\}$ and suppose that $u(z) \geq \lambda_{T(z)}$, $u(y) \geq \lambda_{T(y)}$. Then $z \succ^w_c x$ whilde $\neg[zP^\star x]$.

Example 2: $P^\star \not\subset \succ^w_c$. Consider x such that $u(x) < \lambda_{T(x)}$, $z \notin T(x)$ such that for every $y \in T(x)$, $v(y) > v(z)$. Then $xP^\star z$ while $\neg[x \succ^w_c z]$.

Example 3: $\succ^w_c \not\subset P^R \wedge P^r \not\subset \succ^w_c$. It is easy to see in the example given by Masatlioglu et al. (2012) to show the difference between P^R and $P^\star$ (Example 1, pp. 2191-2192), that while they deduce $xP^R y$ we would concude that $y \succ^w_c$.

3.4 Characterization

In this section, we state our main theorem, which gives a full axiomatic characterization of RSCs, thereby showing the falsifiability of our model.

Restrictions are captured through the removal of options. Hence, expanding menus should reduce the motive to react. In particular, let x be chosen in menu A. Expanding A by adding a set of options B in which x is also chosen should not induce any change in the choice. Otherwise, the reversal from $A \cup B$ to A would be triggered by the loss of an option from $B \setminus A$, which should prevent x from being chosen in B. This is what our first axiom, a standard relaxation of the *Weak Axiom of Revealed Preference*, imposes.[17]

[16] $\succ^w_c = \succ^w_{\mathscr{S}}$ for any minimal RS-structure $\mathscr{S}$ that rationalizes c.

[17] It was already present in Sen (1971), named property γ, and was later used by Manzini and Mariotti (2007) under the name *Expansion*.

EXPANSION (Exp). *For any $x \in X$, $A, B \in \mathscr{X}$, if $x = c(A) = c(B)$, then $x = c(A \cup B)$.*

Note that if *x reacts to the absence of y*, then **Exp** implies that $y = c\{x, y\}$. That is, for a reaction to restriction to be meaningful, it must trigger a choice of an even "worse" option than the one that is no more available. Furthermore, if z plays the same role as in definition 8, then **Exp** also implies that $z = c\{y, z\}$. Therefore, a typical pattern of reaction to restriction implies a binary choice cycle.

It is worth noting though that **Exp** prevents the following case. Let x, y, z be three options that are alphabetically ordered according to a dimension (e.g., political bias). The DM chooses z when the three options are available. But the DM wants to have access to the extreme option x so that when it is removed, y is chosen over z. However, when only x and y are available, the DM still prefers to choose y as it is closer to their first-best option. In some sense, we capture reactions to restriction that involve some minimal link to intrinsic satisfaction. In this case, if having access to x really matters so much to the DM, then they must chooose it in some cases, in particular above y.

Our next axioms involve the relation $\mathbf{R}^c$ and a companion one $\mathbf{P}^c$ that we now define. Assume we observe that *(i)* y is preferred to x, and *(ii)* for each pair z, t such that $t = c\{y, z, t\}$ and $y = c\{y, t\}$, we also observe that $t = c\{x, z, t\}$ and $x = c\{x, t\}$.[18] In this situation we suspect x to be as effective as y to react to any restriction that induces the DM to choose y. Yet, it need not be the case that $x\mathbf{R}^c y$, for there might be no third option z that allow to reveal a reversal as stated by definition 8 — i.e. no z is chosen in $\{x, y, z\}$ while x is chosen in $\{x, z\}$. When this happens, we posit that x *potentially reacts* to the absence of y.

[18] Note that this is stronger than simply requiring that for any z such that $y\mathbf{R}^c z$ we also observe $x\mathbf{R}^c z$.

Definition 11. *Let c be a choice function on $\mathscr{X}$ and $x, y \in X$. We say that x* ***potentially reacts to the absence of*** *y, relative to c, if $y = c\{x, y\}$, there exists z, t such that $t = c\{y, z, t\}$ and $y = c\{y, t\}$, and for any such pair we also observe that $t = c\{x, z, t\}$ and $x = c\{x, t\}$. We denote it $x\mathbf{P}^c y$.*

We now state our axioms. Consider our introductory example. Assume that, when only an expensive phosphate-free detergent is available in their county, the DM reacts to the prohibition by going to the neighbouring county to get some phosphate detergent, while they stay in their own county when a cheap phosphate-free detergent is available. This reveals that "buying phosphate detergent in the neighbouring county" reacts to the absence of "buying phosphate detergent in Miami supermaket", though it is revealed only when the price of the phosphate-free detergent is high. Assume also that, while the DM prefers not to transgress the law when they can buy phosphate in the neighbouring county, they decide to go on the black market when the latter is forbidden, and that they do so whatever the price of the available phosphate-free detergent may be. This reveals that "buying phosphate detergent on the black market" reacts to the absence of "buying phosphate detergent in the neighbouring county". For these behaviours to be consistent, we would like to also observe that if phosphate is first banned in the neighbouring county and then in the DM's county, the DM goes on the black market, thus revealing that "buying phosphate detergent on the black market" also reacts to the absence of "buying phosphate detergent in Miami supermakets". Hence our first axiom requires $\mathbf{R}^c$ to be transitive — note that $\mathbf{P}^c$ is transitive by definition.

R-TRANSITIVITY (R-Tran). *For any $x, y, z \in X$, if $x\mathbf{R}^c y$ and $y\mathbf{R}^c z$, then $x\mathbf{R}^c z$.*

Let us stress that **R-Tran** imposes also transitivity in the *similarity* between options, that is, it prevents the following situation: x is sufficiently close to y and

$x\mathbf{R}^c y$, y is sufficiently close to z and $y\mathbf{R}^c z$, but x and z are too different to consider the possibility of x reacting to the absence of z.

Because $\mathbf{R}^c$ and $\mathbf{P}^c$ are typically incomplete, we also require the negative transitivity of these relations. It requires that if $x\mathbf{R}^c y$, then for any z that would be in-between x and y — i.e., $y = c\{y,z\}$ and $z = c\{z,x\}$ — , it must be that $x\mathbf{R}^c z$ or $z\mathbf{R}^c y$.

R-NEGATIVE TRANSITIVITY (R-NTran). *For any $x,y,z \in X$, such that $y = c\{x,y\}$, $z = c\{y,z\} = c\{x,z\}$:*

(i) if $\neg[x\mathbf{R}^c y]$ and $\neg[y\mathbf{R}^c z]$, then $\neg[x\mathbf{R}^c z]$;

(ii) if $\neg[x\mathbf{P}^c y]$ and $\neg[y\mathbf{P}^c z]$, then $\neg[x\mathbf{P}^c z]$.

To motivate our next axiom, consider an option y that never reacts to the absence of any other option, but whose removal triggers reaction from the DM by choosing x — i.e. $x\mathbf{R}^c y$. This means that as long as the DM has access to y, they never react to some limitation of their opportunity set by choosing y. At the same time, removing y triggers some reaction and motivate them to choose x. Therefore, our interpretation is that y is never chosen because of restriction-related motives: either y satisfies the DM's freedom requirement (interpretation 1); or y as long as y the unavailability of options similar to y is not sufficiently salient to make it more attractive (interpretation 2). Consider a third option z that is chosen over y and such that also $x\mathbf{R}^c z$, then z should be even less chosen because of restriction-related motives. Our third axiom imposes two conditions in that direction. First, any option that reacts to the absence of z must also react to the absence of y. Conversely, any option that reacts to the absence of y might not be good enough to react to the absence of z, but at least z must be chosen over this option.

R-CONSISTENCY (R-Con). *For any $x, y, z \in X$ such that $x\mathbf{R}^c y$, $x\mathbf{R}^c z$, $z = c\{y, z\}$ and there exists no t such that $y\mathbf{R}^c t$, and for any $u \in X$:*

(i) if $u\mathbf{R}^c z$, then $u\mathbf{R}^c y$;

(ii) if $u\mathbf{R}^c y$, then $z = c\{u, z\}$.[19]

To motivate our last axiom, we extend the phosphate example. Suppose that both "buying phosphate on the black market" (x) and "buying phosphate in a neighbouring county" (z) react to the absence of "buying phosphate in Miami supermakets" (t). Add the third option "buying phosphate in a further county" (y): quite naturally, z is chosen over y, and assume further that both z and y are chosen over x. Suppose that the DM considers going on the black market as a reaction to the prohibition in a further county, that is, $x\mathbf{R}^c y$. Said differently, the DM's propensity to choose a phosphate detergent when x is the only one available is greater than when y is available. Because both x and z reacts to the absence of a common option t, then one would expect that similarly the DM's motive to choose a phosphate detergent when y is the only one available is greater than when z available. Hence our third axiom requires that y potentially reacts to the absence of z. The second point says that if in addition $x\mathbf{P}^c z$, that is, whenever the DM considers going in a neighbouring county as a reaction to a restriction, they would also consider going on the black market if necessary, the same conclusion, that is, $y\mathbf{P}^c z$, should follow even if we only observe $x\mathbf{P}^c y$ and not necessarily $x\mathbf{R}^c y$. This axiom imposes some sort of monotonicity in the way the DM reacts to restrictions, in the sense that it forbids any "gap" in their reaction. That is to say, if x, y, z are transitively ranked in binary choices and they are all sometimes chosen as a reaction to the deprivation of a common option t, then the magnitude of the

[19] Point *(ii)* can alternatively be seen as requiring that $u\mathbf{R}^c y$ cannot be revealed through the choice with z, hence $z = c\{u, z\}$, which is consistent with the interpretation that y and z are similar and z offers a better satisfaction.

motivation to react should evolve monotonically from z to x. Hence, as long as $x\mathbf{R}^c y$ or $x\mathbf{P}^c y$, it must be that at least $y\mathbf{P}^c z$.

R-MONOTONICITY (R-Mon). *For any $x, y, z \in X$, such that $z = c\{y,z\}$, $y = c\{x,y\}$:*

(i) if $x\mathbf{R}^c t$ and $z\mathbf{R}^c t$ for some $t \in X$, then $[x\mathbf{R}^c y \implies y\mathbf{P}^c z]$;

(ii) if $x\mathbf{P}^c z$, then $[x\mathbf{P}^c y \implies y\mathbf{P}^c z]$.

We now can state our representation theorem, according to which an RSC is entirely characterized by these five axioms.

Theorem 5. *A choice function c is an RSC if and only if it satisfies **Exp**, **R-Tran**, **R-NTran**, **R-Con** and **R-Mon**.*[20]

Comparisons to existing models. There exist several models that explain similar choice patterns as the ones generated by reactions to restrictions, in particular the choice reversal exhibited in definition 8 (see Manzini and Mariotti, 2007, 2012; Cherepanov, Feddersen, and Sandroni, 2013b; Masatlioglu et al., 2012; Ehlers and Sprumont, 2008; Ok et al., 2015; Lleras, Masatlioglu, Nakajima, and Ozbay, 2017; Apesteguia and Ballester, 2013; Horan, 2016; Ridout, 2021, among others).

In particular, a phenomenon frequently observed and studied is the attraction effect. This is the main focus of Ok et al. (2015). In particular, their definition of *revealed reference* is based on similar choice patterns: $z = c\{x,y,z\}$ and $x = c\{x,z\}$. Their interpretation is however significantly different: they argue that z beats x only with the "help" of y. Hence, while they interpret these reversals as revealing a relationship between y and z, we interpret it as revealing a relationship between x and y.[21] Their model and ours are actually incompatible. Indeed, as we noted

[20] The proof is in Appendix C.2.

[21] More precisely, the application of their definition identifies y as a *revealed reference* of z.

after the statement of **Exp**, observing $x\mathbf{R}^c y$ and satisfying **Exp** imply a binary choice cycle, which is prevented by their *No-Cycle Condition*.

Among the different models that generate similar choice reversals, two have attracted a lot of attention: the *Rational Shortlist Method* (RSM) by Manzini and Mariotti (2007) and the *Choice with Limited Attention* (CLA) by Masatlioglu et al. (2012). Masatlioglu et al. (2012) actually show that these two models are both descriptively and behaviorally distinct. It happens that RSC is a special case of both CLA and RSM. First, our operator $d(.)$ satisfies the unique condition of an *attention filter*; namely, for any A $d(A) = d(A \setminus \{x\})$ whenever $x \notin d(A)$. Let c be an RSC, given that the choice $c(A)$ follows from the maximization of the function v over the set $d(A)$, this shows that c is a CLA. Note however that our interpretations and thus our welfare conclusions are different (see section 3.3.3).

Second, let c be an RS rationalized by $\langle \mathscr{T}, \{\lambda_T\}_T, u, v\rangle$ and define the two orders P_1 and P_2 in the following way: $xP_1y \iff T(x) = T(y) \wedge u(x) > u(y)$; $xP_2y \iff v(x) > v(y)$. In that case, for any menu A, $c(A) = \max(\max(A, P_1), P_2)$. That is, the DM chooses as if she first keeps only options that are the best in each available type, and second, she chooses the best remaining one according to the binary comparisons. Therefore, (P_1, P_2) *sequentially rationalize* c.[22]

3.5 Measuring Freedom

In this section, we adopt the interpretation in terms of reactance. We take the point of view of a DM whose final choices are rationalized by an RS-structure and ask how would they evaluate the freedom offered by the different sets of opportunity they might face. Starting with Jones and Sugden (1982) and Pattanaik and Xu (1990), there has been an important literature about freedom of choice that has

[22] More precisely, the two orders are transitive, hence (P_1, P_2) is a *transitive shortlist method* (Horan, 2016).

proposed a wide variety of freedom measures based on the ranking of opportunity sets (see (see Baujard, 2007, for a survey of this literature). Importantly for us, two dimensions have been pointed out as relevant to the agents' valuation of their freedom: their (potential) preferences over options (see Pattanaik and Xu (1998) — henceforth PX98 — and Sen (1993)) and the similarity between different options (see Pattanaik and Xu (2000) — henceforth PX00 — and Nehring and Puppe (2002)).

According to a (minimal) RS-structure, it is through the interaction between the types and the thresholds that freedom concerns impact choices. This suggests that these two channels should impact the DM's assessment of freedom offered by a given menu. The types represent classes of similar options,[23] suggesting that adding options of a similar type should not increase the DM's freedom of choice. In addition, the thresholds represent the DM's freedom demands. Hence, it seems natural that adding options increases the DM's valuation of freedom only if it gives access to items that satisfy this requirement, that is, above the threshold.

We characterize with two axioms a rule that reflects these arguments. As before, we denote X a finite set of options and $\mathscr{X} := 2^X \setminus \emptyset$ the collection of menus of options in X. Let $\langle \mathscr{T}, F, u, v \rangle$ be a minimal RS-structure defined on X and define F as in (3.2). Finally, $\succsim$ is a complete and transitive binary relation defined on $\mathscr{X}$.

To state our two axioms, we need to introduce the following definition. A menu A is **richer than** a menu B if for any $T \in \mathscr{T}$, if $T \cap F \cap A = \emptyset$, then $T \cap F \cap B = \emptyset$. So A is richer than B means that any type from which no element in F is available in A must also have no feasible options in $F \cap B$. Furthermore we say that A is **strictly richer than** B if A is richer than B but the reverse is not true.

Our first axiom says that (strictly) richer sets are always (strictly) preferred and imposes that it is an equivalence for singletons.

[23] It is actually a specific case of PX00's analysis where the equivalence classes induced by the similarity relation form a partition.

R-DOMINANCE.

(i) For any $A,B \in \mathscr{X}$: if A (strictly) richer than B, then $A(\succ) \succsim B$;

(ii) For any $x,y \in X$: $\{x\} \succ \{y\} \implies \{x\}$ strictly richer than $\{y\}$.

Note that part (i) of the axiom implies monotonicity in the sense of Kreps (1979): for any $A,B \in \mathscr{X}$, $A \supseteq B \implies A \succsim B$. Indeed, in this case, A is trivially richer than B. Part (ii) is an adaptation of Pattanaik and Xu (1990)'s *Indifference Between no Choice Situations*, which simply imposes an indifference between every singleton. They argue that singletons do not offer any freedom of choice, hence they cannot be strictly ranked. This is still true in our case, except if only one the two options is above the threshold of its type (i.e., in F), which is exactly what is implied by (ii).

Our second axiom is an adaptation of the composition axioms used in Pattanaik and Xu's series of papers.

R-COMPOSITION. *For any $A,B,C,D \in \mathscr{X}$, such that $A \cap C = B \cap D = \emptyset$, $C \subseteq T$ and $D \subseteq T'$ for some $T,T' \in \mathscr{T}$, and A is not richer than C: if $A \succsim B$ and $C \succsim D$, then $A \cup C \succsim B \cup D$.*

Combining menus that do not overlap should preserve the ranking. This is however true only if combining really provides additional freedom, which is captured by the requirement that A is not richer that C (see PX00 for a complete discussion of their axiom).

For any menu A, we define $\Phi(A) = \{\mathscr{T}(x) \cap F \cap A | x \in A\}$, the collection of subsets containing every option of one type that is above the threshold and available in A. We can now state our representation theorem (the proof is in Appendix C.3).

Theorem 6. *$\succsim$ satisfies* R-DOMINANCE *and* R-COMPOSITION *iff for any menu A*

and B:

$$A \succsim B \iff \#\Phi(A) \geq \#\Phi(B). \tag{3.3}$$

The interpretation is the following: what matters for the DM is to have access to more options, but only dissimilar objects — as captured by the distinct types — are valued. On top of that, within a certain class of similar options, the DM demands a minimal level of satisfaction to meet her freedom requirements, which is captured by the set F.

This measure is close to PX00's one. In addition to their representation, there is a role for preferences in this evaluation that is captured through the set F. Although PX98 also incorporate preferences, let us stress the key difference. Their starting point is a collection of possible preferences (i.e. complete and transitive orderings over the options) that a reasonable person may have. The resulting measure simply counts in a menu the number of options that are a maximiser of at least one of these preferences over the given menu. This approach integrates preferences relatively to a menu, simply attributing values to options that *could be* chosen in this menu. In contrast, in our approach, preferences are integrated in a more absolute way, in the following sense: below a certain level of satisfaction, even though the DM will have to choose an option, he does not attribute any freedom value to these potentially chosen items.[24] Even more, keeping the RSC in mind, some options that might be chosen later on, simply because of reactance, will not matter in the assessment of freedom, while some unchosen ones will matter.

[24] To illustrate this, our measure can be equal to 0 on some non-empty menus, which is impossible either in PX98 or in PX00.

3.6 Applications

We explore the scope of applicability of RSCs. We first show how our model can give plausible explanations to two observed and empirically supported phenomena: the formation of extreme beliefs — what we will call conspiracy theories — and the backfire effect of integration policies targeted toward minorities, two phenomena that have been related to reactance. We finally study the problem of a principal who must delegate a decision to a better-informed but biased agent who chooses according to an RSC.

3.6.1 Conspiracy Theories

As Sensenig and Brehm (1968) suggest, reactance has its counterpart in the realm of beliefs; namely the boomerang effect for psychologists (Hovland, Janis, and Kelley, 1953) or the backfire effect for political scientists (Nyhan and Reifler, 2010; Wood and Porter, 2019).[25] In the wake of Covid 19 pandemics, scholars argued such effects to be closely related to the formation of conspiracy theories and extreme beliefs (Adiwena, Satyajati, and Hapsari, 2020).[26] We propose to accommodate this mechanism by adapting Che and Mierendorff (2019)'s single period model of attention allocation with reactance.

A DM must choose from two actions, l or r, whose payoffs depend on an unknown state $i \in \{L,R\}$. His prior belief that the state is R is denoted p and we assume that $p \in (0,1/2]$. Before choosing his action, the DM acquires information. To that purpose, he can allocate his attention across four sources of information

[25] The boomerang effect is "a situation in which a persuasive message produces attitude change in the direction opposite to that intended". The backfire effect is a concept from political science that refers to a situation in which evidence contradicting the subjects' prior belief may reinforce their belief in the opposite direction.

[26] The fact that mass media did not give any credit to conspiracy theories has been pointed out as playing a role in reinforcing such theories through reactance.

(e.g. newspapers). Two of them are *L-biased* and the two others are *R-biased*.

The sources are represented by statistical experiments, or signals. The L-biased ones, denoted σ^{LL} and σ^{L}, can only reveal the state R. Symmetrically, the R-biased ones, denoted σ^{RR} and σ^{R}, can only reveal the state L. For $i = L,R$, σ^{ii} is an *extreme* source, whereas σ^{i} is a *moderate* one, i.e. the former is more biased than the latter. Formally, σ^{i} sends signal s^{i} with probability 1 in state i and with probability $1-\lambda$ in state $-i$, and σ^{ii} sends signal s^{i} with probability 1 in state i and with probability $1-\delta$ in state $-i$. We assume that $3/4 > \lambda > \delta = 1/2$. The experiments induced by the moderate sources σ^{L} and σ^{R} are described in table 3.1. The signals σ^{LL} and σ^{RR} are obtained by replacing λ with δ.

σ^{L}

State/signal	s^{L}	s^{R}
L	1	0
R	$1-\lambda$	λ

σ^{R}

State/signal	s^{L}	s^{R}
L	λ	$1-\lambda$
R	0	1

Table 3.1: Experiments induced by the moderate sources.

Initially the DM faces the complete menu $M = \{\sigma^{LL}, \sigma^{L}, \sigma^{R}, \sigma^{RR}\}$. In terms of our representation, the set of L-biased sources and the one of R-biased sources each represent a type of options. For $i = L,R$, σ^{i} is strictly more Blackwell informative than σ^{ii}, therefore the DM will never choose any of the extreme sources when his opportunity set is M, that is: $d(M) = \{\sigma^{L}, \sigma^{R}\}$. The DM's demands from freedom are satisfied when the moderate sources are available, that is,$\lambda_{L} \leq u(\sigma^{L})$ and $\lambda_{R} \leq u(\sigma^{R})$. When facing the menu M, the DM foresees that his payoff from choosing action $a \in \{l,r\}$ in state $i \in \{L,R\}$ is u_{a}^{i} where : $u_{r}^{R} = u_{l}^{L} = 1$, $u_{l}^{R} = u_{r}^{L} = -1$. Hence the DM will prefer action r if and only if his posterior belief is greater than 1/2. One can show that the DM's optimal allocation of attention is to choose the "own-biased news source"; namely the signal biased toward one's prior: in our case σ^{L} given that $p \leq 1/2$. The rationale for this is the following. The prior indicates

action l as the optimal one. Hence, a breakthrough signal s^R from σ^L is more valuable than a breakthrough signal s^L from σ^R. And the biased signal s^L from σ^L is more aligned with the DM's prior belief than s^R from σ^R. Hence, he is better off allocating his attention to σ^L (see Che and Mierendorff, 2019, pp. 2999-3000, for the complete argument).

In the next period, the moderate R-biased source σ^R is no more available, either because the government actually banned this newspaper or simply because the DM perceives that this source is no longer existing: only L-biased or extremely R-biased ones are present. The DM now faces the menu $N = \{\sigma^{LL}, \sigma^L, \sigma^{RR}\}$. He interprets this removal as revealing that the disutility from making a mistake in state L — i.e. choosing action r — is lower than expected: he now foresees it to be $v_r^L = 0$. σ^{RR} is no more removed from consideration by σ^R, hence $d(N) = \{\sigma^L, \sigma^{RR}\}$. His utility from choosing σ^L is unchanged while the one attached to σ^{RR} is $v(\sigma^{RR}) = p + (1-p)\delta$ (for p sufficiently close to 1/2 such that after signal s^R from σ^{RR}, the DM chooses action r).

As a consequence, some DMs with prior beliefs sufficiently close to 1/2, who would have chosen news source σ^L in menu M, will choose the extreme source σ^{RR} in menu N and their default option becomes r.

Proposition 11. *There exists $p^\star < 1/2$ such that if $p \in [p^\star, 1/2]$:*

(i) The DM prefers σ^{RR} to σ^L in menu N;

(ii) After a realisation of signal s^R from σ^{RR}, the DM chooses action r.[27]

This is in strong opposition as what would be obtained without reactance. Indeed, if the DM does not modify his utility when the menu shrinks, by removing σ^R, some DMs with prior belief strictly higher than 1/2 would now choose the source σ^L instead and action l after a signal s^L.

[27] All proofs of this section can be found in Appendix C.4.

3.6.2 Integration Policy Backlash

Can forced assimilation policy foster the integration of immigrants communities? While Alesina and Reich (2015)'s theory of nation building assumes that repressing the cultural practices of minorities spurs homogeneity, Bisin and Verdier (2001) suggest that the success of such policy may be mitigated by an increasing effort of parents to influence their children's cultural trait. In this application we show that, with reactance, one can even predict this policy to yield a backlash effect: the repressed immigrants react to repression by becoming more prompt to self-isolation. This additionally provides a rational to the persistence of reactance as an evolutionary efficient behavior.

Such a backslash effect has been recently documented by several papers. Some evidence suggests that the "burkha ban" in France in 2004 has strengthened the religious identity of French-Muslims (Abdelgadir and Fouka, 2020). Fouka (2020) shows that, in states which prohibited German Schools in the aftermath of World War I, German-Americans "were less likely to volunteer in World War II and more likely to marry within their ethnic group and to choose decidedly German names for their offspring".

To show how this backlash operates, we complement Bisin and Verdier (2001)'s account of cultural transmission with a reactance mechanism: as the repression increases, parents' educational freedom decreases and, reacting to this repression, they may endeavour to influence their children even more.[28] There are two cultural traits $\{m, M\}$ — for minority and Majority. The proportion of the minority q is assumed to be positive but lower than 1/2. Each generation is composed of parents who have only one child. Intergenerational transmission results from two socialization mechanisms. First, by vertical socialization the parents may

[28] For simplicity, we adopt a continuous setting, while our own framework is discrete. The ideas would be exactly the same with a discrete setting.

directly transmit their cultural trait i with probability d^i. If, with probability $1-d^i$, vertical socialization fails, then horizontal transmission occurs and the child adopts the traits of a random individual in society. Hence, the probability that a child from the minority be socialized by her parent's trait is:

$$P(d^i) \equiv d^i + (1-d^i)q. \tag{3.4}$$

As Bisin and Verdier (2001), we argue that parents endeavour to influence their child. They have a unit of time to allocate between their effort to fix d^i — which costs $(d^i)^\beta$ unit of time, with $\beta > 1$ — and a leisure activity $t^i \in [0,1]$, whose cost and utility are t^i. In addition, the government can implement a repressive policy $g^i \geq 1$ that may increase the parents' cost of influencing their child: a pair (t^i, d^i) costs $t^i + (d^i)^\beta g^i$ units of time for the parents. We posit that parents get a utility of 0 when their child is socialized to the other trait, while they get a utility $V(g^i)$ when she is socialized to their own trait. Hence, their expected utility of their child's socialization is $P(d^i)V(g^i)$. This means that, given a repressive policy, parents choose options $(t^i, d^i) \in [0,1]^2$ from the menu

$$K_{g^i} \equiv \{(t^i, d^i) : t^i + (d^i)^\beta g^i \leq 1\},$$

to maximize

$$t^i + P(d^i)V(g^i), \tag{3.5}$$

In what follows, we also assume that V has the following shape:

$$V(g) = \begin{cases} \hat{V} & \text{if } \hat{g} \geq g \\ \hat{V}\frac{g^\lambda}{\hat{g}} & \hat{g} < g \end{cases}$$

For some $\hat{g} > 1$ with $\lambda > 1$ and $\hat{V} > 1$. Hence, after a threshold $\hat{g}$, the more repressive is the policy g, the greater is $V(g)$. The interpretation is that parents react to the repressive policy when they feel that their freedom to educate their child is threatened. In other words, more repression may create incentives to dedicate more resources to transmit their traits to their children. Note that λ represents some kind of reactance rate since as it increases, parents' willingness to influence their child also increases.

From the first order condition, we obtain that the unique equilibrium educational effort — the program (3.5) being concave — must satisfy:

$$d^{i\star}(g^i, q) = \left(\frac{1-q}{\beta}\frac{V(g^i)}{g^i}\right)^{\frac{1}{\beta-1}} \tag{3.6}$$

Given the shape of V, $d^\star$ strictly decreases with g on $(1, \hat{g})$ and strictly increases with g on $(\hat{g}, +\infty)$. In other words, when the repressive policy exceeds $\hat{g}$, the more repression, the more parents invest in having their child socialized to their own trait. This suggests that reactance is at work in this model. In the following lemma, we establish the precise connection between this model and our reactance framework.

Lemma 1. *The function C defined on $\{K_g\}_{g\geq 1}$, such that for all g*

$$C(K_g) = \{(t,d) \in K_g : (t,d) \textit{ solves } (3.5)\}.$$

is a well-defined choice function and there exists an RSC C' defined on all compact subsets of $[0,1]^2$ such that $C(K_g) = C'(K_g)$ for all $g \geq 1$.[29]

Assuming the repressive policy to solely concern the minority (i.e. $g^M = 1$),

[29] For convenience, we construct an RS-structure on this infinite collection of compact sets. Obviously, analogous results could be obtained by making the set of possible policies g and the menus K_g finite.

what does reactance imply for the population dynamics in this model? Let time $\tau \in [0, +\infty)$ be continuous and q_τ be the share of the population with the minority cultural trait at time τ. Then, we have[30]

$$\dot{q} = q(1-q)\Big(d^{m\star}\big(g^m, q\big) - d^{M\star}\big(1, 1-q\big)\Big).$$

Given (3.6), d satisfies the *cultural substitution property*.[31] This implies that q converges to some $q^\star \in (0,1)$, which satisfies $d^{m\star}\big(g^m, q\big) = d^{M\star}\big(1, 1-q\big)$ (see Bisin and Verdier, 2001, Proposition 1). Hence,

$$q^\star(g^m) = \frac{V(g^m)/g^m}{V(1) + V(g^m)/g^m} \tag{3.7}$$

Given that $V(g)/g$ increases with g when $g \in (\hat{g}, +\infty)$ this means that repressive policy increases the size of the minority. This prediction contrasts with Alesina and Reich (2015)'s suggestions.

Noting that reactance is presumably a characteristic cultural trait (Jonas, Graupmann, Kayser, Zanna, Traut-Mattausch, and Frey, 2009), this model also provides a rationale for why reactance can be evolutionary efficient. Minorities which are more prompt to exhibit reactance are more likely to survive to repressive attempts to hinder their cultural practices.

To make precise this comparative statics statement, consider two minorities: one with a high reactance rate λ^H and one with a low reactance rate $\lambda^L < \lambda^H$. Denoting by $q_L^\star(.)$ and $q_H^\star(.)$ the equilibrium population share for these two minorities, the following proposition establishes that $q^\star$ is always higher for the high-reactance minority.

[30] See Bisin and Verdier (2001, equation (3), footnote 9) for discussions about this differential equation.

[31] In Bisin and Verdier (2001, Definition 1), this property states that d is continuous, decreasing with q, and $d = 0$ when $q = 1$.

Proposition 12. *For all* $g > \hat{g}$, $q_H^\star(g) > q_L^\star(g)$.

3.6.3 Optimal Delegation and Reactance

We consider a typical delegation problem: a principal can constrain the decision set of an informed but biased agent, but cannot commit to contingent monetary transfers (see Holmstrom, 1980; Alonso and Matouschek, 2008, for a detailed review of the literature). In any organization (administrations, companies, etc.), many rules govern what agents can or cannot do, with the purpose of reducing agency costs incurred by principals while benefiting as much as possible from better-informed agents. One can think for instance of a head of a company who delegates stock management to plant managers, a regulator who delegates pricing decisions to a monopolist with unknown costs, or a manager who delegates pricing decision to sales persons.

Formally, a *principal* (she) has the legal right to take an action among a finite set $A = \{a^{LL}, a^L, a^R, a^{RR}\}$. The payoffs delivered by each action depends on the realization of a binary state of the world $\theta \in \{L, R\}$. While the principal only knows the probability $p \in [0,1]$ that the state is R, an *agent* (he) is privately informed of the realization θ. The principal cannot use contingent transfers and must decide the set of actions among which the agent will choose.

Preferences. The principal's payoff for action a in state θ is the real number $\pi_\theta(a)$. Her preferred action is a^θ in state θ and her second favorite action $a^{\theta\theta}$. Her payoffs are written in table 3.2. The agent behaves according to an RS-structure with state-dependent utility and reaction functions. In both states, the types are $T^L = \{a^{LL}, a^L\}$ and $T^R = \{a^R, a^{RR}\}$ and $\lambda_L = u(a^{LL}), \lambda_R = u(a^{RR})$. The utility functions u_L, u_R and the reaction functions v_L, v_R are such that the agent reacts to the absence of $a^{\theta\theta}$ by choosing a^θ. In both states, he is more prone to restore the

absence of a^{RR}. The functions are specified in table 3.3.

Principal	R	$\pi_R(a^R) > \pi_R(a^{RR}) > \pi_R(a^L) > \pi_R(a^{LL})$
	L	$\pi_L(a^L) > \pi_L(a^{LL}) > \pi_L(a^R) > \pi_L(a^{RR})$

Table 3.2: Principal's Payoffs.

Optimal Delegation. Denote $\mathscr{A} = 2^A \setminus \emptyset$ the set of *menus* of action. For any $M \in \mathscr{A}$, $a_\theta(M)$ is the (unique) action chosen by the agent in state θ when facing menu M. For any prior belief $p \in [0,1]$, the objective of the principal is to solve the following maximization program, whose value is denoted $V(p)$:

$$V(p) \equiv \max_{M \in \mathscr{A}} (1-p)\pi_L\big(a_L(M)\big) + p\pi_R\big(a_R(M)\big). \tag{3.8}$$

A **delegation strategy** is a mapping from the set of beliefs to the set of menus: $\sigma : [0,1] \longrightarrow \mathscr{A}$. If for any p, $(1-p)\pi_L\big(a_L(\sigma(p))\big) + p\pi_R\big(a_R(\sigma(p))\big) = V(p)$, we say that the delegation strategy σ is **optimal**.

We are interested in the effect of RSC on optimal delegation strategies by the principal, and consequently on the agent's welfare. Without any reaction to restrictions, given that the agent's interest is sufficiently aligned with the principal's ($u_R(a^R) > u_R(a^L)$ and $u_L(a^L) > u_L(a^R)$), for any prior $p \in [0,1]$, the optimal delegation is to let the agent choose among the set of actions $\{a^L, a^R\}$.[32] This strategy cannot be optimal with the RSC because the agent would always choose a^R and therefore, for p sufficiently close to 0, offering a^L as the only possible action is better for the principal. For moderate p, it might be better to let the agent choose among the whole set of actions (or equivalently among his preferred actions $\{a^{LL}, a^{RR}\}$) given that in state $\theta = L, R$, $a^{\theta\theta}$ is the second best action for the

[32] Here we assume that the utility functions u_L and u_R would be the ones used if the agent was not responsive to restrictions. Of course, this is a slight abuse of what we can identify from choices given proposition 9.

Agent		
	R	$v_R(a^R) > v_R(a^L) > u_R(a^{RR}) > u_R(a^{LL}) > u_R(a^R) > u_R(a^L)$
	L	$v_L(a^R) > v_L(a^L) > u_L(a^{LL}) > u_L(a^{RR}) > u_L(a^L) > u_L(a^R)$

Table 3.3: Agent's Utility and reaction functions.

principal. It happens that it depends on the magnitude of the principal's payoffs, as summarized in proposition 13. Define the following beliefs:

$$\bar{p} = \frac{\pi_L(a^{LL}) - \pi_L(a^R)}{\pi_L(a^{LL}) - \pi_L(a^R) + \pi_R(a^R) - \pi_r(a^{RR})},$$
$$\underline{p} = \frac{\pi_L(a^L) - \pi_L(a^{LL})}{\pi_L(a^L) - \pi_L(a^{LL}) + \pi_R(a^{RR}) - \pi_R(a^L)},$$
$$\hat{p} = \frac{\pi_L(a^L) - \pi_L(a^R)}{\pi_L(a^L) - \pi_L(a^R) + \pi_R(a^R) - \pi_R(a^L)}.$$

Proposition 13. *An optimal delegation strategy $\sigma^\star$ must induce the following actions.*

1. *If $\underline{p} < \bar{p}$:*

   (i) $a_L(\sigma^\star(p)) = a_R(\sigma^\star(p)) = a^L$ *for* $p < \underline{p}$;

   (ii) $a_L(\sigma^\star(p)) = a^{LL}$ *and* $a_R(\sigma^\star(p)) = a^{RR}$ *for* $\underline{p} < p < \bar{p}$;

   (iii) $a_L(\sigma^\star(p)) = a_R(\sigma^\star(p)) = a^R$ *for* $p > \bar{p}$;

   and it can induce either of the two possibilities respectively at the boundary beliefs $\underline{p}$ and $\bar{p}$.

2. *If $\underline{p} \geq \bar{p}$:*

   (i) $a_L(\sigma^\star(p)) = a_R(\sigma^\star(p)) = a^L$ *for* $p < \hat{p}$;

   (ii) $a_L(\sigma^\star(p)) = a_R(\sigma^\star(p)) = a^R$ *for* $p > \hat{p}$;

   and it can induce either of the two possibilities at the boundary belief $\hat{p}$.

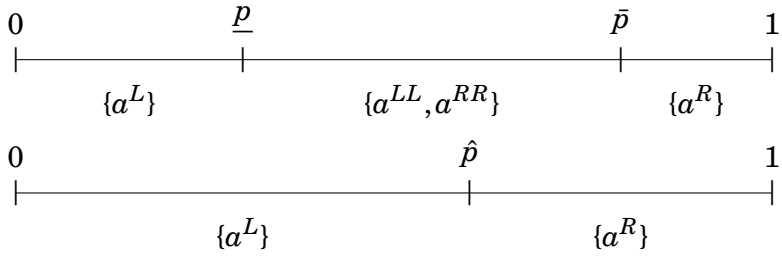

Figure 3.1: Optimal Delegation Strategies.

Two possible optimal strategies are depicted in figure 3.1, implementing the actions described in proposition 13. In each strategy, the principal is indifferent between the two possible menus at boundary beliefs $\underline{p}, \bar{p}$ and $\hat{p}$. These strategies are the most direct ones, in the sense that each menu does not contain irrelevant actions that are never chosen by the agent. The logic behind this result is that RSC makes the agent's threat to choose bad actions (for himself) credible. Hence, the principal reacts either by constraining even more the agent's opportunity set; or on the contrary by offering him a greater satisfaction.

Agent's Welfare. If we measure the agent's material welfare through the utility functions u_L and u_R, we see from proposition 13 that the effect of the RSC is ambiguous. In the case where $\underline{p} \geq \bar{p}$, the effect is only negative, as the agent only has access to a unique action that is not among her best actions. But if $\underline{p} < \bar{p}$, then while there is still this negative effect when $p \leq \underline{p}$ or $p \geq \bar{p}$, on the contrary, for middle beliefs, reactance forces the principal to let the agent choose among her best options $\{a^{LL}, a^{RR}\}$.

Chapter 4

Price Discrimination with Redistributive Concerns

Joint with Daniel M.A. Barreto and Victor Augias (Sciences-Po).

4.1 Introduction

Consumers are continuously leaving traces of their identities on the internet, be it through social media activity, search-engine utilization, online-purchasing and so on. The vast amount of consumer data that is generated and collected has acquired the status of a highly-valued good, as it allows firms to tailor advertisements and prices to different consumers. In practice, the availability of consumer data **segments** consumers: observing that a given consumer has certain characteristics allows firms to fine-tune how they interact with people that share those characteristics. Adjusting how coarse-grained the information available about consumers is thus impacts how they will be segmented, what sort of digital market interactions they will have and what prices they will pay. This suggests room for regulatory oversight.

As shown by Bergemann, Brooks, and Morris (2015) (henceforth BBM), consumer segmentation and price discrimination can induce a wide range of welfare outcomes. It can not only be used to increase economic surplus—by creating segments with prices that allow more consumers to buy—, but can also be performed in a way that ensures that all created surplus accrues to consumers — that is, that maximizes consumer surplus. This is done by creating segments that pool together consumers with high and low willingness to pay, thus allowing higher willingness to pay consumers to benefit from lower prices. This finding suggests that implementing such consumer-surplus maximizing segmentations could be the target of data regulation designed to prioritize consumer welfare. However, an important aspect of price discrimination that remains overlooked by the literature is its **distributive effect**: since different consumers pay different prices, this practice defines how surplus is distributed *across* consumers, raising questions about how it can benefit poorer consumers relative to richer ones. Indeed, if willing-

ness to pay and wealth are positively related, then consumer-surplus maximizing segmentations tend to benefit richer consumers.

In this paper we provide a normative analysis of the distributive impacts of market segmentation. Our aim is to study how this practice impacts different consumers and how it should be performed under the objective of increasing consumer welfare while prioritizing poorer consumers. Our results draw qualitative characteristics of segmentations that achieve this goal, which can be used to guide future regulation. Importantly, our analysis also shows that the prioritization of poorer consumers can be inconsistent with the maximization of total consumer surplus: raising the surplus of poorer consumers may only be possible while granting additional profits to the producer.

We consider a setting in which a monopolist sells a good on a market composed of heterogeneous consumers, each of whom can consume at most one unit and is characterized by their willingness to pay for the good. A social planner can provide information about consumers' willingness to pay to the monopolist. The information provision strategy effectively divides the aggregate pool of consumers into different **segments**, each of which can be priced differently by the producer. The social planner's objective is to maximize a weighted sum of consumers' surplus. As in Dworczak, Kominers, and Akbarpour (2021), we consider weights that are decreasing on the consumer's willingness to pay, capturing the notion of a redistributive motive under the assumption that consumers with higher willingness to pay are on average richer than those with lower willigness to pay.

We first establish that, as long as the social weights are non-negative, maximizing the social planner's redistributive objective never comes at the cost of efficiency, that is, never sacrifices total achievable social surplus. Therefore, the redistribution of surplus among consumers through market segmentation never implies any deadweight loss. We then show that, for some aggregate markets

of consumers, a sufficiently strong redistributive objective cannot be met while still maximizing total consumer surplus. Indeed, in the process of redistributing surplus from richer to poorer consumers, some of the surplus might "leak" to the monopolist in the form of additional profits. We characterize the set of markets for which this is the case and denote them as rent markets. For no-rent markets, on the contrary, we show that *any* redistributive objective can be met while still maximizing total consumer surplus. In this case, our analysis selects one among the infinitely-many consumer surplus maximizing segmentations established by BBM. These insights are illustrated through a three-type example in section 4.3.

Our analysis also provides insights on how to construct optimal segmentations. We show that, in no-rent markets, optimal redistributive segmentations exhibit a stunningly simple form: they simply divide consumers into one discount segment and one residual segment. The discount segment pools together all consumers who would not consume and some who would consume but would not get any surplus under the uniform price, whereas the residual segment pools all of the remaining consumers. In rent markets, we show how that optimal segmentations under sufficiently strong redistributive preferences (SRP) divide consumers into contiguous segments based on their willingness to pay, having consumers with the same WTP belong to at most two different segments. This allows us to construct a procedure that generates SRP-optimal segmentations, which is discussed in section 4.4.2.

Related Literature. Third-degree price discrimination and its welfare effects are the subject of an extensive literature, with early analysis dating back to Pigou (1920) and Robinson (1933). Subsequent literature (Schmalensee, 1981; Varian, 1985) has studied under what conditions a market segmentation increases total surplus.

More recently, a literature incorporating an information design approach has revisited the question of welfare impacts of third-degree price discrimination by analyzing all feasible segmentations of a market. BBM analyze a setting with a monopolist selling a single good and characterize attainable pairs of consumer and producer surplus, showing that any distribution of total surplus over consumers and producer that guarantee and least the uniform-price profit for the producer is attainable. Their analysis has been extended to multi-product settings by Haghpanah and Siegel (2022a,b); the authors establish that any inefficient market can be Pareto improved by a two-market segmentation. Elliott, Galeotti, Koh, and Li (2021) and Ali, Lewis, and Vasserman (2022) extend the analysis for imperfect competition settings, emphasizing the interaction between the information design and competition. Hidir and Vellodi (2020) study market segmentation in a setting where the monopolist can offer one from a continuum of goods to each consumer, such that consumers, upon disclosing their information, face a trade-off between being offered their best option and having to pay a fine-tuned price. Finally, Roesler and Szentes (2017) and Ravid, Roesler, and Szentes (2022) study the inverse problem of information design to a buyer who is uncertain about the value of a good.

Our paper also dialogues with a recent literature on mechanism design and redistribution, most notably with Dworczak et al. (2021) and Akbarpour, Dworczak, and Kominers (2020). The paper closest in spirit to ours is Dube and Misra (2022), who study experimentally the welfare implications of personalized pricing implemented through machine learning. The authors find a negative impact of personalized pricing on total consumer surplus, but note that a majority of consumers benefit from price reductions under personalization, pointing that under some inequality-averse weighted-welfare functions, personalization increases welfare. Our analysis in this paper provides a theoretical foundation on the use of price

discrimination as a tool for redistribution and helps to understand their results.

4.2 Model

A monopolist (he) sells a good to a continuum of mass one of buyers, each of whom can consume at most one unit. We normalize the marginal cost of production of the good to zero. The consumers privately observe their type v, which represents their willingness to pay for the good, and which can take K possible values $\{v_1,\dots,v_K\} \equiv V$, where:

$$0 < v_1 < \cdots < v_K.$$

A *market* μ is a distribution over the valuations and we denote the set of all markets:

$$M \equiv \Delta(V) = \left\{ \mu \in \mathbb{R}^K \;\middle|\; \sum_{k=1}^{K} \mu_k = 1 \text{ and } \mu_k \geq 0 \text{ for all } k \in \{1,\dots,K\} \right\}.$$

We say that a price v_k is **optimal for market** $\mu \in M$ if it maximizes the expected revenue of the monopolist when facing market μ[1], that is:

$$v_k \sum_{i=k}^{K} \mu_i \geq v_j \sum_{i=j}^{K} \mu_i, \quad \forall j \in \{1,\dots,K\}.$$

Let M_k denote the set of markets where price v_k is optimal:

$$M_k = \left\{ \mu \in M \;\middle|\; v_k \in \arg\max_{v_i \in V} v_i \sum_{j=i}^{K} \mu_j \right\}.$$

In the remaining of the paper we will hold the aggregate market fixed and denote it by $\mu^\star \in M$.

[1]Note that we can restrict the action set of the monopolist to be equal to V, since any price $p \notin V$ is dominated by some $v \in V$.

Segmentation. The consumers' types are perfectly observed by a social planner (she) who can **segment** consumers, that is, a divide the aggregate market into different (sub-)markets. Formally, a segmentation is a simple probability distribution on M which averages to the aggregate market $\mu^\star$. The set of possible segmentations of a given aggregate market $\mu^\star$ is:

$$\Sigma(\mu^\star) \equiv \left\{ \sigma \in \Delta(M) \,\middle|\, \sum_{\mu \in \text{supp}(\sigma)} \mu\,\sigma(\mu) = \mu^\star, |\text{supp}(\sigma)| < \infty \right\}.^{2}$$

Given a segmentation σ, the monopolist is able to price differently at each segment μ in the support of σ. As will become clear in section 4.4, segments with more than one optimal price play a key role in our results, such that we focus on the following pricing rule $p : M \to V$ applied by the monopolist:

$$p(\mu) = \min \left\{ \arg\max_{k \in \{1,\dots,K\}} v_k \sum_{i=k}^{K} \mu_i \right\}.$$

That is, the monopolist charges at each segment the smallest among his optimal prices in that segment. This pricing rule is chosen simply for equilibrium selection purposes, and our results still hold qualitatively if the monopolist selects among optimal prices in some other way.

We can therefore write the *utility* of a consumer of type v_k in market μ as:

$$U_k(\mu) \equiv \max\left\{0, v_k - p(\mu)\right\}.$$

Social objective. The social planner's objective is to maximize a weighted sum of consumers' surplus, with positive weights $\lambda \in \mathbb{R}_+^K$. Her preferences exhibit aversion to inequality by putting greater weight on surplus extracted by poorer consumers. By making the simple assumption that consumers with lower willingness to pay

[2] Where $\text{supp}(p)$ is the support of a distribution p.

are also on average poorer, this is captured through weights that are decreasing in v_k.[3] For a given market μ, the weighted total consumers' surplus is given by:

$$W(\mu) \equiv \sum_{k=1}^{K} \lambda_k \mu_k U_k(\mu).$$

Focusing on the pricing rule p makes the function W upper semi-continuous. Hence, for any aggregate market $\mu^\star$, the social planner's objective is given by the following well-defined maximization program, whose value is denoted $V(\mu^\star)$:

$$\max_{\sigma \in \Sigma(\mu^\star)} \sum_{\mu \in \text{supp}(\sigma)} \sigma(\mu) W(\mu). \tag{S}$$

Given an aggregate market $\mu^\star$, a segmentation $\sigma \in \Sigma(\mu^\star)$ is **optimal** if

$$\sum_{\mu \in \text{supp}(\sigma)} \sigma(\mu) W(\mu) = V(\mu^\star).$$

Efficiency. Every consumer has a value for the good that is strictly greater than the marginal cost of production. Hence a market μ is **efficient** if every consumer can buy the good, that is, if the lowest optimal price for the seller allows everyone to consume: $p(\mu) = \min \text{ supp}(\mu)$. A segmentation σ is **efficient** if it is only supported on efficient markets.

Consumer Surplus Maximizing Segmentations. If $\lambda = (1, \ldots, 1)$, program (S) maximizes total (or average) consumer surplus over all possible segmentations. A segmentation that solves this optimization problem is named **CS-maximizing**. We have known since BBM that, in this case, the optimal segmentation (i) is

[3] We follow here the approach by Dworczak et al. (2021) and refer the readers to their work for a detailed justification.

efficient—and hence achieves the maximum feasible social surplus[4]—, and (ii) does not give the monopolist any additional profit as compared to the situation where he must set a uniform price for the whole aggregate market.

Typically, for an interior aggregate market $\mu^\star$, there exists infinitely many CS-maximizing segmentations. In section 4.4.3, we characterize the set of aggregate markets for which the optimal (redistributive) segmentation is also CS-maximizing, thus providing a natural way to select among CS-maximizing segmentations for these markets.

Informational Rents. We say that a segmentation σ generates a **rent** to the monopolist if, under the segmentation, the monopolist's profit is strictly greater than the profit he would have under the uniform price. Formally, let's define the monopolist profit at a given market μ as:

$$\pi(\mu) = p(\mu) \sum_{k \in C(p(\mu))} \mu_k$$

where $C(p) = \{k \in \{1,\dots,K\} | v_k \geq p\}$ is the set of consumer types that buy the good given a price p. We denote the profit of the monopolist under a given segmentation σ as:

$$\Pi(\sigma) = \sum_{\mu \in \text{supp}(\sigma)} \sigma(\mu)\pi(\mu)$$

We know from BBM that market segmentations can only weakly increase the profit of a monopolist, that is, $\Pi(\sigma) \geq \pi(\mu^\star),\ \forall \sigma \in \Sigma(\mu^\star)$. Whenever this inequality holds strictly for a given σ, we say that this segmentation induces a rent to the monopolist.

[4] For a given market μ, the maximum feasible social surplus is given by

$$s(\mu) = \sum_k \mu_k v_k.$$

Note that a segmentation of μ achieves $s(\mu)$ if and only if it is efficient.

Discussion of the Model

Information Provision as Segmentation. In digital markets, information provision about consumers often occurs through the assignment of *labels* to different consumers. Indeed, one could think of a model in which the social planner adopts a labeling strategy $\Psi : V \to \Delta(L)$, where L is the set of labels that she distributes. The meaning of each label is then pinned down by the social planner's strategy, and the monopolist optimally chooses different prices for consumers with different labels.

Such a model is equivalent to ours. Indeed, any segmentation $\sigma \in \Sigma(\mu^\star)$ can be implemented by some labeling strategy Ψ, and any labeling strategy Ψ implements some segmentation $\sigma \in \Sigma(\mu^\star)$. The approach of working directly in the space of feasible distributions over distributions of types, rather than in the space of distributions of signals, is standard in the information design literature (Kamenica, 2019).

Continuum of Consumers. While we consider a setting with a continuum of consumers, our model is equivalent to one in which there is a discrete number of consumers, with types independently distributed according to $\mu^\star$.

4.3 Three-Value Case

In this section, we illustrate our model and some of the results from the following sections in the simple three-value case. Let's consider three types, $V = \{1,2,3\}$, such that $K = 3$ and $v_k = k$. We can conveniently depict the set of markets (i.e distributions over V) M as the two-dimension probability simplex (see Mas-Colell, Whinston, and Green, 1995, p.169). It is represented by an equilateral triangle, as depicted in figure 4.1, where each vertex represents a degenerate market on a

value $v \in V$, denoted by the Dirac measure δ_v.

In the left panel of figure 4.1 are drawn the three regions where the different prices $\{1,2,3\}$ are set according to the pricing rule p.[5] In the right panel, an aggregate market $\mu^\star = (0.3, 0.4, 0.3)$ is represented, which is in the interior of the region M_2, meaning that v_2 is a strictly optimal price for $\mu^\star$. Two possible segmentations are depicted: the one in green dashed lines, that segments $\mu^\star$ into the three degenerate markets (which implements first-degree price discrimination); and the one in black dotted lines, that splits $\mu^\star$ into three markets μ', μ'' and μ'''. Any splitting of $\mu^\star$ into a set of points $S \subset M$ represents a feasible segmentation, as long as $\mu^\star \in co(S)$.[6] The remaining of this section is devoted to constructing segmentations that are optimal given different weights $(\lambda_1, \lambda_2, \lambda_3)$ that the social planner might have, with $\lambda_1 \geq \lambda_2 \geq \lambda_3$. Note that consumers of type v_1 never get any consumer surplus (since the monopolist never charges a price lower than their willingness to pay), such that the optimal segmentation trades-off surplus obtained by types v_2 and v_3. We will focus, without loss of generality, on direct segmentations, i.e. segmentations in which there is not more than one segment with a given price.

A first step for finding the optimal segmentation of $\mu^\star$ is to observe that any optimal segmentation must be efficient. To see that, consider the black dotted segmentation in the right panel of figure 4.1. Both μ' and μ'' are efficient, since all the consumers in these segments are able to buy the good. The remaining segment μ''', however, is not efficient, as it contains some consumers with type v_1 and v_2 who are not able to consume under that segment's price. One could solve that by re-segmenting μ''' in the following way: creating a segment μ'''_b containing all of the types v_1 and v_2 and some of the types v_3 that used to belong to μ''', and another

[5] Formally, for any k, $M_k = \text{cl}(p^{-1}(v_k))$, where $\text{cl}(S)$ denotes the topological closure of a generic set S.

[6] For any set S, $co(S)$ denotes the convex hull of S

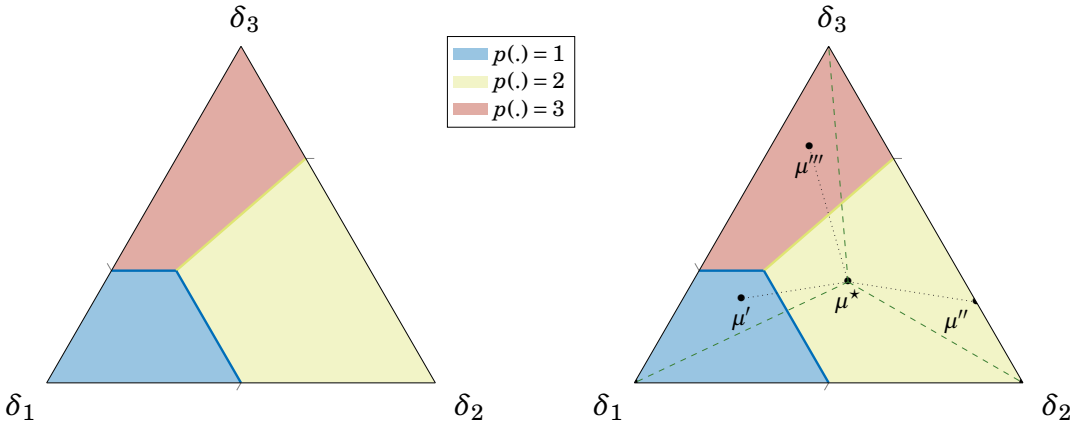

Figure 4.1: The Simplex representing M and two feasible segmentations.

segment δ_3 containing only the remaining types v_3. Note that the amount of type v_3 in μ_b''' can be adjusted to ensure that this segment will have price v_1. That way, both of the resulting segments will be efficient. Furthermore, this re-segmentation of μ''' *unambiguously* increases consumer welfare, since it has no impact on the welfare of consumers in μ' and μ'' and (weakly) increases the surplus of every consumer previously belonging to μ'''.

Indeed, a welfare-increasing segmentation can be performed to any inefficient market. This narrows down the search for an optimal segmentation, as we know that it must be supported *only* on efficient segments. The left panel of figure 4.2 depicts, in orange, the efficient markets. These are: the degenerate market δ_3; the set of markets in region M_2 that have no consumer with value 1; and the entire region M_1.

We can further narrow down our search by noting that, in an optimal segmentation, the segment with price v_1 must not belong to the interior of region M_1. To see that, consider the right panel of figure 4.2. In it are depicted two segmentations: σ_a, which splits $\mu^\star$ into μ_a and μ', and σ_b, which splits $\mu^\star$ into μ_b and μ'. Segmentation σ_b is always preferred over σ_a for two reasons. First, μ_b has a higher share of types v_2 and v_3 than μ_a. Since these are the only two types that

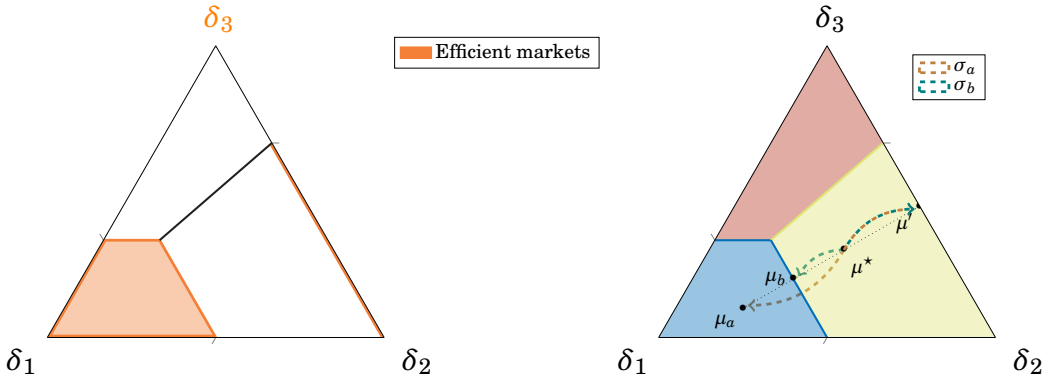

Figure 4.2: Efficient Markets and Segmentations.

are extracting surplus on the segment whose price is v_1, having a higher share of them increases the social planner's objective. Second, μ_b is "closer" to $\mu^\star$, which means that $\sigma_b(\mu_b) > \sigma_a(\mu_a)$. That means that segmentation σ_b is able to include a bigger mass of consumers in the segment where they will extract the largest surplus, thus also increasing the social planner's objective.

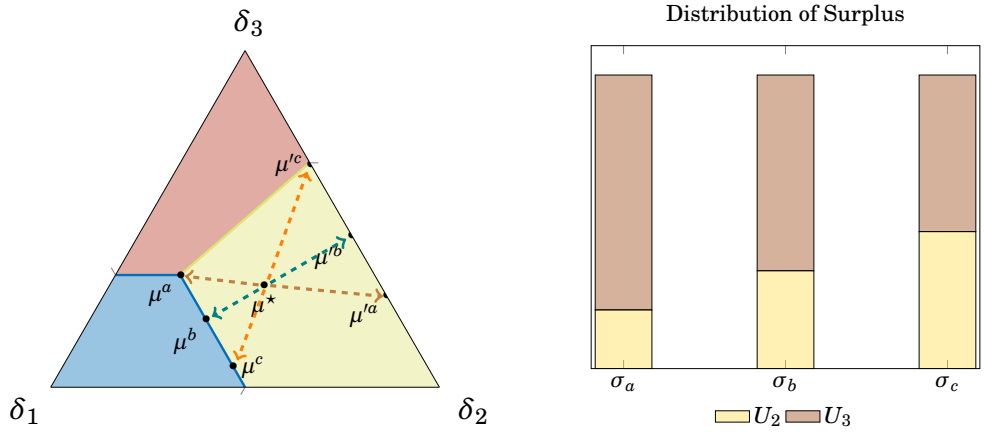

Figure 4.3: Distribution of Surplus among CS-Maximizing Segmentations.

The argument outlined above illustrates how every segmentation generating a segment on the interior of region M_1 must be dominated by some segmentation that instead generates a segment on the boundary of regions M_1 and M_2. This amounts to saying that any optimal segmentation must include a segment in which

the monopolist is indifferent between charging price v_1 or charging some other price. The intuition for that is simple: if the monopolist strictly prefers to charge price v_1 in that segment, then there's still room for "fitting" other types in that segment in a Pareto improving way.

Having established these properties of optimal segmentations, we can now compare some candidate policies. Figure 4.3 depicts three different segmentations that are all efficient and generate one segment on the boundary between M_1 and M_2. Two things should be noted when comparing the three segmentations. First, that $\sigma_a(\mu_a) = \sigma_b(\mu_b) = \sigma_c(\mu_c)$ and $\sigma_a(\mu'_a) = \sigma_b(\mu'_b) = \sigma_c(\mu'_c)$, such that all three segmentations include the same mass of consumers in the segment with price v_1 and in the segment with price v_2. Thus, the three segmentations differ only in the composition of types in each of these segments.

Second and most importantly, it should be noted that all three segmentations in figure 4.3 are CS-maximizing. This follows from the fact that i) they maximize total (consumer + producer) surplus, since they are all efficient, and ii) the monopolist does not get any of the surplus that is created from the segmentation [7].

Indeed, there are uncountably many CS-maximizing segmentations of $\mu^\star$. For a social planner with $\lambda_1 = \lambda_2 = \lambda_3$, all of these are optimal segmentations. For a social planner with $\lambda_2 > \lambda_3$, however, they are not equivalent. As the right panel of figure 4.3 shows, segmentations σ_a, σ_b and σ_c differ in how they distribute consumer surplus among different consumers. In particular, σ_c distributes a larger share of consumer surplus to consumers of type v_2, as μ^c has a higher share of these types than μ^a or μ^b. Therefore, a social planner valuing the surplus of types

[7] One way of seeing this is as follows: A decision-maker strictly benefits from observing a piece of information if, as a result of this observation, she is able to make better decisions than she would have made absent this information. In our setting, this amounts to the monopolist being able to, as a result of the segmentation, choose *different* prices than the uniform price, at markets in which these different prices are *strictly* preferred over the uniform price. Since price v_2 belongs to the set of optimal prices in every segment generated by the segmentations in figure 4.3, the monopolist does not strictly benefit from them.

v_2 more than that of types v_3 prefers σ_c over the two other segmentations depicted in figure 4.3.

Can σ_c be improved upon? One potential way of doing so is to further increase the share of type v_2 in μ^c, which could be achieved by exchanging the remaining types v_3 in μ_c against types v_2 present in μ'^c. While such an exchange increases the surplus of types v_2, since now more of them pay a price of v_1, it also has an even stronger negative impact on the surplus of types v_3, as now there would be sufficiently many of them in segment μ'_c for the monopolist to want to increase the price in that segment. That would lead to a segmentation that is no longer CS-maximizing, and that instead grants additional profits to the monopolist. The result below establishes when this exchange is desirable from the social planner's perspective.

Result 1. *Let* $\mu^\star = (0.3, 0.4, 0.3)$*:*

1. *for* $\frac{v_2 - v_1}{v_3 + v_2 - v_1} < \frac{\lambda_3}{\lambda_2} < 1$ *the optimal segmentation is CS-maximizing and generates two segments: one containing types* $\{v_1, v_2, v_3\}$ *and the other one only containing types* $\{v_2, v_3\}$*. This segmentation is represented in the left panel of figure 4.4;*

2. *for* $0 < \frac{\lambda_3}{\lambda_2} < \frac{v_2 - v_1}{v_3 + v_2 - v_1}$*, the optimal segmentation is not CS-maximizing and generates three segments: the first containing types* $\{v_1, v_2\}$*, the second containing types* $\{v_2, v_3\}$ *and the third containing only types* $\{v_3\}$*. This segmentation is represented in the right panel of figure 4.4.*

An important consequence of this result is that if the social planner's preferences are sufficiently redistributive, the optimal segmentation might give a *rent* (i.e. an additional profit) to the monopolist. By packing more consumers with lower types together, the social planner also makes higher types more distinguishable, thus allowing the monopolist to raise their prices. The above example illustrates

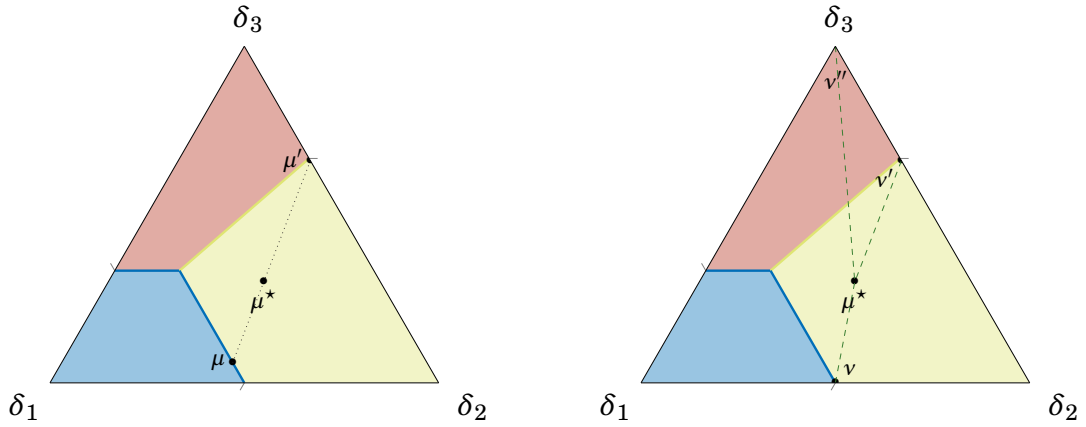

Figure 4.4: Optimal Segmentations.

the main argument of the paper: while personalized pricing can redistribute surplus without any loss of efficiency, sometimes part of the surplus created in the process might need to go to the monopolist.

However, the necessity of granting rents to the monopolist in order to satisfy redistributive objectives is not true for any aggregate market. Consider for instance the aggregate market $\mu^\star = (0.2, 0.65, 0.15)$, represented in the left panel of figure 4.5. The optimal segmentation of this market given **any** preferences $\lambda_2 \geq \lambda_3$ is the one depicted in the figure: it always generates a segment with $\{v_1, v_2\}$ and another one with $\{v_2, v_3\}$, and this segmentation is always CS-maximizing. Satisfying a redistributive objective never requires granting rents to the monopolist because this aggregate market contains sufficiently many consumers of type v_2, such that even after pooling as many as possible of them with types v_1 in segment μ, there are still sufficiently many types v_2 left to ensure that types v_3 will not be too distinguishable in segment μ'.

The result below characterizes the set of aggegate markets that, under a sufficiently strong redistributive motive, would require granting rents to the monopolist. We denote this set as the **rent region**.

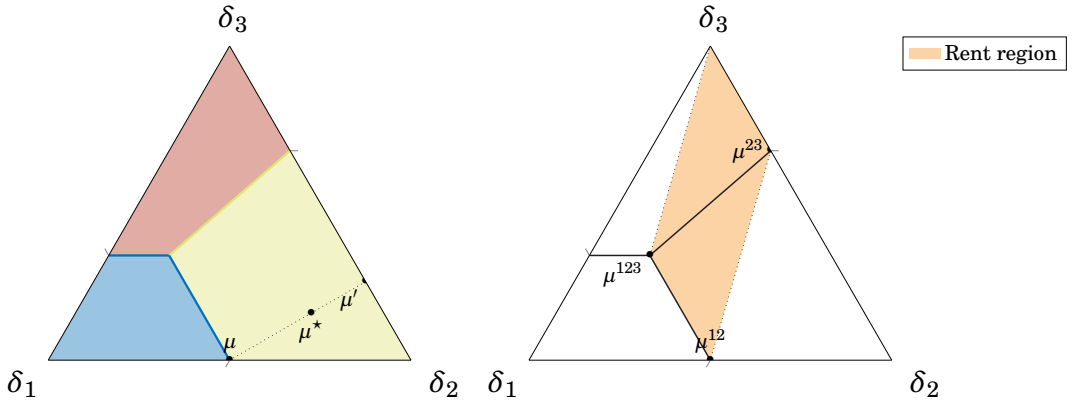

Figure 4.5: Rent Region.

Result 2. *The rent region is:*

$$\text{Int}\Big(\text{conv}\big(\{\delta_3, \mu^{123}, \mu^{12}, \mu^{23}\}\big)\Big).^{8}$$

This result is illustrated in the right panel of figure 4.5, where the rent region is depicted in orange. Equivalently, the complement of this set denotes the aggregate markets for which any redistributive objective can be met without granting rents to the monopolist — that is, while maximizing total consumer surplus—. We call this set the **no-rent region**.

The following section generalizes the insights presented through this example. Section 4.4.1 generalizes the fact that optimal segmentations are efficient and include discount segments supported at markets at which the monopolist is indifferent between more than one price, while section 4.4.2 establishes properties of optimal segmentations when the redistributive motive is sufficiently strong and shows how to construct optimal segmentations in this case. Finally, section 4.4.3 characterizes generally the no-rent and rent regions and shows that optimal segmentations for markets belonging to the no-rent region exhibit a very simple form,

[8] Int(S) and conv(S) are respectively the *interior* and the *convex hull* of the set S.

with only one discount segment and one uniform price segment.

4.4 Optimal Segmentations

We now turn to the analysis of the general case. In section 4.4.1 we derive general properties of optimal segmentations — that is, characteristics that are present in optimal segmentations given any decreasing weights $\boldsymbol{\lambda}$ that the social planner might have. Section 4.4.2 then constructs optimal segmentations under strongly redistributive preferences: when the weight assigned to lower types is sufficiently larger than the weight assigned to higher types. Finally, section 4.4.3 characterizes the set of aggregate markets for which satisfying a redistributive objective might require granting additional profits to the monopolist.

4.4.1 General Properties

Our first result echoes our analysis of efficiency in the three-value case and establishes that i) we can always restrict ourselves to efficient segmentations—as long as the weights are non-negative; ii) if the weights are all strictly positive (i.e. if $\lambda_K > 0$ under our assumption of decreasing weights), only efficient segmentations can be optimal.

Proposition 14. *For any aggregate market $\mu^\star$ and any weights $\boldsymbol{\lambda} \in \mathbb{R}^K_+$ (not necessarily decreasing), there exists an efficient optimal segmentation of $\mu^\star$. Furthermore, if every weight is strictly positive ($\boldsymbol{\lambda} \in \mathbb{R}^K_{++}$), any optimal segmentation is efficient.*

This result is a direct consequence of Proposition 1 in Haghpanah and Siegel (2022b)—which itself follows from the proof of Theorem 1 in BBM. Indeed, their result states that any inefficient market can be segmented in a Pareto improving manner, that is, in a way that weakly increases the surplus of all consumers. This implies that, as long as the social planner does not assign a negative weight to any consumer, there must be an efficient optimal segmentation. As a consequence, the

social planner's redistributive objective never comes at the expense of efficiency.

Direct segmentation. A segmentation σ is **direct** if all segments in σ have different prices, that is, for any $\mu, \mu' \in \text{supp}(\sigma)$, $p(\mu) \neq p(\mu')$. Our next lemma shows that it is without loss of generality to focus on direct segmentations.

Lemma 2. *For any aggregate market $\mu^\star$ and any segmentation $\sigma \in \Sigma(\mu^\star)$, there exists a direct segmentation $\sigma' \in \Sigma(\mu^\star)$ such that,*

$$\sum_{\mu \in \text{supp}(\sigma)} \sigma(\mu) W(\mu) = \sum_{\mu \in \text{supp}(\sigma')} \sigma'(\mu) W(\mu).$$

We further show that there always exists an optimal and direct segmentation that is only supported on the boundaries of price regions $\{M_k\}_k$. For this, denote for any aggregate market $\mu^\star$, $I(\mu^\star) \equiv \{k \mid v_k \in \text{supp}(\mu^\star)\}$, the set of indices k such that v_k is in the support of $\mu^\star$.

Lemma 3. *For any aggregate market $\mu^\star$ that is not efficient, there exists an optimal direct segmentation supported on boundaries of sets $\{M_k\}_{k \in I(\mu^\star)}$.*

4.4.2 Strongly Redistributive Social Preferences

In this section, we derive some qualitative characteristics of the optimal segmentation when the social planner's preferences are *strongly redistributive*, that is, when the weights $\boldsymbol{\lambda}$ are strongly decreasing on the type v.

Definition 12. *The weights $\boldsymbol{\lambda}$ are κ-**strongly redistributive** if, for any $k < k' \leq K-1$, $\frac{\lambda_k}{\lambda_{k'}} \geq \kappa$.*

That is, a social planner exhibits κ-strongly redistributive preferences (κ-SRP) if the weight she assigns to a consumer of type v_k is at least κ times larger than the weight she assigns to any consumer of type greater than v_k.

Before stating the main result of this section, let us formally define the *dominance* ordering between any two sets.

Definition 13. *Let $X, Y \subset \mathbb{R}$, X **dominates**[9] Y, denoted $X \geqslant_D Y$, if for any $x \in X$ and any $y \in Y$, $x \geq y$.*

Proposition 15. *For any aggregate market $\mu^\star$ in the interior of M, there exists $\underline{\kappa}$ such that if $\boldsymbol{\lambda}$'s are $\underline{\kappa}$-strongly redistributive, then for any optimal direct segmentation $\sigma \in \Sigma(\mu^\star)$ and any markets $\mu, \mu' \in \text{supp}(\sigma), \mu \neq \mu'$: either $\text{supp}(\mu) \geqslant_D \text{supp}(\mu')$ or $\text{supp}(\mu') \geqslant_D \text{supp}(\mu)$.*

The result stated above establishes that, when the social planner's preferences exhibit a sufficiently strong taste for redistribution, optimal segmentations divide the type space V into contiguous overlapping intervals, with the overlap between any two segments being composed of at most one type. The following corollary is a direct consequence of proposition 15:

Corollary 4. *For any aggregate market $\mu^\star$ in the interior of M, there exists $\underline{\kappa}$ such that if $\boldsymbol{\lambda}$'s are $\underline{\kappa}$-strongly redistributive, then for any optimal direct segmentation $\sigma \in \Sigma(\mu^\star)$, any market $\mu \in \text{supp}(\sigma)$ and any k such that $\min\{\text{supp}(\mu)\} < v_k < \max\{\text{supp}(\mu)\}$: $\sigma(\mu)\mu_k = \mu^\star_k$.*

The above result states that any segment μ belonging to a segmentation that is optimal under sufficiently strong redistributive preferences contains *all* of the consumers with types strictly in-between $\min\{supp(\mu)\}$ and $\max\{supp(\mu)\}$. Together with proposition 15, it implies that, under κ-SRP optimal segmentations, every consumer type v will belong to *at most* two segments: either it will belong to the interior of the support of a segment μ, such that all consumers of this type have surplus $v - \min(supp(\mu))$, or it will be the boundary type between two segments μ

[9] Note that this definition of dominance is stronger than the notion of dominance in the strong set order (Topkis, 1998).

and μ', such that a fraction of these consumers (those belonging to segment μ) gets surplus $v - \min(supp(\mu))$ and the rest gets no surplus.

These results along with proposition 14 completely pin down the κ-SRP optimal direct segmentation. One can construct it by employing the following procedure, presented as follows through steps:

- **Step i)** Start by creating a segment — call it μ_a — with all consumers of type v_1.
- **Step ii)** Proceed to including in μ_a, successively, all consumers of type v_2, then all of the types v_3, and so on. From proposition 14 we know that μ_a must be efficient, meaning that we must have $p(\mu_a) = v_1$. As such, the process of inclusion of types higher than v_1 must be halted at the point in which adding a new consumer in μ_a would result in v_1 no longer being an optimal price in this segment. We denote as $v_{(a|b)}$ the type that was being included when the process was halted.
- **Step iii)** Create a new segment — call it μ_b — with all of the remaining types $v_{(a|b)}$.
- **Step iv)** Proceed to including in μ_b, successively, all of consumers of type $v_{(a|b)+1}$, then all of the types $v_{(a|b)+2}$, and so on. Halt this process at the point in which adding a new consumer in μ_b would result in $v_{(a|b)}$ no longer being an optimal price in this segment. We denote as $v_{(b|c)}$ the type that was being included when the process was halted.
- **Step v)** Create a new segment with all of the remaining types $v_{(b|c)}$. Repeat the process described in the last steps until every consumer has been allocated to a segment.

4.4.3 Optimal Segmentations and Informational Rents

This section explores the question of when does an optimal segmentation maximize total consumer surplus or, conversely, when it induces a rent for the monopolist.

Say that an aggregate market $\mu^\star$ belongs to the **rent region** if there exists some $\underline{\kappa}$ such that if the social planner has $\underline{\kappa}$-strongly redistributive preferences, the optimal segmentation leaves a rent for the monopolist. Conversely, denote **no-rent region** the set of aggregate markets for which any optimal segmentation also maximizes total consumer surplus. The following result establishes a necessary condition for an aggregate market to belong to the no-rent region:

Proposition 16. *Let μ^* be an aggregate market with uniform price v_u. If $\mu^\star$ belongs to the no-rent region, then its optimal direct segmentation σ generates two segments μ^s and μ^r, with:*

$$\mu^s = (\frac{\mu_1^*}{\sigma}, \frac{\mu_2^*}{\sigma}, \dots, \mu_u^s, 0, \dots, 0),$$
$$\mu^r = (0, 0, \dots, \mu_u^r, \frac{\mu_{u+1}^*}{1-\sigma}, \dots, \frac{\mu_K^*}{1-\sigma}),$$

where $\sigma = \frac{\mu_u^r - \mu_u^*}{\mu_u^r - \mu_u^s}$ and $\mu_u^s = \frac{v_1}{v_u}$.

From corollary 4 we know that optimal segmentations under strong redistributive preferences admit an almost-partitional structure: contiguous types are pooled together, and each type is present in *at most* two segments. On the other hand, we have as a necessary and sufficient condition for total consumer surplus to be maximized that the segmentation is i) efficient and ii) the uniform price v_u is an optimal price in *every* segment generated by this segmentation. Condition i) ensures that total surplus is maximized, while condition ii) ensures that producer surplus is kept at it's uniform price level, ensuring that all of the surplus created by the segmentation goes to consumers. Since condition ii) can only be satisfied

if type v_u belongs in the support of all segments, we get that the conditions for optimality under strong redistributive preferences and for consumer-surplus to be maximized can only be simultaneously met if the optimal segmentation pools all types from v_1 to v_u into one segment and all types v_u to v_K into another, thus obtaining proposition 16.

This result establishes that, for markets in the no-rent region, optimal segmentations have an extremely simple structure: they only generate a discount segment with price v_1, pooling all the types who would not consume under the uniform price and some of the types v_u, and a residual segment with price v_u, containing all of the remaining consumers.

It is also interesting to note that, for a market in the no-rent region, this segmentation should be optimal given *any* decreasing $\boldsymbol{\lambda}$'s. To see this, first note that, given that the market belongs to the no-rent region, the optimal segmentation maximizes total consumer surplus. Due to the structure of the segmentation described in proposition 16, all of the surplus that is generated by the segmentation is given to consumers with types below or equal to v_u, all of which get the maximum surplus they could potentially get. Since it is impossible to raise the surplus of any type below v_u, and impossible to raise the surplus of types above v_u without redistributing from lower to higher types, this segmentation must be optimal whenever the weights assigned to different consumers are (weakly) decreasing on the type.

In the following proposition we characterize the no-rent region:

Proposition 17. *The no-rent region in price region M_u is given by:*

$$NRR_u = \{\mu \in M_u : A\mu \leq z\},$$

with

$$A = \begin{bmatrix} S & O_S \\ O_R & R \end{bmatrix} \in \mathbb{R}^{K-2\times K} \text{ and } z = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -\frac{v_1}{v_{u+1}(v_u - v_1)} \\ \vdots \\ -\frac{v_1}{v_K(v_u - v_1)} \end{bmatrix} \in \mathbb{R}^{K-2}$$

where O_S and O_R are null matrices with, respectively, dimensions $(u-2)\times(u-1)$ and $(K-u)\times(K+1-u)$, and

$$S = \begin{bmatrix} -\alpha(2) & 1-\alpha(2) & \cdots & 1-\alpha(2) & 1-\alpha(2) \\ -\alpha(3) & -\alpha(3) & \cdots & 1-\alpha(3) & 1-\alpha(3) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\alpha(u-2) & -\alpha(u-2) & \cdots & 1-\alpha(u-2) & 1-\alpha(u-2) \\ -\alpha(u-1) & -\alpha(u-1) & \cdots & -\alpha(u-1) & 1-\alpha(u-1) \end{bmatrix} \in \mathbb{R}^{(u-2)\times(u-1)},$$

$$R = \begin{bmatrix} -\beta(u+1) & 1-\beta(u+1) & \cdots & 1-\beta(u+1) & 1-\beta(u+1) \\ -\beta(u+2) & -\beta(u+2) & \cdots & 1-\beta(u+2) & 1-\beta(u+2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\beta(K-1) & -\beta(K-1) & \cdots & 1-\beta(K-1) & 1-\beta(K-1) \\ -\beta(K) & -\beta(K) & \cdots & -\beta(K) & 1-\beta(K) \end{bmatrix} \in \mathbb{R}^{(K-u)\times(K+1-u)}$$

for $\alpha(j) = \frac{v_1(v_u - v_j)}{v_j(v_u - v_1)}$ and $\beta(j) = \frac{v_u^2}{v_j(v_u - v_1)}$.

Proposition 17 defines the no-rent region as a polytope inside each price region. An immediate corollary is that the rent region is defined as the complement of the no-rent region within each price region:

Corollary 5. *Consider an aggregate market $\mu^* \in M_u$. If $\mu^* \notin NRR_u$, then there exists $\underline{\kappa}$ such that for any $\kappa \geq \underline{\kappa}$, if $(\lambda_k)_k$ are strongly redistributive at level κ, no*

optimal segmentation is surplus-maximizing (i.e all optimal segmentations grant a rent to the monopolist).

The results in this section establish that there are essentially two types of markets: those for which redistribution can be done only within consumers, while keeping total consumer surplus maximal, and those for which increasing the surplus of lower types past a certain point necessarily decreases the total pie of surplus accruing to consumers and grants additional profits to the monopolist.

Appendices

Appendix A

Proofs of Chapter 1

A.1 Proof of Proposition 1

Proof. We say that a pair $(\mathbf{m}, \geqslant)$ **rationalizes** $\succsim$ if for any $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \succsim \mathbf{y} \iff \mathbf{x} \circ \mathbf{m} \geqslant \mathbf{y} \circ \mathbf{m}$. The proof of the necessity is straightforward and therefore omitted.

(*Sufficiency*). Assume that $(\succsim_t)_t$ satisfy Restricted Reversals. First, we fix a period t and show that we can indeed construct an ordering $\geqslant_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$. We define the two following binary relations on $X(\mathbf{m}_t)$:

$$>_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \succ_t \mathbf{y}\},$$
$$\simeq_t = \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}_t) \ : \ \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}_t, \mathbf{b} = \mathbf{y} \circ \mathbf{m}_t, \text{ and } \mathbf{x} \sim_t \mathbf{y}\}.$$

By definition, $\simeq_t$ is reflexive and symmetric. We show that $>_t$ is irreflexive, i.e. for any $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \neq \mathbf{y}$ and $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$, $\mathbf{x} \sim_t \mathbf{y}$. Indeed, $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$ implies that $\mathbf{x}$ and $\mathbf{y}$ do not differ on any attributes k such that $m_t^k = 1$. Hence, if by contradiction we had $\mathbf{x} \succ_t \mathbf{y}$, then there should be a subset of attributes on which $\mathbf{x}$ and $\mathbf{y}$ differ (i.e. with $m_t^k = 0$) and that are revealed relevant. This contradicts the definition of $\mathbf{m}_t$.

Now let $\mathbf{a},\mathbf{b} \in X(\mathbf{m}_t)$, with $\mathbf{a} \neq \mathbf{b}$, and $\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}' \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_t = \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_t = \mathbf{b}$. Applying Restricted Reversal with $t = t'$, we obtain $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}' \succsim_t \mathbf{y}'$. Given that $>_t$ is irreflexive, this establishes that it is asymmetric. It also proves that $>_t \cap \simeq_t = \emptyset$.

Therefore, the relation $\geqslant_t := \simeq_t \cup >_t$ is complete on $X(\mathbf{m}_t)$ (by the completeness of $\succsim_t$) ; $\simeq_t$ and $>_t$ being respectively its symmetric and asymmetric parts. Furthermore, $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$.

Second, we show that for any two distinct periods t and t', $\geqslant_t$ does not contradict $\geqslant_{t'}$. Let $\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}'$ be such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{x}' \circ \mathbf{m}_{t'} =: \mathbf{a}$ and $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m}_{t'} =: \mathbf{b}$. Then by Restricted Reversal we have,

$$\mathbf{a} \geqslant_t \mathbf{b} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant_t \mathbf{y} \circ \mathbf{m}_t \iff \mathbf{x} \succsim_t \mathbf{y} \overbrace{\iff}^{\text{Restricted Reversal}} \mathbf{x}' \succsim_{t'} \mathbf{y}'$$
$$\iff \mathbf{x}' \circ \mathbf{m}_{t'} \geqslant_{t'} \mathbf{y}' \circ \mathbf{m}_{t'} \iff \mathbf{a} \geqslant_{t'} \mathbf{b}$$

Finally, define $\geqslant := \bigcup_t \geqslant_t$. By the previous argument, $\geqslant \cap X^2(\mathbf{m}_t) = \geqslant_t$, so for any t, $(\mathbf{m}_t, \geqslant)$ rationalizes $\succsim_t$. Furthermore $\geqslant$ can be innocuously completed on $\bar{X}$. □

A.2 Proof of Theorem 1

Proof. (Necessity). We only prove the necessity of Acyclicity. Let t and t' such that $t+1 < t'$ and assume $\mathbf{m}_{t'} \neq \mathbf{m}_{t+1}$. By equation (2.2), $\mathbf{m}_{t'} \rhd \mathbf{m}_{t+1}$ and, thus, $\mathbf{m}_{t'} \notin R(\mathbf{m}_t, \mathbf{a}_t)$. Hence, $m_t^k \neq m_{t'}^k$ and $a_t^k = 0$ for some attribute k. Yet, $a_t^k = 0$ and $\mathbf{m}_{t+1} \in R(\mathbf{m}_t, \mathbf{a}_t)$ imply that $m_t^k = m_{t+1}^k$, and thus $m_{t+1}^k \neq m_{t'}^k$

(Sufficiency). We know from proposition 1 that there exists an attribute ordering $\geqslant \subseteq \bar{X}^2$, such that for any period t, $(\mathbf{m}_t, \geqslant)$ rationalizes $\succsim_t$. Define the sequence of awareness as $\mathbf{a}_t = |\mathbf{m}_{t+1} - \mathbf{m}_t|$ for any t; and the revealed meta-preference relation

$\triangleright$ as follows: $\mathbf{m} \triangleright \mathbf{m}'$ if $\mathbf{m} \neq \mathbf{m}'$ and there exists t, such that $\mathbf{m} = \mathbf{m}_t$ and,

$$\mathbf{m}' \in \bigcup_{t' : t' < t} R(\mathbf{m}_{t'}, \mathbf{a}_{t'}).$$

We verify that $\triangleright$ is asymmetric. Suppose that $\mathbf{m} \triangleright \mathbf{m}'$ and take $t' < t$, such that $\mathbf{m}' \in R(\mathbf{m}_{t'}, \mathbf{a}_{t'})$ and $\mathbf{m} = \mathbf{m}_t$. First let us show that there cannot be $t'' > t$ such that $\mathbf{m}' = \mathbf{m}_{t''}$. Assume by contradiction that such a t'' exists. Then, we have

$$|\mathbf{m}_{t''} - \mathbf{m}_{t'}| \underbrace{=}_{\text{Def. } \mathbf{m}_{t''}} |\mathbf{m}' - \mathbf{m}_{t'}| \underbrace{\leq}_{\mathbf{m}' \in R(\mathbf{m}_{t'}, |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|)} |\mathbf{m}_{t'+1} - \mathbf{m}_{t'}|$$

where $|\cdot|$ is the element-wise absolute value: for any vector $\mathbf{b} \in \mathbb{R}^K$, $|\mathbf{b}| = (|b^1|, \ldots, |b^K|)$. This means that for all k, if $m^k_{t'} = m^k_{t'+1}$, then $m^k_{t''} = m^k_{t'} = m^k_{t'+1}$. Thus Acyclicity implies that $\mathbf{m}_{t'+1} = \mathbf{m}_{t''} = \mathbf{m}'$. But, then we still have that $\mathbf{m}' \in R(\mathbf{m}_{t'+1}, |\mathbf{m}_{t'+2} - \mathbf{m}_{t'+1}|)$ so that, applying the previous reasoning inductively, we obtain $\mathbf{m}' = \mathbf{m}_{t'+2} = \mathbf{m}_{t'+3} = \ldots \mathbf{m}_t = \mathbf{m} \neq \mathbf{m}'$. A contradiction. Second assume by contradiction that $\mathbf{m}' = \mathbf{m}_{t'''}$ and $\mathbf{m} \in R(\mathbf{m}_{t''}, |\mathbf{m}_{t''+1} - \mathbf{m}_{t''}|)$ for some t'', t''' such that $t'' < t''' < t$. By the same argument, Acyclicity would then imply that $\mathbf{m} = \mathbf{m}_{t''+2} = \mathbf{m}_{t''+3} = \ldots \mathbf{m}_{t'''} = \mathbf{m}' \neq \mathbf{m}$. A contradiction.

We now verify that $\triangleright$ is transitive. Suppose that $\mathbf{m} \triangleright \mathbf{m}'$ and $\mathbf{m}' \triangleright \mathbf{m}''$. Then there exist t, t' with $t > t'$, such that, $\mathbf{m} = \mathbf{m}_t$ and $\mathbf{m}' = \mathbf{m}_{t'}$. Moreover,

$$\mathbf{m}'' \in \bigcup_{t''' : t''' < t'} R(\mathbf{m}_{t'''}, \mathbf{a}_{t'''}) \subseteq \bigcup_{t'' : t'' < t} R(\mathbf{m}_{t''}, \mathbf{a}_{t''})$$

where the inclusion follows from $t > t'$. We conclude that $\mathbf{m} \triangleright \mathbf{m}''$, implying the transitivity of $\triangleright$.

Moreover, by the definition of $\triangleright$, $\mathbf{m}_{t+1} = \max(R(\mathbf{m}_t, \mathbf{a}_t), \triangleright)$. By Szpilrajn's theorem, the meta-preference can be completed on $\{0,1\}^K \times \{0,1\}^K$. $\square$

A.3 Proof of Proposition 2

Proof. For any t, any $\mathbf{a},\mathbf{b} \in X(\mathbf{m}_t)$, $\mathbf{a} \geqslant \mathbf{b}$ if and only if there exist, $\mathbf{x},\mathbf{y} \in X$ such that $\mathbf{x} \circ \mathbf{m}_t = \mathbf{a}$, $\mathbf{y} \circ \mathbf{m}_t = \mathbf{b}$ and $\mathbf{x} \succsim_t \mathbf{y}$, which in turn is true if and only if $\mathbf{a} \geqslant' \mathbf{b}$. Which establishes that $\geqslant \cap X^2(\mathbf{m_t}) = \geqslant' \cap X^2(\mathbf{m_t})$ for any t. Therefore, any completion of $\geqslant \cap \geqslant'$ together with $\mathbf{m}_t$ rationalizes $\succsim_t$ for any t.

We next show that by considering $\rhd \cap \rhd'$ and the sequence of awareness $(\mathbf{a}_t \circ \mathbf{a}'_t)_t$, we can rationalize the meta-choices of each period t. Fix a period t, and suppose that being at $\mathbf{m}_t$, DM faces the meta-menu $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$. Note that $R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t) = R(\mathbf{m}_t, \mathbf{a}_t) \cap R(\mathbf{m}_t, \mathbf{a}'_t)$. Hence it implies that $(\mathbf{m}_{t+1}, \mathbf{m}) \in \rhd \cap \rhd'$ for any $\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)$. This completes the proof than any completion of $\rhd \cap \rhd'$, together with $(\geqslant \cap \geqslant', \mathbf{m}_t, \mathbf{a}_t \circ \mathbf{a}'_t)_t$ rationalize $(\succsim_t)_t$. □

A.4 Proof of Proposition 3

Proof. To show that Perfect Deliberation is necessary, assume that $(a_t)_t$ and $\rhd$ represent $(\succsim_t)_t$ with $(a_t)_t$ growing. Let t, t' such that $t < t'$, $\mathbf{m}_{t'} \neq \mathbf{m}_t$. Hence, the set $H = \{k \in K : m_t^k \neq m_{t'}^k\}$ is not empty. If, by contradiction, for all $k \in H$, we have $m_t^k \neq m_1^k$, then given that $(a_t)_t$ grows, for all $k \in H$, $a_{t-1}^k = 1$. Hence, $\mathbf{m}_{t'} \in R(\mathbf{m}_{t-1}, \mathbf{a}_{t-1})$ and $\max(R(\mathbf{m}_{t-1}, \mathbf{a}_{t-1}), \rhd) = \mathbf{m}_t \neq \mathbf{m}_{t'}$, while $\mathbf{m}_{t'} \rhd \mathbf{m}_t$. A contradiction.

To show that Perfect Deliberation (with Sufficient Reason) is sufficient, note that it implies Deliberation. Thus, there exists $\rhd$ such that $\mathbf{m}_t \rhd \mathbf{m}_{t-1}$ for any t. Consider the growing sequence $(\mathbf{a}_t)_t$ defined by

$$\mathbf{a}_t = \mathbf{a}_{t-1} + (1 - \mathbf{a}_{t-1}) \cdot (|\mathbf{m}_{t+1} - \mathbf{m}_t|)$$

and[1] $\mathbf{a}_1 = |\mathbf{m}_2 - \mathbf{m}_1|$. Assume by contradiction that $(\mathbf{a}_t)_t$, together with $\rhd$, does not

[1] 1 is the vector $(1, \ldots, 1)$.

represent $(\succsim_t)_t$. Noting that, by definition of $(\mathbf{a}_t)_t$, $\mathbf{m}_{t+1} \in R(\mathbf{m}_t, \mathbf{a}_t)$, this can be the case only if for some t, t' such that $t' > t$ we have $\mathbf{m}_{t+1} \neq \mathbf{m}_{t'}$, $\mathbf{m}_{t'} \rhd \mathbf{m}_{t+1}$, and $\mathbf{m}_{t'} \in R(\mathbf{m}_t, \mathbf{a}_t)$. Note that, since $\rhd$ is transitive, $\mathbf{m}_{t+1} \neq \mathbf{m}_{t'}$ implies that for any $t'' \leq t+1$, we have $\mathbf{m}_{t''} \neq \mathbf{m}_{t'}$. Hence, by Perfect Deliberation, there exists k such that for all $t'' \leq t+1 < t'$ $m_1^k = m_{t''}^k \neq m_{t'}^k$. Hence, for such k, we have $m_{t+1}^k \neq m_{t'}^k$ and, by definition of $(\mathbf{a}_t)_t$, $a_t^k = 0$. This contradicts $\mathbf{m}_{t'} \in R(\mathbf{m}_t, \mathbf{a}_t)$.

□

A.5 Proof of Proposition 4

Proof. (*Necessity*.) Suppose there exists a complete preorder $\geqslant$ such that for every t, $(\mathbf{m}_t, \geqslant)$ represents $\succsim_t$. Take any $\{t_1, \ldots, t_n\}$ and any $\{\mathbf{x_k}, \mathbf{x'_k}\}_{k=1,\ldots,n}$ such that, for $k = 1, \ldots, n-1$: $\mathbf{x}'_k \circ \mathbf{m}_{t_k} = \mathbf{x}_{k+1} \circ \mathbf{m}_{t_{k+1}}$, $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1}$, and for every $k \leq n-1$, $\mathbf{x}_k \succsim_{t_k} \mathbf{x}'_k$. The latter implies that $\mathbf{x}_k \circ \mathbf{m}_{t_k} \geqslant \mathbf{x}'_k \circ \mathbf{m}_{t_k}$. Hence by the transitivity of $\geqslant$, we can conclude that $\mathbf{x}'_n \circ \mathbf{m}_{t_n} = \mathbf{x}_1 \circ \mathbf{m}_{t_1} \geqslant \mathbf{x}_{n-1} \circ \mathbf{m}_{t_{n-1}} = \mathbf{x}_n \circ \mathbf{m}_{t_n}$, i.e. $\mathbf{x}'_n \succsim_{t_n} \mathbf{x}_n$. Hence Strong Restricted Reversal is satisfied.

(*Sufficiency*). We fix a period t and show that we can construct a complete preorder $\geqslant_t \subseteq X^2(\mathbf{m}_t)$ such that $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$. We define $\geqslant_t$ in the same way as in the proof of proposition 1. Given that Strong Restricted Reversal implies Restricted Reversal, the same arguments apply and we conclude that $(\mathbf{m}_t, \geqslant_t)$ rationalizes $\succsim_t$. Furthermore, the transitivity of $\geqslant_t$ is a direct consequence of the transitivity of $\succsim_t$. We need now to construct a complete preorder $\geqslant$ on $\bar{X}$ that is time-independent.

From the proof of proposition 1, we know that for any two distinct periods t and t', $\geqslant_t$ does not contradict $\geqslant_{t'}$. We define $\geqslant_{1;T} := \bigcup_t \geqslant_t$. We know therefore that for any t, $\geqslant_{1;T} \cap X^2(\mathbf{m}_t) = \geqslant_t$.

We next show that the transitive closure of $\geqslant_{1;T}$, denoted $\geqslant_{1;T}^C$, can rationalize

the sequence $(\succsim_t)_t$ together with the sequence $(\mathbf{m}_t)_t$. Namely, we show that for any period t, any $\mathbf{a},\mathbf{b} \in X(\mathbf{m}_t)$, if $\mathbf{b} >_t \mathbf{a}$, there cannot be a sequence $(\mathbf{a}_k)_{k=1,\dots,n}$ and $(t_k)_{1\leq k\leq n-1}$ such that $\mathbf{a}_1 = \mathbf{a}$, $\mathbf{a}_n = \mathbf{b}$ and $\mathbf{a}_k \geqslant_{t_k} \mathbf{a}_{k+1}$. Let suppose by contradiction the existence of such a sequence. If $t_1 \neq t$, then complete the sequence with $\mathbf{a}_0 = \mathbf{a}$ and $t_0 = t$; similarly, if $t_{n-1} \neq t$, then complete the sequence with $\mathbf{a}_{n+1} = \mathbf{b}$ and $t_n = t$. Therefore, w.l.o.g we consider the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$.

If $t_k = t_{k'}$ for some $k \neq k'$, we show that we can restrict to a subsequence $(\mathbf{a}_{\tau(k)})_{k=1,\dots,n+1}$ with $\tau(0) = 0$, $\tau(n+1) = n+1$, such that $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$. Let suppose that $t_k = t_{k'}$ with $k < k'$ and that for any $k \leq i, j < k'$, if $i \neq j$ then $t_i \neq t_j$. Let's consider the sequence $(\mathbf{a}_i)_{k\leq i\leq k'+1}$. There exists a sequence $(\mathbf{x}_i, \mathbf{y}_{i+1})_{k\leq i\leq k'}$ such that $\mathbf{x}_k \circ \mathbf{m}_{t_k} = \mathbf{a}_k$, $\mathbf{y}_{k'+1} \circ \mathbf{m}_{t_{k'}} = \mathbf{a}_{k'+1}$, for any $k \leq i \leq k'-1$, $\mathbf{y}_{i+1} \circ \mathbf{m}_{t_i} = \mathbf{x}_{i+1} \circ \mathbf{m}_{t_{i+1}} = \mathbf{a}_{i+1}$, and for any $k \leq i \leq k'$, $\mathbf{x}_i \succsim_{t_i} \mathbf{y}_{i+1}$. By applying Strong Restricted Reversal, this must be that $\mathbf{x}_k \succsim_{t_k} \mathbf{y}_{k'+1}$, i.e. $\mathbf{a}_k \geqslant_{t_k} \mathbf{a}_{k'+1}$. Therefore, from the sequence $(\mathbf{a}_k)_{k=0,\dots,n+1}$, we can construct a subsequence $(\mathbf{a}_{\tau(k)})_{k=0,\dots,n+1}$, with $\tau(0) = 0$, $\tau(n+1) = n+1$, $\tau(i) \neq \tau(j) \implies t_{\tau(i)} \neq t_{\tau(j)}$, and such that for any k with $\tau(k) \neq \tau(k+1)$, $\mathbf{a}_{\tau(k)} \geqslant_{\tau(k)} \mathbf{a}_{\tau(k+1)}$. From a similar reasoning, we conclude by Strong Restricted Reversal that $\mathbf{a} \geqslant_t \mathbf{b}$, a contradiction.

By an implication of Szpilrajn's theorem (see Corollary A.1 in Ok (2007)), there exists a complete, transitive and reflexive binary relation that extends $\geqslant^C_{1;T}$. We denote it $\geqslant$. We proved that for any t, $X^2(\mathbf{m}_t) \cap \geqslant = \geqslant_t$, hence $(\mathbf{m}_t, \geqslant)$ rationalizes $\succsim_t$. □

Lemma 4. *Assume richness and the transitivity of $\succsim_t$. If $M \subset \{1,\dots,K\}$ is revealed relevant, then M is a singleton.*

Proof. Let M be revealed relevant and suppose by contradiction that $|M| = n > 1$. This means that there exists $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}^{-M} = \mathbf{y}^{-M}$ and $x^k \neq y^k$ for any $k \in M$, with $\mathbf{x} \not\sim_t \mathbf{y}$; and for every $M' \subsetneq M$ and every $\mathbf{w}, \mathbf{z} \in X$ with $\mathbf{w}^{-M'} = \mathbf{z}^{-M'}$, $\mathbf{w} \sim_t \mathbf{z}$. By the richness assumption, there exists a sequence of alternatives $\mathbf{z}_1, \dots, \mathbf{z}_n \in X$

such that $\mathbf{z}_1 = \mathbf{x}$, $\mathbf{z}_n = \mathbf{y}$ and $z_i^{-k} = z_{i+1}^{-k}$ for some $k \in M$, for all $i = 1, \ldots, n-1$. By assumption, it must be that $\mathbf{z}_i \sim_t \mathbf{z}_{i+1}$ for all $i = 1, \ldots, n-1$, which by transitivity would imply that $\mathbf{x} \sim_t \mathbf{y}$, a contradiction. □

A.6 Proof of Proposition 6

Proof. First, note that $(1, \ldots, 1)$ can rationalize $\succsim_t$. In this case, our representation at t coincides with standard preference maximization because for any $\mathbf{x} \in X$, $\mathbf{x} \circ (1, \ldots, 1) = \mathbf{x}$. Identifying $\geqslant_t$ with $\succsim_t$ yields the desired result.

Second, we show that for any $\mathbf{m}' \ngeq \mathbf{m}_t$, $(\mathbf{m}', \geqslant_t')$ cannot rationalize $\succsim_t$ for some $\geqslant_t'$. By contradiction, suppose that there exists such $\mathbf{m}'$. Given lemma 4 and the definition of $\mathbf{m}_t$, there exists an attribute k such that $m_t^k - m'^k = 1$, and some alternatives $\mathbf{x}, \mathbf{y}$ such that $\mathbf{x}^{-k} = \mathbf{y}^{-k}$, $x^k \neq y^k$ and $\mathbf{x} \nsucc_t \mathbf{y}$ for some $\mathbf{x}, \mathbf{y} \in X$. Given that $\mathbf{x} \circ \mathbf{m}' = \mathbf{y} \circ \mathbf{m}'$, this contradicts the fact that $(\mathbf{m}', \geqslant_t')$ rationalizes $\succsim_t$.

Finally, we prove that for any $\mathbf{m}' > \mathbf{m}_t$, there exists $\geqslant_t'$ such that $(\mathbf{m}', \geqslant_t')$ rationalizes $\succsim_t$. Define:

$$\begin{aligned}
>_t' &= \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}') \,:\, \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}', \mathbf{b} = \mathbf{y} \circ \mathbf{m}', \text{ and } \mathbf{x} \succ_t \mathbf{y}\}, \\
\simeq_t' &= \{(\mathbf{a}, \mathbf{b}) \in X^2(\mathbf{m}') \,:\, \exists\, \mathbf{x}, \mathbf{y} \in X, \mathbf{a} = \mathbf{x} \circ \mathbf{m}', \mathbf{b} = \mathbf{y} \circ \mathbf{m}', \text{ and } \mathbf{x} \sim_t \mathbf{y}\}.
\end{aligned}$$

A similar reasoning as in the proof of proposition 1 establishes that $(\mathbf{m}', \geqslant_t')$ rationalizes $\succsim_t$. □

A.7 Proof of Theorem 2

Proof. The proof of the necessity of Justified Indifference is left to the readers. By proposition 4, $\mathbf{m}_t \in \mathscr{M}(\succsim_t)$, so that there exists $\geqslant$ such that for all $\mathbf{x}, \mathbf{y} \in X$

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geqslant \mathbf{y} \circ \mathbf{m}_t \tag{A.1}$$

Moreover, the contraposition of Justified Indifference implies that for all $\mathbf{x}, \mathbf{y} \in X$, if $\mathbf{x} \circ \mathbf{m}_t \neq \mathbf{y} \circ \mathbf{m}_t$, then either $\mathbf{x} \succ_t \mathbf{y}$ or $\mathbf{x} \succ_t \mathbf{y}$. Hence, for any $\mathbf{x}, \mathbf{y} \in X$ such that $\mathbf{x} \circ \mathbf{m}_t \neq \mathbf{y} \circ \mathbf{m}_t$, (A.1) implies that $\mathbf{x} \circ \mathbf{m}_t > \mathbf{y} \circ \mathbf{m}_t$. Hence, $\mathbf{m}_t \in \mathscr{M}^\star(\succsim_t)$.

Now assume by contradiction that there exists $\mathbf{m} \in \mathscr{M}^\star(\succsim_t)$ with $\mathbf{m} \neq \mathbf{m}_t$. We know from proposition 4 that there exists i such that $m_t^i = 0$ and $m^i = 1$. From the fact that there should be two alternatives in X that differ on i and the richness assumption it can easily be shown that there exist two alternatives $\mathbf{x}, \mathbf{y} \in X$ such that $x^i \neq y^i$ and $\mathbf{x}^{-i} = \mathbf{y}^{-i}$. This means that $\mathbf{x} \circ \mathbf{m} \neq \mathbf{y} \circ \mathbf{m}$. Given that $\mathbf{m} \in \mathscr{M}^\star(\succsim_t)$, this means that there exists $>$ such that either $\mathbf{x} \circ \mathbf{m} > \mathbf{y} \circ \mathbf{m}$ or $\mathbf{x} \circ \mathbf{m} < \mathbf{y} \circ \mathbf{m}$ that rationalizes $\succsim_t$. Hence, we either have $\mathbf{x} \succ_t \mathbf{y}$ or $\mathbf{y} \succ_t \mathbf{x}$. Furthermore, $\mathbf{x} \circ \mathbf{m}_t = \mathbf{y} \circ \mathbf{m}_t$, which, given that $\mathbf{m}_t \in \mathscr{M}^\star(\succsim_t)$, implies that $\mathbf{x} \sim_t \mathbf{y}$, a contradiction. □

A.8 Proof of Theorem 3

Proof. *(Necessity)* Assume that there exists a sequence of awareness $(\mathbf{a}_t)_t$ and a function $u : \mathbb{R}^K \times \{0,1\}^K \longrightarrow \mathbb{R}$ such that for all t, all $\mathbf{m}_t$ and all $\mathbf{x}$,

$$\mathbf{x} \precsim_t \mathbf{x}' \iff u(\mathbf{x} \circ \mathbf{m}_t) \leq u(\mathbf{x}' \circ \mathbf{m}_t)$$

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\arg\max} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m}).$$

Step 1: We show that for all t if $\mathbf{a} \gg_t \mathbf{b}$, then $u(\mathbf{a}) \geq u(\mathbf{b})$.

Assume first that there exists $\mathbf{x},\mathbf{y} \in X$ such that $\mathbf{a} = \mathbf{x} \circ \mathbf{m}_t$ and $\mathbf{b} = \mathbf{y} \circ \mathbf{m}_t$. Then, $\mathbf{x} \succsim_t \mathbf{y}$ implies that $u(\mathbf{a}) = u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{y} \circ \mathbf{m}_t) = u(\mathbf{b})$. Now assume that there exists $\mathbf{x},\mathbf{y} \in X$ such that $\mathbf{a} = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{b} = \mathbf{y} \circ \mathbf{m}$ for some $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, and $\mathbf{x} \in \max(X, \precsim_t)$. Then $R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|) \subset R(\mathbf{m}_{t-1}, \mathbf{a}_t)$ and hence $u(\mathbf{x} \circ \mathbf{m}_t) = \max_{\mathbf{m} \in R(\mathbf{m}_{t-1}, \mathbf{a}_t)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m}) \geq \max_{\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m}) \geq u(\mathbf{y} \circ \mathbf{m}) = u(\mathbf{b})$.

Step 2. We show that Motivated Restricted Reversals holds.

Suppose we have $\{t_1, \ldots, t_n\}$ and $\{\mathbf{a_k}\}_{k=1,\ldots,n} \in (\mathbb{R}^K)^n$ such that $\mathbf{a}_{k+1} \gg_{t_k} \mathbf{a}_k$ for $k = 1, \ldots, n-1$, and $\mathbf{a}_1 \gg_t \mathbf{a}_n$. Given *Step 1.*, this implies that

$$u(\mathbf{a}_n) \geq \cdots \geq u(\mathbf{a}_2) \geq u(\mathbf{a}_1) \geq u(\mathbf{a}_n)$$

Hence, $u(\mathbf{a}_n) = u(\mathbf{a}_1)$. Moreover, from $\mathbf{a}_1 \gg_t \mathbf{a}_n$ we know that either there exists $\mathbf{x},\mathbf{y} \in X$ such that $\mathbf{a}_1 = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}_t$, and $\mathbf{x} \succsim_t \mathbf{y}$; or there exist $\mathbf{x},\mathbf{y} \in X$ such that $\mathbf{a}_1 = \mathbf{x} \circ \mathbf{m}_t$, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}$ for some $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|)$, and $\mathbf{x} \in \max(X, \succsim_t)$. In the first case, since $u(\mathbf{a}_n) = u(\mathbf{a}_1)$, we have $\mathbf{x} \sim_t \mathbf{y}$ and, therefore, $\mathbf{a}_n = \mathbf{y} \circ \mathbf{m}_t \gg_t \mathbf{x} \circ \mathbf{m}_t = \mathbf{a}_1$. In the second case, if $\mathbf{m} = \mathbf{m}_t$ we are back to the first case; otherwise, if $\mathbf{m} \neq \mathbf{m}_t$, since $u(\mathbf{y} \circ \mathbf{m}) = u(\mathbf{a}_n) = u(\mathbf{a}_1) = u(\mathbf{x} \circ \mathbf{m}_t)$ and $\mathbf{m}_t \in \arg\max_{\mathbf{m} \in R(\mathbf{m}_{t-1}, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$, we have that the set $\arg\max_{\mathbf{m} \in R(\mathbf{m}_{t-1}, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$ contains at least two elements. This contradicts $\{\mathbf{m}_t\} = \arg\max_{\mathbf{m} \in R(\mathbf{m}_{t+1}, \mathbf{a}_t)} \max_{\mathbf{x} \in X} u(\mathbf{x} \circ \mathbf{m})$.

Step 3. We show that Motivated Tie-Breaking holds.

Let period t, $\mathbf{x} \in \max(X, \succsim_t)$, $\mathbf{y},\mathbf{y}' \in X$ such that $\mathbf{y} \circ \mathbf{m}_t = \mathbf{y}' \circ \mathbf{m} \circ \mathbf{m}_t$ and $\mathbf{y}' \in \max(X, \precsim_t)$. We have that $u(\mathbf{y} \circ \mathbf{m}_t) = u(\mathbf{y}' \circ \mathbf{m} \circ \mathbf{m}_t)$. But then since $u(\mathbf{y}' \circ \mathbf{m}_t) = \max_{\mathbf{m} \in R(\mathbf{m}_t, |\mathbf{m}_t - \mathbf{m}_{t-1}|)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m})$, if $\mathbf{m} \neq \mathbf{m}_t$, then the set $\arg\max_{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m})$ contains at least two elements. This contradicts the fact that

$$\{\mathbf{m}_t\} = \arg\max_{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)} \max_{\mathbf{x}' \in X} u(\mathbf{x}' \circ \mathbf{m}).$$

(Sufficiency) Define $\geq^*$ as follows

$$\geq^* = \bigcup_{n\in\mathbb{N}} \bigcup_{t_1,t_2,\ldots,t_n,t_{n+1}} \{(\mathbf{a},\mathbf{b}) \in \mathbb{R}^2 : \exists (\mathbf{a}_k)_{k\leq n} \in \mathbb{R}^n, \mathbf{a} \gg_{t_{n+1}} \mathbf{a}_n \gg_{t_n} \ldots \mathbf{a}_1 \gg_{t_1} \mathbf{b}\}$$

Step 1: We show that for all $\mathbf{x},\mathbf{y} \in X$, $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}\circ\mathbf{m}_t \geq^* \mathbf{y}\circ\mathbf{m}_t$.

First note that, by definition of $\gg_t$, $\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x}\circ\mathbf{m}_t \gg_t \mathbf{y}\circ\mathbf{m}_t$. Hence, we only need to prove that $\mathbf{x}\circ\mathbf{m}_t \gg_t \mathbf{y}\circ\mathbf{m}_t \iff \mathbf{x}\circ\mathbf{m}_t \geq^* \mathbf{y}\circ\mathbf{m}_t$. That $\mathbf{x}\circ\mathbf{m}_t \gg_t \mathbf{y}\circ\mathbf{m}_t \implies \mathbf{x}\circ\mathbf{m}_t \geq^* \mathbf{y}\circ\mathbf{m}_t$ directly follows from the definition of $\geq^*$. To show the converse, assume $\mathbf{x}\circ\mathbf{m}_t \geq^* \mathbf{y}\circ\mathbf{m}_t$. This means that there exist $(n,(t_k)_{1\leq k\leq n+1},(\mathbf{a}_k)_{1\leq k\leq n})$ such that

$$\mathbf{y}\circ\mathbf{m}_t \ll_{t_1} \mathbf{a}_1 \ll_{t_2} \cdots \ll_{t_{n+1}} \mathbf{x}\circ\mathbf{m}_t \tag{A.2}$$

Given that $(\mathbf{x}\circ\mathbf{m}_t,\mathbf{y}\circ\mathbf{m}_t) \in X^2(\mathbf{m}_t)$ and the completeness of $\precsim_t$ we either have that $\mathbf{x}\circ\mathbf{m}_t \gg_t \mathbf{y}\circ\mathbf{m}_t$ or $\mathbf{x}\circ\mathbf{m}_t \ll_t \mathbf{y}\circ\mathbf{m}_t$. If the former case holds there is nothing left to prove. If the later case holds, then, by Motivated Restricted Reversals and (A.2), so does the former.

Step 2: We show that if $\mathbf{x} \in \max(X,\precsim_t)$*, then for all* $\mathbf{m} \in R(\mathbf{m}_{t-1},|\mathbf{m}_t-\mathbf{m}_{t-1}|)\setminus\{\mathbf{m}_t\}$*, and all* $\mathbf{y} \in X$*, we have* $\mathbf{y}\circ\mathbf{m} <^* \mathbf{x}\circ\mathbf{m}_t$.

Assume that $\mathbf{x} \in \max(X,\precsim_t)$. Hence, for all $\mathbf{m} \in R(\mathbf{m}_{t-1},|\mathbf{m}_t-\mathbf{m}_{t-1}|)$ and all $\mathbf{y} \in X$, $\mathbf{y}\circ\mathbf{m} \ll_t \mathbf{x}\circ\mathbf{m}_t$, which implies that $\mathbf{y}\circ\mathbf{m} \leq^* \mathbf{x}\circ\mathbf{m}_t$. By contradiction, suppose that for some $\mathbf{m} \in R(\mathbf{m}_{t-1},|\mathbf{m}_t-\mathbf{m}_{t-1}|)\setminus\{\mathbf{m}_t\}$ and some $\mathbf{y} \in X$, $\mathbf{x}\circ\mathbf{m}_t \leq^* \mathbf{y}\circ\mathbf{m}$. This implies that there exist $(n,(t_k)_{1\leq k\leq n+1},(\mathbf{a}_k)_{1\leq k\leq n})$ such that, $\mathbf{x}\circ\mathbf{m}_t \ll_{t_1} \mathbf{a}_1 \ll_{t_2} \cdots \ll_{t_{n+1}} \mathbf{y}\circ\mathbf{m}$.

From this and the fact that $\mathbf{y}\circ\mathbf{m} \ll_t \mathbf{x}\circ\mathbf{m}_t$, it follows from Motivated Restricted Reversals that $\mathbf{x}\circ\mathbf{m}_t \ll_t \mathbf{y}\circ\mathbf{m}$. Hence by definition of $\ll_t$, there exists $\mathbf{y}' \in X$ such that $(\mathbf{y}\circ\mathbf{m})\circ\mathbf{m}_t = \mathbf{y}'\circ\mathbf{m}_t$ and $\mathbf{y}' \succsim_t \mathbf{x}$. By Motivated Ties Breaking this implies that $\mathbf{m} = \mathbf{m}_t$, a contradiction.

Step 3: We conclude.

Now note that, by construction, $\geq^*$ is transitive and reflexive. Thus, it is a preorder. By an extension of Szpilrajn's theorem (see Corollary A.1 in Ok (2007)) we can complete $\geq^*$ to obtain a complete preorder. This means that there exists utility function u representing $\geq^*$. By *step 1,* we thus have that for all t and all $\mathbf{x}, \mathbf{y} \in X$,

$$\mathbf{x} \succsim_t \mathbf{y} \iff \mathbf{x} \circ \mathbf{m}_t \geq^* \mathbf{y} \circ \mathbf{m}_t \iff u(\mathbf{x} \circ \mathbf{m}_t) \geq u(\mathbf{y} \circ \mathbf{m}_t).$$

By *step 2*, we have that for all t, all $\mathbf{x}, \mathbf{y} \in X$, and all $\mathbf{m} \in R(\mathbf{m}_{t-1}, |\mathbf{m}_t - \mathbf{m}_{t-1}|) \backslash \{\mathbf{m}_t\}$,

$$\mathbf{x} \in \max(X, \precsim_t) \implies \mathbf{x} \circ \mathbf{m}_t >^* \mathbf{y} \circ \mathbf{m} \iff u(\mathbf{x} \circ \mathbf{m}_t) > u(\mathbf{y} \circ \mathbf{m}).$$

Hence, taking $\mathbf{a}_t = |\mathbf{m}_{t+1} - \mathbf{m}_t|$ for all t, we obtain,

$$\{\mathbf{m}_{t+1}\} = \underset{\mathbf{m} \in R(\mathbf{m}_t, \mathbf{a}_t)}{\arg\max} \max_{\mathbf{x} \in X} u(\mathbf{x}, \mathbf{m}).$$

□

Appendix B

Proof of Theorem 4

We first establish a couple of lemmata that show how Strong Restricted Reversal* (SRR*) puts structure on the set of candidate sequences of relevant attributes.

For any $(t_1,\dots,t_n)\in\{1,\dots,T\}^n$, we denote $F^{(t_1,\dots,t_n)}\subseteq\mathbf{M}^{2n}$ all the sequences of sets of relevant attributes that satisfy the conditions stated in SRR*. Namely, $(M_k,M'_k)_{k=1,\dots,n}\in F^{(t_1,\dots,t_n)}$ if $M_k,M'_k\in\mathscr{M}(\succsim_{t_k})$ for $k=1,\dots,n$ and for any $(x_k,x'_k)_{k=1,\dots,n}$ such that:

$$\begin{aligned} &x'_k\cap M'_k = x_{k+1}\cap M_{k+1} \quad \text{for } k=1,\dots,n-1, \text{ and}\\ &x'_n\cap M'_n = x_1\cap M_1. \end{aligned}$$

If $x_k\succsim_{t_k}x'_k$ *for* $k=1,\dots,n-1$, then $x'_n\succsim_{t_n}x_n$.

Lemma 5. *For any* $(t_1,\dots,t_n)$, $(\bar M_k,\bar M'_k)_k\in F^{(t_1,\dots,t_n)}$, *and* $(M_k,M'_k)_{k=1,\dots,n}$ *: if* $M_k\subseteq\bar M_k$, $M'_k\subseteq\bar M'_k$ *and* $M_k,M'_k\in\mathscr{M}(\succsim_{t_k})$ *for* $k=1,\dots,n$, *then* $(M_k,M'_k)_k\in F^{(t_1,\dots,t_n)}$.

Proof. Let $(x_k, x'_k)_{k=1,\dots,n}$ be such that:

$$x'_k \cap M'_k = x_{k+1} \cap M_{k+1} \quad \text{for } k = 1,\dots,n-1, \tag{B.1}$$

$$x'_n \cap M'_n = x_1 \cap M_1 \quad \text{and} \tag{B.2}$$

$$x_k \succsim_{t_k} x'_k \quad \text{for } k = 1,\dots,n-1. \tag{B.3}$$

By Perfect Instantiation, there exist $(\alpha_k, \alpha'_k)_{k=1,\dots,n}$ such that, $\alpha_k = x_k \cap M_k$ and $\alpha'_k = x'_k \cap M'_k$ for $k = 1,\dots,n$. Since $M_k \subseteq \bar{M}_k$ and $M'_k \subseteq \bar{M}'_k$:

$$\alpha'_k \cap \bar{M}'_k = x'_k \cap M'_k = x_{k+1} \cap M_{k+1} = \alpha_{k+1} \cap \bar{M}_{k+1} \quad \text{for } k = 1,\dots,n-1, \text{ and} \tag{B.4}$$

$$\alpha'_n \cap \bar{M}'_n = x'_n \cap M'_n = x_1 \cap M_1 = \alpha_1 \cap \bar{M}_1. \tag{B.5}$$

Moreover, since $M_k, M'_k \in \mathscr{M}(\succsim_{t_k})$, $\alpha_k \cap M_k = x_k \cap M_k$ and $\alpha' \cap M'_k = x' \cap M'_k$, it must be that $\alpha_k \sim_{t_k} x_k$ and $\alpha'_k \sim_{t_k} x'_k$. Therefore transitivity implies that $x_k \succsim_{t_k} x'_k \iff \alpha_k \succsim_{t_k} \alpha'_k$, and thus (B.3) means that:

$$\alpha_k \succsim_{t_k} \alpha'_k \quad \text{for } k = 1,\dots n-1. \tag{B.6}$$

Hence, given that $(\bar{M}_k, \bar{M}'_k)_k \in F^{(t_1,\dots,t_n)}$, (B.4), (B.5) and (B.6) imply that $\alpha'_n \succsim_{t_n} \alpha_n$, which in turn implies that $x'_n \succsim_{t_n} x_n$. □

Lemma 6. *For any* $(t_1,\dots,t_n)$, $(M_k, M'_k)_k \in F^{(t_1,\dots,t_n)}$ *and* $(B_k, B'_k)_{k=1,\dots,n}$*: if* $M'_k \cap (B_{k+1} \cup B'_k) = M_{k+1} \cap (B_{k+1} \cup B'_k) = \emptyset$ *for every* $k \in \{1,\dots,n-1\}$ *and* $M'_n \cap (B_1 \cup B'_n) = M_1 \cap (B_1 \cup B'_n) = \emptyset$*, then* $(M_k \cup B_k, M'_k \cup B'_k) \in F^{(t_1,\dots,t_n)}$.

Proof. Let $(x_k, x'_k)_{k=1,\ldots,n}$ be such that:

$$x'_k \cap (M'_k \cup B'_k) = x_{k+1} \cap (M_{k+1} \cup B_{k+1}) \quad \text{for } k \in \{1,\ldots,n-1\}, \tag{B.7}$$

$$x'_n \cap (M'_n \cup B'_n) = x_1 \cap (M_1 \cup B_1) \quad \text{and} \tag{B.8}$$

$$x_k \succsim_{t_k} x'_k \quad \text{for } k = 1,\ldots,n-1. \tag{B.9}$$

Because $M'_k \cap (B_{k+1} \cup B'_k) = M_{k+1} \cap (B_{k+1} \cup B'_k) = \emptyset$, for every $k \in \{1,\ldots,n-1\}$,

$$x'_k \cap B_{k+1} = x_{k+1} \cap B'_k = \emptyset,$$

which together with (B.7) imply that,

$$\begin{aligned} & x'_k \cap B'_k = x_{k+1} \cap B_{k+1} = \emptyset, \\ \text{that is,} \quad & x'_k \cap (B'_k \cup B_{k+1}) = x_{k+1} \cap (B'_k \cup B_{k+1}) = \emptyset. \end{aligned}$$

Hence:

$$\begin{aligned} & (x'_k \cap (M'_k \cup B'_k)) \setminus (B'_k \cup B_{k+1}) = x'_k \cap M'_k \\ \text{and} \quad & x_{k+1} \cap (M_{k+1} \cup B_{k+1}) \setminus (B'_k \cup B_{k+1}) = x_{k+1} \cap M_{k+1}, \\ \text{that is,} \quad & x'_k \cap M'_k = x_{k+1} \cap M_{k+1}. \end{aligned} \tag{B.10}$$

Similarly, given that $M'_n \cap (B_1 \cup B'_n) = M_1 \cap (B_1 \cup B'_n) = \emptyset$,

$$x'_n \cap (B'_n \cup B_1) = x_1 \cap (B'_n \cup B_1) = \emptyset.$$

Hence:

$$
\begin{aligned}
& (x'_n \cap (M'_n \cup B'_n)) \setminus (B'_n \cup B_1) = x'_n \cap M'_n \\
\text{and} \quad & x_1 \cap (M_1 \cup B_1) \setminus (B'_n \cup B_1) = x_1 \cap M1, \\
\text{that is,} \quad & x'_n \cap M'_n = x_1 \cap M_1. \qquad \text{(B.11)}
\end{aligned}
$$

The desired conclusion finally follows from (B.9), (B.10), (B.11) and the fact that $(M_k, M'_k)_k \in F^{(t_1,\ldots,t_n)}$. □

Proof of Theorem 4. *(Necessity).* Suppose that $(\succsim_t)_t$ is represented by general deliberate preference change. Let define the following explanation E: for all t, t', $t < t'$,

$$E_{t,t'} := M_t \triangle M_{t'}.$$

First note that, given this definition, No Explanatory Gap is trivially satisfied. To show that $(E_{t,t'})_{t<t'}$ satisfies Acyclic Explanation, consider t, τ with $t+1 < \tau$ such that $E_{t,\tau} \subseteq E_{t,t+1}$. A necessary consequence of the meta-choice is that $M_\tau \rhd M_{t+1} \rhd M_t$. Furthermore, the operator R can be defined equivalently as $R(M,A) = \{M' \mid M \triangle M' \subseteq A\}$. Hence, $A_t \supseteq E_{t,t+1} \supseteq E_{t,\tau}$, and $M_\tau \in R(M_t, A_t)$. Therefore, $M_{t+1} \rhd M_\tau$ and given the antisymmetry of $\rhd$, $M_{t+1} = M_\tau$, that is, $E_{t+1,\tau} = \emptyset$.

We finally show that $(E_{t,t'})_{t<t'}$ satisfies SRR*. From the proof of Proposition 3 in BGG, one can see that for any $(t_1,\ldots,t_n)$, $(M_{t_k}, M_{t_k})_k \in F^{(t_1,\ldots,t_n)}$. Fix a sequence $(t_1,\ldots,t_n)$ and define:

$$
\begin{aligned}
\tilde{M}_1 &= V^E_{t_1,t_n} \cup \underline{M}_{t_1}, \\
\tilde{M}_k &= V^E_{t_k,t_{k-1}} \cup \underline{M}_{t_k} \quad \text{for } k = 2,\ldots,n, \\
\tilde{M}'_k &= V^E_{t_k,t_{k+1}} \cup \underline{M}_{t_k} \quad \text{for } k = 1,\ldots,n-1 \text{ and} \\
\tilde{M}'_n &= V^E_{t_n,t_1} \cup \underline{M}_{t_n}.
\end{aligned}
$$

Given the definition of E, $V^E_{t,t'} = M_t \cap \underline{M}_{t'}$ and $V^E_{t',t} = M_{t'} \cap \underline{M}_t$; so $V^E_{t,t'} \cup \underline{M}_t \subseteq M_t$ and $V^E_{t',t} \cup \underline{M}_{t'} \subseteq M_{t'}$. Therefore, for any k, $\tilde{M}_k, \tilde{M}'_k \subseteq M_{t_k}$. Lemma 5 implies that $(\tilde{M}_k, \tilde{M}'_k)_k \in F^{(t_1,\dots,t_n)}$, that is, SRR* is satisfied.

(Sufficiency). Suppose that there exists an explanation E that satisfies SRR*, No Explanatory Gap (NEG) and Acyclic Explanation (AE).

Step 1: we build candidate sequences $(M_t, A_t)_t$.

Definition 14. *A sequence* $(M_t)_t$ *is* ***consistent*** *with* E *if for any* t, $M_t \in \mathscr{M}(\succsim_t)$, *and for any* $t < t'$, $M_t \triangle M_{t'} = E_{t,t'}$.

We want to show the existence of a consistent sequence. Define the following set for any $2 \leq \tau \leq T$:

$$\mathscr{C}^\tau_1(E) = \bigcap_{t=2}^{\tau} \bigcup_{M' \in \mathscr{M}(\succsim_t)} \{M \in \mathscr{M}(\succsim_1) \,:\, M \triangle M' = E_{1,t}\}. \tag{B.12}$$

Suppose that $\mathscr{C}^\tau_1(E) \neq \emptyset$ and fix $M_1 \in \mathscr{C}^\tau_1(E)$. M_1 and E uniquely define a sequence $(M_t)_{t=1,\dots,\tau}$ such that for any $2 \leq t \leq \tau$,

$$M_t = M_1 \triangle E_{1,t}. \tag{B.13}$$

Lemma 7. *The resulting sequence from (B.13) is consistent with* $(E_{t,t'})_{1 \leq t < t' \leq \tau}$.

Proof. By definition of this sequence, for any $t \geq 2$, $E_{1,t} = M_1 \triangle M_t$. Given the definition of $\mathscr{C}^\tau_1(E)$ and $M_1 \in \mathscr{C}^\tau_1(E)$, this means that $M_t \in \mathscr{M}(\succsim_t)$.

Furthermore, note that NEG implies that for any $t < t'$, $E_{1,t} \triangle E_{t,t'} = E_{1,t'}$, which is equivalent to $E_{t,t'} = E_{1,t} \triangle E_{1,t'}$. Hence:

$$M_t \triangle M_{t'} = (M_1 \triangle E_{1,t}) \triangle (M_1 \triangle E_{1,t'}) = E_{1,t} \triangle E_{1,t'} = E_{t,t'}.$$

□

Lemma 8. $\mathscr{C}_1^T(E) \neq \emptyset$.

Proof. We prove by induction on τ that $\mathscr{C}_1^\tau(E) \neq \emptyset$ for every $2 \leq \tau \leq T$.

The initialization for $\tau = 2$ directly follows from the definition of an explanation E. Fix $2 \leq \tau < T$ and suppose that $\mathscr{C}_1^\tau(E) \neq \emptyset$. We want to show that $\mathscr{C}_1^{\tau+1}(E) \neq \emptyset$.

Fix $M_1 \in \mathscr{C}_1^\tau(E)$ and construct a sequence $(M_t)_{t=1,\dots,\tau+1}$ as in (B.13). If $M_{\tau+1} \in \mathscr{M}(\succsim_{\tau+1})$, this ends the proof. Suppose on the contrary that $M_{\tau+1} \notin \mathscr{M}(\succsim_{\tau+1})$. From Proposition 6, this means that there exists $m \in \underline{M}_{\tau+1} \setminus M_{\tau+1}$. Construct then the following sequence $(\tilde{M}_t)_{t=1,\dots,\tau+1}$:

$$\tilde{M}_{\tau+1} = M_{\tau+1} + m,$$
$$\tilde{M}_t = \begin{cases} M_t - m & \text{if } m \in E_{t,\tau+1} \\ M_t + m & \text{if } m \notin E_{t,\tau+1} \end{cases} \quad \text{for } t = 1, \dots, \tau.$$

Note that for any $t \leq \tau$, if $m \in E_{t,\tau+1}$ this means that $m \in M_t$. Given that $m \in \underline{M}_{\tau+1}$ it must be that $m \notin \underline{M}_t$ as otherwise E would be ill-defined. Therefore, $(M_t - m) \in \mathscr{M}(\succsim_t)$, and more generally for any $t \leq \tau$, $\tilde{M}_t \in \mathscr{M}(\succsim_t)$. Furthermore, for any $t \leq \tau$, $\tilde{M}_1 \triangle \tilde{M}_t = E_{1,t}$. Indeed, by NEG $E_{1,t} = E_{1,\tau+1} \triangle E_{t,\tau+1}$. Hence, $\tilde{M}_1 \in \mathscr{C}_1^\tau(E)$. Similarly, $\tilde{M}_1 \triangle \tilde{M}_1 = E_{1,\tau+1}$. If $\tilde{M}_{\tau+1} \in \mathscr{M}(\succsim_{\tau+1})$, this ends the proof, if not, it means that there exists $m \in \underline{M}_{\tau+1} \setminus \tilde{M}_{\tau+1}$ and it suffices to reiterate the same process again, until you reach a point when the newly constructed sequence $(\hat{M}_t)$ is such that $\hat{M}_{\tau+1} \in \mathscr{M}(\succsim_{\tau+1})$. □

Finally, lemmata 7 and 8 together imply that there exists a consistent sequence, denote it $(M_t)_{t=1,\dots,T}$. The candidate sequence $(A_t)_{t=1,\dots,T-1}$ is simply $A_t = E_{t,t+1}$, that is $A_t = M_t \triangle M_{t+1}$.

Step 2: we show that (M_t) satisfies Strong Restricted Reversal (SRR) and Acyclicity from BGG, and conclude.

From the proof of Proposition 3 in BGG, it is enough to show that our candidate sequence (M_t) satisfies SRR and Acyclicity.

(Acyclicity). Let t and $\tau > t+1$ and suppose that $M_{t+1} \neq M_\tau$, that is, $E_{t+1,\tau} = M_{t+1} \triangle M_\tau \neq \emptyset$. AE implies that $E_{t,\tau} \not\subseteq E_{t,t+1}$, that there exists $m \in (M_t \triangle M_\tau) \setminus (M_t \triangle M_{t+1})$, which is exactly the statement of Acyclicity in BGG (in terms of sets).

(SRR). We want to show that for any $(t_1,\dots,t_n)$, $(M_{t_k}, M_{t_k})_k \in F^{(t_1,\dots,t_n)}$. Fix a sequence $(t_1,\dots,t_n)$ and define:

$$\begin{aligned}
\tilde{M}_1 &= V^E_{t_1,t_n} \cup \underline{M}_{t_1}, \\
\tilde{M}_k &= V^E_{t_k,t_{k-1}} \cup \underline{M}_{t_k} \quad \text{for } k = 2,\dots,n, \\
\tilde{M}'_k &= V^E_{t_k,t_{k+1}} \cup \underline{M}_{t_k} \quad \text{for } k = 1,\dots,n-1 \text{ and} \\
\tilde{M}'_n &= V^E_{t_n,t_1} \cup \underline{M}_{t_n}.
\end{aligned}$$

SRR* means that $(\tilde{M}_k, \tilde{M}'_k) \in F^{(t_1,\dots,t_n)}$.

Note first that fora any t,t', $V^E_{t,t'} = \underline{M}_{t'} \setminus E_{t,t'} = \underline{M}_{t'} \setminus (M_t \triangle M_{t'})$. Given that $\underline{M}_{t'} \subseteq M_{t'}$, $V^E_{t,t'} = \underline{M}_{t'} \cap M_t \subseteq M_t$ and $\underline{M}_t \cup V^E_{t,t'} = (\underline{M}_t \cup \underline{M}_{t'}) \cap M_t$. So for any k $\tilde{M}_k, \tilde{M}'_k \subseteq M_{t_k}$.

We simply need to show that for any $k \in \{1,\dots,n-1\}$, $(M_{t_k} \setminus \tilde{M}'_k) \cap \tilde{M}_{k+1} = (M_{t_{k+1}} \setminus \tilde{M}_{k+1}) \cap \tilde{M}'_k = \emptyset$, and $(M_{t_n} \setminus \tilde{M}'_n) \cap \tilde{M}_1 = (M_{t_1} \setminus \tilde{M}_1) \cap \tilde{M}'_n = \emptyset$. The conclusion will follow from lemma 6.

Let t,t', $m \in M_t \setminus (\underline{M}_t \cup V^E_{t,t'}) = M_t \setminus (\underline{M}_t \cup (\underline{M}_{t'} \cap M_t)) = M_t \setminus (\underline{M}_t \cup \underline{M}_{t'})$. Remember that $\underline{M}_{t'} \cup V^E_{t',t} = \underline{M}_{t'} \cup (\underline{M}_t \cap M_{t'})$, hence, $m \notin \underline{M}_{t'} \cup V^E_{t',t}$, which ends the proof that $(M_{t_k}, M_{t_k})_k \in F^{(t_1,\dots,t_n)}$ by applying lemma 6.

The rest of the proof simply replicates the construction from Theorem 1 and Proposition 3 in BGG, not using the sequence of revealed relevant attributes $(\underline{M}_t)_t$ but the candidate sequence $(M_t)_t$. □

Appendix C

Proofs of Chapter 3

C.1 Proofs of Section 3.3

Proof of Proposition 7. Let c be an RSC rationalized by $\langle \mathscr{T}, \{\lambda_T\}, u, v \rangle$. The *if* part is left to the reader as it simple results from an application of the choice procedure of an RSC. We prove the *only if* part. Consider x, y, z such that $z = c\{x, y, z\}$ and $x = c\{x, z\}$, so $x \mathbf{R}^c y$. One can easily check that it is not possible that x, y, z are either all in the same type, or all in different types: in both cases, the choice results from the maximization of a unique function, so it must satisfy the *Weak Axiom of Revealed Preferences*. Hence exactly two among them must be of the same type, denote it T. Given that $z = c\{x, y, z\}$, this means that $z \in d\{x, y, z\}$ and z is the best element of its type according to u. Consequently, x is not of the same type as z as otherwise this would imply that $z = c\{x, z\}$. This also implies that $z \in d\{x, z\}$ and thus $v(x) > v(z)$. Hence $x \notin d\{x, y, z\}$, as otherwise it would imply $x = c\{x, y, z\}$, which is only possible if $x, y \in T$ and $u(y) > u(x)$. Therefore, $y \in d\{x, y, z\}$ and $z = c\{x, y, z\}$ implies that $v(z) > v(y)$. □

Proof of Proposition 8. This simply follows from the proof of sufficiency of theorem 5. □

Proof of Corollary 2. Point *(i)*. Let $\langle \mathcal{T}, \{\lambda_T\}_T, u, v\rangle$ a minimal RS-structure that rationalizes c, $T \neq T_0^c$ and $x \in T$. There must exist $x \in T$ such that $u(x) < \lambda_T$, as otherwise we would conclude from proposition 7 that $T = T_0^c$. Hence, there exists a sequence of options $(x_k)_{k=1}^n$ in T such that and $\lambda_T > u(x_1) > \cdots > u(x_n)$ and $\{x \in T : u(x) < \lambda_T\} = \{x_1, \ldots, x_n\}$.

From point iii in the definition of an RSC, there exists $k^\star \in \{1, \ldots, n\}$ such that $v(x_{k^\star}) = \max v\big(\{x_1, \ldots, x_n\}\big)$ and for any $k' < k < k^\star$, or $k' > k > k^\star$, $v(x_{k'}) < v(x_k) < v(x_{k^\star})$. Hence, it is sufficient to show that $x_1 \mathbf{R}^c x_T$ and $x_n \mathbf{R}^c x_T$.

First, $x_1 \in T \neq T_0^c$ and $u(x_1) < \lambda_T$ mean that there exists $y \in T$ such that $x_1 \mathbf{R}^c y$ as otherwise this would contradict the definition of λ_T from proposition 8. Hence, from proposition 7, there exists also $z \notin T$, such that $u(x_1) < u(y)$ and $v(x_1) > v(z) > v(y)$. Furthermore, given the definition of x_1, $u(y) \geq \lambda_T$, hence $v(y) = u(y)$. Therefore, $v(x_1) > v(z) > \lambda_T = v(x_T)$ and $u(x_1) < \lambda_T = u(x_T)$, that is, $x_1 \mathbf{R}^c x_T$.

Second, given that $u(x_n) = \min u(T)$ and $x_n \in T \neq T_0^c$, there exists $y \in T$ such that $x_1 \mathbf{R}^c y$ as otherwise this would contradict the definition of T_0^c. Hence, there exists also $z \notin T$, such that $u(x_1) < u(y)$ and $v(x_1) > v(z) > v(y)$. This implies that $u(y) > u(x_{k^\star})$, which in turn implies that $v(y) \geq v(x_T) = \lambda_T$. We can conlude similarly as in the previous paragraph that $x_n \mathbf{R}^c x_T$.

Point *(ii)* directly follows from point *(i)* and propositions 7 and 8. □

Proof of Proposition 9. The proof of *(i)* directly follows from the proof of sufficiency of theorem 5. The proof of the *if* part of *(ii)* is not complicated and thus left to the readers. We only prove the *only if* part.

The fact that f must be increasing on $u(T)$ for every $T \in \mathcal{T}$ simply follows from the fact that the function u represents the binary choices within each type. $f|_{u(F)} = g|_{u(F)}$ is a direct consequence of the requirement that utility and the reaction functions be equal on F.

We now prove that g must be increasing on $v(X)$. Suppose by contradiction that there exists $x, y \in X$ such that $v(x) > v(y)$ but $g \circ v(x) \leq g \circ v(y)$. Note that there must exist a type, denote it T, such that $x, y \in T$, as otherwise $v(x) > v(y) \implies x = c\{x, y\}$, which cannot be accommodated by $\langle \mathscr{T}, F, f \circ u, g \circ v \rangle$ if $g \circ v(x) \leq g \circ v(y)$. Define $x^\star = \arg\max v(T \setminus F)$ and x_T as in corollary 2.

(1) Consider the case where $u(x) > u(y)$. If $u(y) \geq \lambda_T$, this would mean that $x, y \in F$, in which case, given that $u|_F = v|_F$, $f|_{u(F)} = g|_{u(F)}$ and f is increasing on $u(T)$, it is impossible that $g \circ v(x) \leq g \circ v(y)$. If $u(y) \leq \lambda_T$, it means that $y \notin F$. The fact that $v(x) > v(y)$ implies that there exists $z \notin T$ such that $x = c\{x, z\}$ while $z = c\{y, z\}$, that is $v(x) > v(z) > v(y)$. Hence, $g \circ v(x) \leq g \circ v(y)$ cannot accommodate these choices.

(2) Consider the case where $u(x) < u(y)$. Then necessarily $u(x) < \lambda_T$, that is $x \notin F$. Given that $v(x) > v(y)$, there exists no $z \notin T$ such that $z = c\{x, z\}$ while $y = c\{y, z\}$. Conversely, if there exists $z \notin T$ such that $x = c\{x, z\}$ while $z = c\{y, z\}$, that is $v(y) > v(z) > v(x)$, then again $g \circ v(x) \leq g \circ v(y)$ cannot accommodate these choices. If it is note the case, that is for every $z \notin T$ such that $x = c\{x, z\} \iff y = c\{y, z\}$, then $g \circ v(x) \leq g \circ v(y)$ does not satisfy the requirement of RS*-structure. □

C.2 Proof of Theorem 5

Proof of the necessity. Let $\mathscr{S} = \langle \mathscr{T}, \{\lambda_T\}_T, u, v \rangle$ be an RS-structure that rationalizes c. We denote $T(x)$ the type of the option x. Furthermore, for any $x, y, z \in X$, if $z = c\{x, y, z\}$ and $x = c\{x, z\}$, we write $x \mathbf{R}^c_z y$. Hence the definition of P^c can now be written: $x P^c y \iff$ there exists z, t such that $y R^c_t z$ and for any such pair $x R^c_t z$.

The following lemma shows that WARP is satisfied for each collection of menus that contains options of the same type.

Lemma 9. *For any $T \in \mathscr{T}$ and $A \subset B \subseteq T$, if $c(B) \in A$, then $c(A) = c(B)$.*

Proof. Let $c(B) = x$. Given that $B \subseteq T$, $x = c(B)$ implies that $c(B) = d(B)$, that is, $u(x) > u(y)$ for any $y \in B$, $y \neq x$. Because $A \subset B$, it means that $x = d(A)$, and therefore $x = c(A)$. □

Exp. Let $x \in X$ and $A, B \in \mathscr{X}$ such that $x = c(A) = c(B)$. This means that $x \in d(A) \cap d(B)$. Hence, $u(x) > u(y)$ for all $y \in (A \cup B) \cap T(x)$, $y \neq x$, which implies that $x \in d(A \cup B)$. Moreover, $x = c(A) = c(B)$ implies that $v(x) > v(z)$ for all $z \neq x$ such that $z \in d(A) \cup d(B)$. Besides, $d(A \cup B) \subseteq d(A) \cup d(B)$, hence $v(x) > v(z)$ for all $z \neq x$, $z \in d(A \cup B)$. Hence, $x = c(A \cup B)$.

R-Tran. Let $x, y, z \in X$ such that $x\mathbf{R}^c y$ and $y\mathbf{R}^c z$. By definition,

$$x\mathbf{R}^c y \implies \exists t \in X, t = c\{x, y, t\} \text{ and } x = c\{x, t\},$$
$$y\mathbf{R}^c z \implies \exists t' \in X, t' = c\{y, z, t'\} \text{ and } y = c\{y, t'\}.$$

Proposition 7 implies that $T(x) = T(y) = T(z)$. Coupled with lemma 9, this shows that $T(t') \neq T(x) \neq T(t)$. Given that **Exp** is satisfied, we also know that $z = c\{z, y\}$ and $y = c\{y, x\}$. Hence, $u(z) > u(y) > u(x)$ and $v(x) > v(t) > v(y) > v(t') > v(z)$. Therefore, $t = c\{x, z, t\}$ and $x = c\{x, t\}$, which means that $x\mathbf{R}^c z$.

R-NTran (i). Let $x, y, z \in X$ such that $y = c\{x, y\}$, $z = c\{y, z\} = c\{x, z\}$, $\neg[x\mathbf{R}^c y]$ and $\neg[y\mathbf{R}^c z]$. Assume by contradiction that $x\mathbf{R}^c z$. Then there exists t such that $t = c\{x, z, t\}$ and $x = c\{x, t\}$ and, by proposition 7 and lemma 9, $T(x) = T(z) \neq T(t)$. Moreover, we have that $v(x) > v(t) > v(z)$ so that $v(x) > v(z)$. Hence, if $T(y) \neq T(x)$,

then $z = c\{y,z\} = c\{x,z\}$ implies that

$$v(y) \underbrace{>}_{y=c\{x,y\}} v(x) > v(z) \underbrace{>}_{z=c\{y,z\}} v(y), \tag{C.1}$$

a contradiction. Now if $T(y) = T(x)$, then $u(z) > u(y) > u(x)$. But then either $v(y) > v(t)$, and then $y\mathbf{R}^c z$, or $v(t) > v(y)$ and then $x\mathbf{R}^c y$. In both cases we have a contradiction.

R-NTran (ii). Let $x,y,z \in X$ such that $y = c\{x,y\}$, $z = c\{y,z\} = c\{x,z\}$, $\neg[x\mathbf{P}^c y]$ and $\neg[y\mathbf{P}^c z]$. Assume by contradiction that $x\mathbf{P}^c z$. Hence, there exists t,u such that $z\mathbf{R}^c_u t$ and for any such t,u, $x\mathbf{R}^c_u t$. By proposition 7, $T(z) = T(t) = T(x) \equiv T$, $u(t) > u(z) > u(x)$ and $v(z), v(x) > v(t)$, which implies that $u(z) < \lambda_T$. Suppose that $y \notin T$, hence $v(z) > v(y) > v(x)$. Let t be such that both $xR^c t$ and $zR^c t$. Then it must be that $zR^c_y t$ but $\neg[xR^c_y t]$, which contradicts the fact that $xP^c z$. Suppose then that $y \in T$. This means that $u(z) > u(y) > u(x)$. Because $u(z) < \lambda_T$, then $u(y) < \lambda_T$. Given point iii in definition 7, we have that $v(y) > \min\{v(x), v(z)\}$. If $v(y) > v(z)$, then proposition 7 implies that $y\mathbf{P}^c z$, a contradiction. Hence $v(z) > v(y) > v(x)$. Let t,u be such that, $z\mathbf{R}^c_u t$, then $x\mathbf{R}^c_u t$, which means that $v(x) > v(u) > v(t)$, and therefore $v(y) > v(u) > v(t)$, and hence $y\mathbf{R}^c_u t$. This proves that $y\mathbf{P}^c z$, again a contradiction.

R-Con. Let $x,y,z \in X$ such that $x\mathbf{R}^c y$, $x\mathbf{R}^c z$, $z = c\{y,z\}$, and there exists no t such that $y\mathbf{R}^c t$. Proposition 7 implies that $T(z) = T(x) = T(y) \equiv T$ and $u(z) > u(y)$. Furthermore, by proposition 8, it is without loss of generality to assume that $\mathscr{S}$ is minimal, and by proposition 7, $y\mathbf{R}^c t$ for no t implies that $u(y) \geq \lambda_T$, and hence $u(z) \geq \lambda_T$. Let $u \in X$ such that $u\mathbf{R}^c z$, so $u \in T$ and there exists $t \notin T$ such that $v(u) > v(t) > v(z) = u(z) > u(u)$. Hence $u(u) < \lambda_T \leq u(y)$. This means that $v(u) > v(t) > v(y) = u(y) > u(u)$. Hence, $u\mathbf{R}^c y$. This completes the proof of *(i)* in **R-Con**. Now assume that $u\mathbf{R}^c y$. This means that $u \in T$ and $u(z) > u(y) > u(u)$,

hence $z = c\{x,z\}$, which proves *(ii)* in **R-Con**.

R-Mon. Let $x,y,z \in X$ such that $z = c\{y,z\}$, $y = c\{x,y\}$, $x\mathbf{R}^c t$, and $z\mathbf{R}^c t$ for some t. Assume that $x\mathbf{R}^c y$. By proposition 7, this means that $T(x) = T(y) = T(t) = T(z) \equiv T$, $\lambda_T > u(z) > u(y) > u(x)$, $v(y) < v(x)$. Point (iii) of definition 7 implies that $v(y) > v(z)$, from which we can conclude that $y\mathbf{P}^c z$. This proves *(i)*. Assume now that $x\mathbf{P}^c z$ and $x\mathbf{P}^c y$. By proposition 7, $T(x) = T(y) = T(z) \equiv T$ adn $\lambda_T > u(z) > u(y) > u(x)$. Point iii of definition 7 implies that $v(y) > \min\{v(x), v(z)\}$. If $v(y) > v(x)$, then for any t,u such that $z\mathbf{R}^c_u t$, given that $x\mathbf{R}^c_u t$, it must be that $y\mathbf{R}^c_u t$ and hence $y\mathbf{P}^c z$. Similarly, if $v(y) > v(z)$, $y\mathbf{P}^c z$ follows directly, which ends the proof of *(ii)*. □

Proof of the sufficiency. Let define the binary relation $\succ \subset X^2$ by $x \succ y$ if and only if $x = c\{x,y\}$ or $x = y$. It is clear that $\succ$ is complete and antisymmetric. For any transitive and complete binary relation $>$ defined on a set A, we write $\max(A,>) \equiv \{x \in A \mid x > y, \forall\ y \in A\}$. When $>$ is a linear order, with a slight abuse of notation, when no confusion can be made, we indifferently write $\max(A,>)$ for the singleton or for the element of the singleton. We define analogously $\min(A,>)$.

Lemma 10. *Let K a subset of X such that,*

$$\big((x,y) \in K^2 \iff \neg[x\mathbf{R}^c y] \text{ and } \neg[y\mathbf{R}^c x]\big), \tag{C.2}$$

then $\succ$ restricted to K^2 is a linear order and for all $K' \subseteq K$, $c(K) \succ y$ for all $y \in K'$.

Proof. Let K satisfying (C.2), $x,y,z \in K$ and $x \succ y \succ z$. Suppose by contradiction that $z \succ x$. If $x = c\{x,y,z\}$, then $z\mathbf{R}^c y$, which contradicts that $(y,z) \in K^2$ and K satisfies (C.2). The same reasoning applies if either $y = c\{x,y,z\}$ or $z = c\{x,y,z\}$. Hence, we conclude that $x \succ z$.

Moreover, let $K' \subseteq K$, the transitivity of $\succ$ on K, implies that there exists $x \in K'$ such that $x \succ y$ for any $y \in K'$. By **Exp**, $x = c(K')$. □

Define now

$$X^{\downarrow} = \bigcup_{y \in X} \{x \in X : x\mathbf{R}^c y\}$$
$$X^{\uparrow} = \bigcup_{y \in X} \bigcap_{t \in X} \{x \in X : y\mathbf{R}^c x, \neg[x\mathbf{R}^c t]\}$$

Let $\tilde{X} = X^{\uparrow} \cup X^{\downarrow}$ and for all $x \in X^{\downarrow}$, $R^{\uparrow}(x) = \{y \in X^{\uparrow} : x\mathbf{R}^c y\}$.

Lemma 11. *If $x \in X^{\downarrow}$, then $R^{\uparrow}(x) \neq \emptyset$.*

Proof. Let $x \in X^{\downarrow}$, i.e. $x\mathbf{R}^c y$ for some $y \in X$. If $y \in X^{\uparrow}$, this terminates the proof. Suppose that $y \notin X^{\uparrow}$, then there exists z_1 such that $y\mathbf{R}^c z_1$, which by **R-Tran** implies that $x\mathbf{R}^c z_1$. Either $z_1 \in X^{\uparrow}$, which ends the proofs, or there exists z_2 such that $z_1\mathbf{R}^c z_2$, which again by **R-Tran** implies that $x\mathbf{R}^c z_2$. At each step k, we replicate the same reasoning. Because X is finite, there must exist n such that for all $t \in X$, $\neg[z_n\mathbf{R}^c t]$, i.e., $z_n \in X^{\uparrow}$. Yet, **R-Tran** also implies that $x\mathbf{R}^c z_n$. Hence, $R^{\uparrow}(x) \neq \emptyset$. □

Note that lemma 10 implies that $\succ$ is transitive on $X^{\uparrow}$. Hence, lemma 11 implies the existence, for all $x \in X^{\downarrow}$, of $m(x)$, defined by:

$$m(x) \equiv \min(R^{\uparrow}(x), \succ). \tag{C.3}$$

Lemma 12. *For all $x, y \in X^{\downarrow}$, if $R^{\uparrow}(x) \cap R^{\uparrow}(y) \neq \emptyset$, then $m(x) = m(y)$;*

Proof. Let $x, y \in X^{\downarrow}$. Assume there exists $t \in R^{\uparrow}(x) \cap R^{\uparrow}(y)$ and let $t' = m(x)$. We show that $t' \in R^{\uparrow}(y)$. If $t = t'$ there is nothing to prove. If $t \neq t'$, then by definition of t' and since $t \in R^{\uparrow}(x)$, we have $t \succ t'$. Given that $t' \in X^{\uparrow}$ we have that $\neg[t'\mathbf{R}^c z]$ for any $z \in X$. Since $x\mathbf{R}^c t$, $x\mathbf{R}^c t'$ and $t \succ t'$, by **R-Con**(i) $y\mathbf{R}^c t$, implies that $y\mathbf{R}^c t'$, i.e. $t' \in R^{\uparrow}(y)$.

We prove symmetrically that $m(y)$ belongs to $R^{\uparrow}(x)$. Hence, by definition $m(x) \succ m(y)$ and $m(y) \succ m(x)$. Given that $\succ$ is a linear order on $x^{\uparrow}$, $m(x) = m(y)$. □

Since X is finite there exists n^* such that we can index the set $\{m(x) : x \in X^{\downarrow}\}$ by a sequence $(m(i))_{1 \leq i \leq n^*}$ such that $i \neq j \iff m(i) \neq m(j)$. Define now for all $1 \leq i \leq n^*$:

$$
\begin{aligned}
&T^{\downarrow}(i) = \{x \in X^{\downarrow} : x \mathbf{R}^c m(i)\},\\
&T^{\uparrow}(i) = \{x \in X^{\uparrow} : \exists y \in X^{\downarrow}, y \mathbf{R}^c m(i), y \mathbf{R}^c x\}, \text{and}\\
&T(i) = T^{\uparrow}(i) \cup T^{\downarrow}(i).
\end{aligned}
$$

Define finally:

$$
T(0) \equiv X \setminus \tilde{X} = \bigcap_{y \in X} \bigcap_{t \in X} \{x \in X : \neg[x \mathbf{R}^c y], \neg[t \mathbf{R}^c x]\}.
$$

These will define the types. We denote $\mathcal{T} = \{T(i) : 0 \leq i \leq n^*\}$ the collection of types.

Lemma 13. *$\mathcal{T}$ forms a partition of X.*

Proof. Given the definition of $T(0)$, it is sufficient to show that the collection $\{T(i) : 1 \leq i \leq n^*\}$ partitions $\tilde{X}$.

We first show that $\tilde{X} = \bigcup_{1 \leq i \leq n^*} T(i)$. Note that for all $1 \leq i \leq n^*$, if $x \in T(i)$, then there exists y such that $x \mathbf{R}^c y$ or $y \mathbf{R}^c x$, so that $x \in \tilde{X}$. Hence, $\bigcup_{i \leq n} T(i) \subseteq \tilde{X}$. Similarly, if $x \in \tilde{X}$, then either $x \in X^{\downarrow}$ or $x \in X^{\uparrow}$. If $x \in X^{\downarrow}$, then $x \mathbf{R}^c y$ for some $y \in X$ and by (C.3), $x \mathbf{R}^c m(x)$, i.e. $x \in T^{\downarrow}(i)$ for some $1 \leq i \leq n^*$. If $x \in X^{\uparrow}$, then $y \mathbf{R}^c x$ for some $y \in X$ and $\neg[x \mathbf{R}^c z]$ for all $z \in X$. But then $x \in R^{\uparrow}(y)$ and (C.3) implies that $y \mathbf{R}^c m(y) = m(i)$ for some $1 \leq i \leq n^*$. Therefore $x \in T^{\uparrow}(i)$. Hence, in both cases, $x \in \bigcup_{1 \leq i \leq n^*} T(i)$.

We now assume that for some $1 \leq i, j \leq n^*$, $x \in T(i) \cap T(j)$, and show that this implies $i = j$.

Case 1: Assume $x \in T^{\downarrow}(i)$. Then, because $X^{\uparrow}$ and $X^{\downarrow}$ are disjoint, x necessarily belongs to $T^{\downarrow}(j)$. $x \in T^{\downarrow}(i)$ means that $x\mathbf{R}^c m(i)$. By definition of the $m(i)$'s, there exists y such that $m(y) = m(i)$. Applying lemma 12, we conclude that $m(x) = m(y) = m(i)$. Similarly, we prove that $m(x) = m(j)$. Hence $m(i) = m(j)$ and therefore $i = j$.

Case 2: Assume $x \in T^{\uparrow}(i)$. Then, because $X^{\uparrow}$ and $X^{\downarrow}$ are disjoint, $x \in T^{\uparrow}(j)$. Hence, there exists $y_i, y_j \in X^{\downarrow}$ such that $y_i\mathbf{R}^c m(i)$, $y_j\mathbf{R}^c m(j)$, $y_i\mathbf{R}^c x$, and $y_j\mathbf{R}^c x$. Hence, $x \in R^{\uparrow}(y_i) \cap R^{\uparrow}(y_j)$, which by lemma 12, implies that $m(y_i) = m(y_j)$. Using the same argument as in *case 1*, we conclude that $m(i) = m(y_i) = m(y_j) = m(j)$, which means that $i = j$. □

Note that, given lemma 13, $T(x)$ is well defined as the type of the option $x \in X$, i.e. $T(x) = T(i) \iff x \in T(i)$.

Lemma 14. *For all* $1 \leq i \leq n^*$, $x \in T^{\downarrow}(i)$ *and* $y \in T^{\uparrow}(i)$, $x \prec y$.

Proof. If $y = m(i)$ this follows directly. If $y \neq m(i)$, there exists $z \in X$ such that $z\mathbf{R}^c y$ and $z\mathbf{R}^c m(i)$. Hence, $y, m(i) \in R^{\uparrow}(z)$ and $y \succ m(i)$. Moreover, $m(i) \in X^{\uparrow}$ so that $\neg[m(i)\mathbf{R}^c t]$ for any t. Since $x\mathbf{R}^c m(i)$, **R-Con**(ii) implies that $y \succ x$. □

Lemma 15. *For any* $x, y \in X^{\downarrow}$, *if* $x\mathbf{R}^c y$, *then* $x\mathbf{P}^c y$.

Proof. Let z, t such that $y\mathbf{R}^c_t z$. By**R-Tran**, it must be that $xR^c z$. Suppose first that $x \succ t$, then $x \succ t \succ z \succ x$. Then, if $zR^c t$, **R-Tran** implies that $yR^c t$, a contradiction. If $tR^c x$, **R-Tran** implies that $tR^c z$, a contradiction. Hence it must be that $xR^c_t z$, which means that $xP^c y$.

Suppose now that $t \succ x$. Let t' such that $xR^c_{t'} y$. Hence $x \succ t' \succ y$.

Case 1: $t' \succ t$. Then $x \succ t' \succ t \succ x$. First note that $t' \succ y \succ t$ and $\neg[tR^c y]$ (as otherwise this would imply $tR^c z$), $\neg[yR^c t']$ (as otherwise this would imply $xR^c t'$). Hence **R-NTran**(i) implies that $\neg[tR^c t']$.

1.1: $t'R^c x$. This implies that $t'R^c z$. Hence $y \succ x \succ t'$. Given that in addition $yR^c z$ we can apply **R-Mon**(i) to conclude that $xP^c y$, which contradicts that $t \succ x$.

1.2: $xR^c t$. Given that $t \succ y$, this means that by lemma 14, $t \in T^{\downarrow}(y)$ (given that $y \in T^{\downarrow}(y)$). Given that both $xR^c z$ and $yR^c z$, we can apply **R-Mon**(i) to conlcude that $tP^c y$, which means that $tR^c z$ a contradiction.

Case 2: $t \succ t'$. Then $y \succ t \succ t' \succ y$. First note that $t \succ x \succ t'$ and $\neg[t'R^c x]$ (as otherwise this would imply $t'R^c y$), $\neg[xR^c t]$ (as otherwise this would imply $xP^c y$ as in case 1.2). Hence **R-NTran**(i) implies that $\neg[tR^c t']$.

2.1: $yR^c t'$. This implies that $xR^c t'$, a contradiction.

2.2: $t \succ y$. This implies that $tR^c z$, again a contradiction.

Hence it must be that $x \succ t$ and therefore $xP^c y$. □

We now prove that $\succ$ is transitive on every type $T(i)$.

Lemma 16. *For all* $0 \leq i \leq n^*$, *the relation* $\succ$ *is transitive on* $T(i)$.

Proof. That $\succ$ is transitive on $T(0)$ is a direct consequence of lemma 10. We now focus on $1 \leq i \leq n^*$. We first show that $\succ$ is transitive on $T^{\downarrow}(i)$. Let $x, y, z \in T^{\downarrow}(i)$ such that $x \succ y \succ z$. Assume by contradiction that $z \succ x$. Suppose (w.l.o.g) that $x = c\{x, y, z\}$. In this case, $z\mathbf{R}^c y$. Given that $x\mathbf{R}^c m(i)$, $z\mathbf{R}^c m(i)$, and $x \succ y \succ z$, **R-Mon**(i) entails $y\mathbf{P}^{\mathbf{c}} x$. But since $y, z \in X^{\downarrow}$, $z\mathbf{R}^c y$ implies $z\mathbf{P}^c y$, by lemma 15. Hence, by the transitivity of P^c (by definition), $z\mathbf{P}^c x$, which contradicts $z \succ x$.

Finally, we prove that $\succ$ is transitive on each type. Let i and $x, y, z \in T(i)$ such that $x \succ y \succ z$. If $x \in T^{\downarrow}(i)$ then, according to the first part of the proof, $y \in T^{\downarrow}(i)$ and therefore similarly $z \in T^{\downarrow}(i)$. Similarly, if $z \in T^{\uparrow}(i)$, the first part of the proof implies that $y \in T^{\uparrow}(i)$, which in turn also triggers that $x \in T^{\uparrow}(i)$. In both cases, we proved that $\succ$ is transitive on $T^{\downarrow}(i)$ and on $T^{\uparrow}(i)$ (a consequence of lemma 10). The last case is if $x \in T^{\uparrow}(i)$ and $z \in T^{\downarrow}(i)$, but then $x \succ z$ follows from lemma 14. □

For any menu A we define:

$$d(A) \equiv \{x \mid x = \max(T(x) \cap A, \succ)\}.$$

Lemma 16 implies that $d(A)$ is well defined. Furthermore, a direct implication of lemma 10 is that $\succ$ is transitive on $d(A)$. Hence we can state the following lemma:

Lemma 17. *For any $A \in \mathscr{X}$,*

$$c(A) = \max(d(A), \succ) \tag{C.4}$$

Proof. For any menu A, denote $i(A) = \#\{i \mid T(i) \cap A \neq \emptyset\}$. We prove that for any $1 \leq n \leq n^* + 1$, for any A such that $i(A) = n$, (C.4) holds.

If $i(A) = 1$, the conclusion follows from lemma 16. Assume now that $i(A) = 2$. Let $x, y \in A$ be such that $T(x) \cap T(y) = \emptyset$, $x = \max(T(x) \cap A, \succ)$, $y = \max(T(y) \cap A, \succ)$, and $y \succ x$. Assume by contradiction that $y \neq c(A)$. By definition of y, $y \succ z$ for any $z \in T(y) \cap A$. Hence, there must exist $z \in T(x)$ such that $z \succ y$ and $y \neq c\{x, y, z\}$, otherwise **Exp** would imply that $y = c(A)$. This implies that $x \succ z \succ y \succ x$. Since $y \neq c\{x, y, z\}$, this is only possible if either $y\mathbf{R}^c z$ or $x\mathbf{R}^c y$, which in any case contradicts that $x, z \notin T(y)$ (given that, according to lemma 13, types partition X). Hence we conclude that $y = c(A)$.

Then fix $3 \leq n \leq n^* + 1$ and let A a menu such that $i(A) = n$. We denote $y = \max(d(A), \succ)$. Given the preceding proof for any menu A' such that $i(A') = 2$, for any $z \in A$, $y = c\Big(\big(T(y) \cup T(z)\big) \cap A\Big)$. This implies by **Exp** that $y = c(A)$. □

Lemma 18. *For any $x, y, z \in X$, if $xR^c yP^c z$, then $xR^c z$.*

Proof. Let $x, y, z \in X$ such that $xR^c yP^c z$. Let t such that $xR^c_t y$. A consequence of lemma 16 is that $t \notin T(x) = T(y) = T(z) \equiv T$. If $t \succ z$ then it suffices to prove that $xR^c z$. Suppose on the contrary that $z \succ t$. We want to show that there exists u such that $zR^c_u t$, contradicting $yP^c z$. Let u such that $zR^c u$, hence $u \in T$ and $u \succ z$. If $t \succ u$, then it suffices to prove that $zR^c_t u$. Suppose by contradiction that for any

such u, $u \succ t$. Then there exists t' such that $zR^c_{t'}u$, in which case $yR^c_{t'}u$ and thus $t' \succ u \succ t \succ y \succ t'$. Suppose first that $t' \succ t$. Then $t' \succ t \succ y \succ t'$. Because $t, t' \notin T$, this implies that $tR^c t'$. But at the same time, given that $t' \succ u \succ t$, $t' \succ t$, $\neg[tR^c u]$ and $\neg[uR^c t']$, we can apply **R-Ntran**(i) to conclude that $\neg[tR^c t']$, a contradiction. An analogous reasoning applies for the other case $t \succ t'$ to obtain a contradiction. This ends the proof. □

For all $0 \leq i \leq n^*$, define $\trianglerighteq_i$ on $T(i)$, for any $x, y \in T(i)$, $x \trianglerighteq_i y$ if one of the following cases is satisfied:

1. $xR^c y$;
2. $x \succ y \wedge \neg[yR^c x]$;
3. $xP^c y \wedge \neg[xR^c y]$.

Denote $\triangleright_i$ the asymmetric part of $\trianglerighteq_i$.

Lemma 19. *For all $0 \leq i \leq n^*$, $\trianglerighteq_i$ is complete and transitive.*

Proof. Note that by definition $\trianglerighteq_0$ is simply equal to $\succ$ on $T(0)$, hence it is complete and transitive. Let, $1 \leq i \leq n^\star$ and $x, y, z \in T(i)$ such that $x \trianglerighteq_i y \trianglerighteq_i z$. We detail all the cases.

1. $x \succ y$. Together with $x \trianglerighteq_i y$, this implies that $\neg[yR^c x]$.
   (a) $\neg[yP^c x]$.
      i. $y \succ z$. This implies that $x \succ z$ and $\neg[zR^c y]$, which in turn implies that $\neg[zR^c x]$ by **R-NTran**(i). Which implies that $x \trianglerighteq_i z$.
      ii. $z \succ y$.
         A. $yR^c z$.

i. $x \succ z$. $zR^c x \implies yR^c x$, a contradiction, hence $\neg[zR^c x]$ and $x \trianglerighteq_i z$.

ii. $z \succ x$. Given that $\neg[yR^c x]$ and $yR^c z$, **R-NTran**(i) implies $xR^c z$, and therefore $x \trianglerighteq_i z$.

B. $\neg[yR^c z]$. This implies that $yP^c z$.

i. $x \succ z$. $zR^c x$ and $yP^c z$ would imply that $yR^c z$ (lemma 18), a contradiction. Hence $\neg[zR^c x]$ and $x \trianglerighteq_i z$.

ii. $z \succ x$. By **R-NTran**(ii), $yP^c z$ and $\neg[yP^c x]$ imply that $xP^c z$, and thus $x \trianglerighteq_i z$.

(b) $yP^c x$.

i. $y \succ z$. This implies that $\neg[zR^c y]$ and $x \succ z$. Then, $\neg[yR^c x] \implies \neg[zR^c x]$. Hence $x \trianglerighteq_i z$.

ii. $z \succ y$.

A. $yR^c z$.

i. $x \succ z$. $zR^c x$ would imply $yR^c x$, a contradiction. Hence $\neg[zR^c x]$ and $x \trianglerighteq_i z$.

ii. $z \succ x$. If $\neg[x\mathbf{R}^c z]$, then $\neg[y\mathbf{R}^c x]$ implies that $\neg[y\mathbf{R}^c z]$, a contradiction. Hence, $x\mathbf{R}^c z$ and $x \trianglerighteq_i z$.

B. $\neg[y\mathbf{R}^c z]$. This implies that $y\mathbf{P}^c z$.

i. $z \succ x$. By **R-Mon**(ii), $yP^c z$ and $yP^c x$ imply that $xP^c z$. Hence, $x \trianglerighteq_i z$.

ii. $x \succ z$. $zR^c x$ together with $yP^c z$ would imply $yR^c x$, a contradiction. Hence $\neg[zR^c x]$ and thus $x \trianglerighteq_i z$.

2. $y \succ x$.

(a) $xR^c y$.

i. $z \succ y$. This implies that $z \succ x$.

A. $yR^c z$. Then **R-Tran** implies that $xR^c z$, i.e., $x \trianglerighteq_i z$.

B. $\neg[yR^c z]$. This implies that $yP^c z$. $xR^c yP^c z \implies xR^c z$. Hence $x \trianglerighteq_i z$.

ii. $y \succ z$. This implies that $\neg[zR^c y]$.

A. $z \succ x$. If $\neg[xR^c z]$, then $\neg[zR^c y]$ implies that $\neg[xR^c y]$ a contradiction. Hence $xR^c z$ and $x \trianglerighteq_i z$.

B. $x \succ z$. $zR^c x$ would imply $zR^c y$, a contradiction, hence $x \trianglerighteq_i z$.

(b) $\neg[xR^c y]$. This implies that $xP^c y$.

i. $z \succ y$. This implies that $z \succ x$.

A. $yR^c z$. Then $xP^c y$ implies that $xR^c z$, i.e., $x \trianglerighteq_i z$.

B. $\neg[yR^c z]$. This implies that $yP^c z$. $xP^c yP^c z \implies xP^c z$. Hence $x \trianglerighteq_i z$.

ii. $y \succ z$. This implies that $\neg[zR^c y]$.

A. $z \succ x$. $xP^c y$ implies that $y \in T^{\downarrow}(i)$, and so $y \succ z$ implies that $z \in T^{\downarrow}(i)$ (lemma 14). So let $u \in T(i)$ such that $zR^c u$; $u \succ z$ and thus $u \succ x$. Let $t \notin T(i)$ such that $t = c\{u,z,t\}$ and $z \succ t$. Then because $\neg[zR^c y]$, it must be that $y \succ t$, which in turn implies that $x \succ t$ (as if $t \succ x$, we would conclude from **R-NTran**, $\neg[xP^c t]$ and $\neg[tP^c y]$ that $\neg[xP^c y]$). But from lemma 17, we get that $t = c\{x,u,t\}$, hence $xP^c z$ and $x \trianglerighteq_i z$.

B. $x \succ z$. $zR^c x$ would imply $zR^c y$, a contradiction, hence $\neg[zR^c x]$ and $x \trianglerighteq_i z$.

□

Lemma 20. *For any* $1 \leq i \leq n^\star$, $x,y,z \in T^{\downarrow}(i)$ *such that* $x \prec y \prec z$*:*

1. *if* $x \rhd_i y$, *then* $y \trianglerighteq_i z$;

2. *if* $z \rhd_i y$, *then* $y \trianglerighteq_i x$.

Proof. 1. $x \rhd_i y$ implies that $xR^c y$. Given that $x, z \in T^{\downarrow}(i)$, $xR^c m(i)$ and $zR^c m(i)$. We can thus apply **R-Mon**(i) to conclude that $yP^c z$ and thus $y \trianglerighteq_i z$.
2. $z \rhd_i$ implies that $z \succ y \wedge \neg[yP^c z]$. Suppose bvy contradiction that $x \rhd_i y$, i.e., $xR^c y$. Then, similarly as in case 1, **R-Mon**(i) implies that $yP^c z$, a contradiction. Therefore, $\neg[xR^c y]$ and thus $y \trianglerighteq_i x$. □

Denote $\tilde{\trianglerighteq} = \bigcup_i \trianglerighteq_i$. Let $\trianglerighteq$ be the relation on X defined by:

$$\forall x, y \in X, x \trianglerighteq y \iff \begin{cases} x \tilde{\trianglerighteq} y & \text{if } x \in T(y) \\ x \succ y & \text{if } x \notin T(y) \end{cases}$$

Denote $\rhd$ the asymmetric part of $\trianglerighteq$.

Lemma 21. *The relation* $\trianglerighteq$ *is a complete preorder.*

Proof. We only need to prove the transitivity. Let $x, y, z \in X$ such that $x \trianglerighteq y \trianglerighteq z$. If there exists i such that $x, y, z \in T(i)$, then this follows from lemma 19. If $T(x) \cap T(y) = T(x) \cap T(z) = T(y) \cap T(z) = \emptyset$, then this follows from lemma 10.

Suppose we are in the case $T(x) = T(y) \neq T(z)$. Note that this implies that $y \succ z$. Suppose by contradiction that $z \succ x$. If $x \succ y$, this would imply that $yR^c x$, which contradicts $x \trianglerighteq y$, thus $y \succ x$. Given that $\neg[xR^c z]$ and $\neg[zR^c y]$, **R-NTran**(i) implies that $\neg[xR^c y]$. Similarly $\neg[xP^c y]$. Hence $y \rhd x$, a contradiction. Hence $x \succ z$ and thus $x \rhd z$. We deal with the case $T(y) = T(z) \neq T(x)$ similarly.

Suppose finally that we are in the case $T(x) = T(z) \neq T(y)$. Note that this implies that $x \succ y \succ z$. If $z \succ x$, then $xR^c z$ and thus $x \rhd z$. Suppose on the contrary that $x \succ z$. Then **R-NTran** implies that $\neg[zR^c x] \wedge \neg[zP^c x]$, that is $x \rhd z$. □

Now let $F = \bigcup_{1 \leq i \leq n^*} T^{\uparrow}(i) \cup T(0)$. Given that $\succ$ is transitive on F (lemma 10), there exists a function $w : F \longrightarrow \mathbb{R}$ that represents $\succ$ on F. Furthermore, for every $i = 0, \ldots, n^*$, $\succ$ is transitive on $T(i)$ (lemma 16), hence there exists a function $u_i : T(i) \longrightarrow \mathbb{R}$ representing $\succ$ on $T(i)$, and such that $u_T(x) = w(x)$ for every $x \in F$. We now define the function $u : X \longrightarrow \mathbb{R}$ such that for every i, $x \in T(i)$, $u(x) = u_i(x)$. We clearly have, for any menu A,

$$d(A) = \bigcup_{T \in \mathcal{T}} \arg\max_{x \in T \cap A} u(x).$$

Given lemma 21, there exists $v : X \longrightarrow \mathbb{R}$ that represents $\trianglerighteq$. Note that $\trianglerighteq \cap F^2 = \succ \cap F^2$, hence we can force that

$$v|_F = u|_F. \tag{C.5}$$

Lemma 22. *For any menu A,*

$$c(A) = \arg\max_{x \in d(A)} v(x).$$

Proof. For any $x, y \in d(A)$, $T(x) \neq T(y)$, so $\trianglerighteq \cap\, d(A)^2 = \succ \cap\, d(A)^2$. Therefore:

$$\arg\max_{x \in d(A)} v(x) = \max\left(d(A), \trianglerighteq\right) = \max\left(d(A), \succ\right) \underbrace{=}_{lemma\ 17} c(A).$$

□

Let $T \in \mathcal{T}$, so $T = T(i)$ for some $0 \leq i \leq n^\star$. Set $\lambda_T \equiv u(m(i))$.

To complete the proof of the theorem, we check that the tuple $\langle \mathcal{T}, \{\lambda_T\}, u, v \rangle$ so defined is an RS-structure that rationalizes the choice function c. Note first that by definition the order induced by u on each $T \in \mathcal{T}$ is $\succ \cap\, T^2$, which is a linear order (lemma 16). Lemma 22 shows that point i in definition 7 is satisfied. (C.5) show

that point ii is satisfied. Finally, lemma 20 shows that point iii is satisfied. □

C.3 Proof of Theorem 6

Proof. The necessity part of the theorem is left to the readers. We only prove the sufficiency.

(a) We first show that for any A,B such that $A \subseteq T$ and $B \subseteq T'$ for some $T,T' \in \mathcal{T}$, $A \succ B \iff A \cap F \neq \emptyset = B \cap F$. If $T = T'$, this is simply a consequence of part *(i)* of R-Dominance (RD).

Suppose now that $T \neq T'$. Let denote $A' = A \setminus F = \{a_1,\dots,a_n\}$ and $B' = B \setminus F = \{b_1,\dots,b_l\}$ and suppose that both are non-empty. By RD, $\{a_1\} \sim A'$, because both are richer than each other. Similarly $\{b_1\} \sim B'$. Furthermore, RD *(ii)* implies that $\{a_1\} \sim \{b_1\}$; hence, by transitivity, $A' \sim B'$.

Let denote $A'' = A \setminus A'$ and $B'' = B \setminus B'$. If $A'' = B'' = \emptyset$, we conclude from the previous argument that $A \sim B$. Suppose that $A'' \neq \emptyset = B''$, so $B = B'$. By a simple application of RD *(i)*, A is strictly richer than A', so $A \succ A'$, and by transitivity, $A \succ B$.

Assume now that $B'' \neq \emptyset$. By a similar reasoning as for A' and B', one can easily show that $A'' \sim B''$. If $B' = \emptyset$, then $B = B''$, hence $A \sim B$. If $B' \neq \emptyset$, note that $A' \cap A'' = B' \cap B'' = \emptyset$ and neither A' is richer than A'' nor B' is richer than B''. Hence applying twice R-Composition (RC), we obtain that $A \sim B$.

(b) We next show that for any A,B, if $\#\Phi(A) = \#\Phi(B)$, then $A \sim B$. Denote $\Phi(A) = \{A_1,\dots,A_n\}$ and $\Phi(B) = \{B_1,\dots,B_n\}$. By (a), we know that for any i, $A_i \sim B_i$. Noting that $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$, and neither A_1 is richer than A_2 nor B_1 is richer than B_2, by applying twice RC, we get that $A_1 \cup A_2 \sim B_1 \cup B_2$. By reiterating the same argument, we obtain that $\bigcup_i A_i \sim \bigcup_i B_i$. Finally, note that A is richer that $\bigcup_i A_i$ and conversely $\bigcup_i A_i$ is richer than A, hence, by RD, $A \sim \bigcup_i A_i$; similarly

$B \sim \bigcup_i B_i$. Therefore, by transitivity, we obtain that $A \sim B$.

(c) We finally prove that for any A, B, if $\#\Phi(A) > \#\Phi(B)$, then $A \succ B$. Denote $\Phi(A) = \{A_1, \ldots, A_n\}$ and $\Phi(B) = \{B_1, \ldots, B_k\}$, with $k < n$. By (b) $\bigcup_{i=1}^{k} A_i \sim B$. Furthermore, by RD, $\bigcup_{i=1}^{n} A_i \succ \bigcup_{i=1}^{k} A_i$. A similar argument as before shows that $A \sim \bigcup_{i=1}^{n} A_i$ and $B \sim \bigcup_{i=1}^{k} B_i$. Hence by transitivity $A \succ B$.

□

C.4 Proofs of Section 3.6

Proof of Proposition 11. *(i)* Denote $u(\sigma^L)$ and $v(\sigma^{RR})$ the DM's anticipated utility from choosing respectively σ^L and σ^{RR} in the menu N:

$$\begin{aligned} u(\sigma^L) \leq v(\sigma^{RR}) &\iff (1-p) + p\lambda - p(1-\lambda) \leq p + (1-p)\delta \\ &\iff p \geq \frac{1-\delta}{3-2\lambda-\delta} = \frac{1/2}{5/2-2\lambda} \end{aligned}$$

We define $p^\star := \frac{1/2}{5/2-2\lambda}$ and verify that $p^\star < 1/2$:

$$p^\star < 1/2 \iff \lambda < \frac{3}{4}$$

which is true by assumption.

(ii). We first compute the value $q^\star$ of the posterior such that for any $q \geq q^\star$, action r is preferred. $q^\star$ solves $(1-q) - q = q$, hence $q^\star = 1/3$.

Then we simply compare the posterior obtained after the realisation of a signal s^r from the news source σ^{RR} with 1/3. The posterior is, $\frac{p}{p+(1-p)1/2}$, which is greater or equal than p. We are in the case where $p \geq p^\star$, hence it is sufficient to show that $p^\star \geq 1/3$: $p^\star \geq \frac{1}{3} \iff \lambda \geq \delta$ which is true by assumption. □

Proof of Lemma 1. The maximand of the program (3.5) is strictly concave and the

set K_g is compact. Hence, C is well-defined (Weierstrass theorem) and is a choice function.

Now let us build the RS-structure $\langle \mathscr{T}, F, u, v\rangle$ that represents C. Given (3.6), $d^\star$ strictly increases with g if and only if $g > \hat{g}$. Let $g(t,d)$ be the g such that $t + gd^\beta = 1$.

Let us introduce the three following sets:

$$D^\uparrow = \bigcup_{\substack{t\in[0,1]\\ g>\hat{g}}} \{d \in [0,1] : (t,d) = C(K_g)\},$$

$$\forall d \in D^\uparrow, T(d) = \bigcup_{t\in[0,1]} \{(t,d)\},$$

$$T_0 = \bigcup_{\substack{d\notin D^\uparrow\\ t\in[0,1]}} \{(t,d)\}.$$

From these sets we can define the set of types and the freedom set

$$\mathscr{T} = \{T_0\} \cup \bigcup_{d\in D^\uparrow} \{T(d)\} \text{ and } F = T_0 \cup \bigcup_{d\in D^\uparrow} \{(t,d) \in T(d) : g(t,d) \leq \hat{g}\}$$

Now let us define u and v. For each (t,d) we posit $u(t,d) = t + P(d)\hat{V}$ and $v(t,d) = t + P(d)V\big(g(t,d)\big)$.

Given the uniqueness of $g(t,d)$ for each (t,d), v is well-defined. It can easily be shown that $\langle \mathscr{T}, F, u, v\rangle$ is an RS-structure. Consider the choice function C' which is the RSC defined on the compact subsets of $[0,1]^2$ and represented by $\langle \mathscr{T}, F, u, v\rangle$. We claim that for all g, $C(K_g) = C'(K_g)$. To check this claim let (t,d) and g such that $(t,d) = C(K_g)$ and (t',d') such that $(t',d') = C'(K_g)$. Note that $(t,d) = C(K_g)$ implies $g = g(t,d)$. Similarly, $(t',d') = C'(K_g)$ implies that $u(t',d') \geq u(t'',d')$ for all t'' such that $(t'',d') \in K_g$, that is, for all $t'' \leq t'$. Hence, $g = g(t',d')$.

Assume first that $g \leq \hat{g}$. Then, note that $(t',d') \in F$. Suppose that there exists $(t'',d'') \in K_g \setminus F$, this means that $g(t'',d'') > \hat{g}$, and hence there exists $t''' > t''$, such

that $g(t''',d'')=g$, which implies $(t''',d'')\in K_g$ and $u(t''',d'')>u(t'',d'')$. Therefore, only elements in F can be considered for choices in K_g according to the choice procedure (7). Hence, both (t,d) and (t',d') are elements of $\arg\max u(K_g)$. Because the latter is a singleton, $(t,d)=(t',d')$.

Assume next that $g(t,d)>\hat{g}$. Suppose that $(t',d')\in F$, then this implies that $d'\notin D^{\uparrow}$, that is $d'\neq d$. Because, $g(t,d)=g(t',d')$, this in turn implies that $t'\neq t$. Furthermore, by definition, $(t,d)\notin F$, and from $(t',d')=C'(K_g)$, we deduce that $u(t',d')>v(t,d)$. We also know that $v(t',d')\geq u(t',d')$, so $v(t',d')>v(t,d)$, which contradicts that $(t,d)=C(K_g)$. Therefore, $(t',d')\notin F$. Because $(t,d)\notin F$, $(t',d')=C'(K_g)$ and $(t,d)=C(K_g)$ imply that $v(t',d')\geq v(t,d)=\max v(K_g)$. Therefore, $(t',d')\in\arg\max v(K_g)$, and given that this set is a singleton, this implies that $(t',d')=(t,d)$. □

Proof of Proposition 12. This is a straightforward consequence of (3.7). □

Proof of Proposition 13. Action a^L can only be implemented in the absence of both a^R and a^{LL}. In any case, if a^L is chosen in a menu M by the agent, it is chosen in both states L and R, which gives the principal the expected payoff:

$$(1-p)\pi_L(a^L)+p\pi_R(a^L). \tag{C.6}$$

Similarly, action a^R can only be implemented in the absence of a^{RR}, in which case it is chosen in both states L and R, giving the principal the expected payoff:

$$(1-p)\pi_L(a^R)+p\pi_R(a^R). \tag{C.7}$$

From this we can deduce the existence of $p_\star\in(0,1)$ and $p^\star\in(0,1)$ such that: for any $p<p_\star$, the principal strictly prefers a menu M (e.g. $\{a^L\}$) such that $a_\theta(M)=a^L$ for $\theta=L,R$; for any $p>p^\star$, the principal strictly prefers a menu M (e.g. $\{a^R\}$) such

that $a_\theta(M) = a^R$ for $\theta = L, R$. Furthermore, there exists $\hat{p}$, the unique belief such that (C.6) = (C.7).

Only actions a^{LL} and a^{RR} can be simultaneously implemented respectively in state L and R. Given that $\pi_L(a^{LL}) > \pi_L(a^{RR})$ and $\pi_R(a^{RR}) > \pi_R(a^{LL})$, the principal will always prefer a menu implementing both actions (e.g. $\{a^{LL}, a^{RR}\}$) than a menu implementing only one of them. In this, the principal's expected payoff is:

$$(1-p)\pi_L(a^{LL}) + p\pi_R(a^{RR}). \tag{C.8}$$

Hence there exist $\underline{p}$ and $\bar{p}$ such that: (C.6) $\geq$ (C.8) if and only if $p \leq \underline{p}$; and (C.7) $\geq$ (C.8) if and only if $p \geq \bar{p}$.

The conclusions of the proposition follows easily from these observations. □

Appendix D

Proofs of Chapter 4

D.1 Proof of Section 4.4.1

Proof of Lemma 2. Let $\sigma \in \Sigma$ and suppose that there exist $\mu, \mu' \in \text{supp}(\sigma)$ with $p(\mu) = p(\mu')$. Consider the following market:

$$\tilde{\mu} = \frac{\sigma(\mu)}{\sigma(\mu) + \sigma(\mu')} x + \frac{\sigma(\mu')}{\sigma(\mu) + \sigma(\mu')} x'.$$

By the convexity of $X_{p(\mu)}$, $p(\tilde{\mu}) = p(\mu)$. Define σ' in the following way: $\sigma'(\tilde{\mu}) = \sigma(\mu) + \sigma(\mu')$, $\sigma'(\mu) = \sigma'(\mu') = 0$ and $\sigma' = \sigma$ otherwise. Is it easy to check that $\sum_{\mu \in \text{supp}(\sigma)} \sigma(\mu) W(\mu) = \sum_{\mu \in \text{supp}(\sigma')} \sigma'(\mu) W(\mu)$. We can iterate this operation as many times as the number of pairs $\nu, \nu' \in \text{supp}(\sigma')$ such that $p(\nu) = p(\nu')$ to finally obtain the desired conclusion. □

Proof of Lemma 3. Let $\mu^\star$ be an inefficient aggregate market, hence for any optimal segmentation $\sigma \in \Sigma(\mu^\star)$, $|\text{supp}(\sigma)| \geq 2$. Let σ be a direct and optimal segmentation of $\mu^\star$ and $\mu \in \text{supp}(\sigma)$ such that μ is in the interior of $X_{p(\mu)}$. Let ν be any other

market in the support of σ. Consider the market:

$$\xi = \frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\nu)}\mu + \frac{\sigma(\nu)}{\sigma(\mu)+\sigma(\nu)}\nu.$$

Because $\mu^\star$ is inefficient, it is without loss of generality to assume that ξ is also inefficient.

Denote $\bar{\mu}$ (resp. $\bar{\nu}$) the projection of ξ on the boundary of the simplex M in direction of μ (resp. ν). For σ to be optimal, the segmentation of ξ between μ with probability $\frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\nu)}$ and ν with probability $\frac{\sigma(\nu)}{\sigma(\mu)+\sigma(\nu)}$ must be optimal. In particular, it must be optimal among any segmentation on $[\bar{\mu},\bar{\nu}]$.

There exists a one-to-one mapping $f:[\bar{\mu},\bar{\nu}]\to[0,1]$ such that for any $\gamma\in[\bar{\mu},\bar{\nu}]$, $\gamma = f(\gamma)\bar{\mu}+(1-f(\gamma))\bar{\nu}$. Thus, the set $[\bar{\mu},\bar{\nu}]$ can be seen as all the distributions on a binary set of states of the world $\{\bar{\mu},\bar{\nu}\}$, where for any $\gamma\in[\bar{\mu},\bar{\nu}]$, $f(\gamma)$ is the probability of $\bar{\mu}$.

Therefore, the maximization program,

$$\max_{\sigma} \sum_{\gamma\in\text{supp}(\sigma)} \sigma(\gamma)W(\gamma) \tag{$\bar{S}$}$$
$$\text{s.t. } \sigma\in\Sigma^{[\bar{\mu},\bar{\nu}]}(\xi)\equiv\left\{\sigma\in\Delta([\bar{\mu},\bar{\nu}])\;\middle|\;\sum_{\gamma\in\text{supp}(\sigma)}\sigma(\gamma)\gamma=\xi,\ \text{supp}(\sigma)<\infty\right\},$$

is a bayesian persuasion problem (Kamenica and Gentzkow, 2011), with a binary state of the world and a finite number of actions. Hence, applying theorem 1 in Lipnowski and Mathevet (2017), there exists an optimal segmentation only supported on extreme points of sets $M\in\mathscr{M}^{[\bar{\mu},\bar{\nu}]}\equiv\{M_k\cap[\bar{\mu},\bar{\nu}]\mid k\in\{1,\dots,K\}\text{ and }M_k\cap[\bar{\mu},\bar{\nu}]\neq\emptyset\}$. It happens that for any $M\in\mathscr{M}^{[\bar{\mu},\bar{\nu}]}$, so that $M=M_k\cap[\bar{\mu},\bar{\nu}]$ for some k, if γ is an extreme point of M, then it is on the boundary of (M_k).

Let (μ',ν') with respective probabilities $(\alpha,1-\alpha)$ be a solution to ($\bar{S}$) where μ' and ν' are extreme points of some $M\in\mathscr{M}^{[\bar{\mu},\bar{\nu}]}$. We now consider the segmentation

$\bar{\sigma}$ such that $\bar{\sigma}(\gamma) = \sigma(\gamma)$ for all $\gamma \in \text{supp}(\sigma) \setminus \{\mu, \nu\}$, $\bar{\sigma}(\mu') = (\sigma(\mu) + \sigma(\nu))\alpha$, $\bar{\sigma}(\nu') = (\sigma(\mu) + \sigma(\nu))(1-\alpha)$, and $\bar{\sigma} = 0$ otherwise. One can easily check that $\bar{\sigma} \in \Sigma(\mu^\star)$. If $\bar{sigma}$ is not direct, that is, there exists $\gamma \in \text{supp}(\bar{\sigma})$ such that (w.l.o.g.) $p(\gamma) = p(\mu')$, then construct a direct segmentation $\bar{\bar{\sigma}}$ following the same process as in the proof of lemma 2. Then, if $\bar{\bar{\sigma}}$ is not only supported on boundaries of sets $\{M_k\}_{k \in I(\mu^\star)}$, reiterate the same process as above, until you reach the desired conclusion. □

D.2 Proofs of Section 4.4.2.

Proof of Proposition 15. Fix an aggregate market $\mu^\star$ and let $\sigma \in \Sigma(\mu^\star)$ be optimal and direct. Suppose by contradiction that there exist $\mu, \mu' \in \text{supp}(\sigma)$ such that $v_a := \min\{\text{supp}(\mu)\} < \max\{\text{supp}(\mu')\} =: v_d$ and $v_b := \min\{\text{supp}(\mu')\} < \max\{\text{supp}(\mu)\} =: v_c$. Assume further, without loss of generality, that $\min\{\text{supp}(\mu)\} < \min\{\text{supp}(\mu')\}$.

Define $\bar{\mu} := \frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\mu')}\mu + \frac{\sigma(\mu')}{\sigma(\mu)+\sigma(\mu')}\mu'$. A consequence of σ being optimal is that $V(\bar{\mu}) = \frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\mu')}W(\mu) + \frac{\sigma(\mu')}{\sigma(\mu)+\sigma(\mu')}W(\mu')$. The proof consists in showing that we can improve on this splitting of $\bar{\mu}$ and thus obtains a contradiction.

Define, for small $\epsilon > 0$, $\check{\mu}, \check{\mu}'$ as follows:

$$\check{\mu}_k = \begin{cases} \mu_k + \epsilon & \text{if } k = b \\ \mu_k - \epsilon & \text{if } k = c \\ \mu_k & \text{otherwise.} \end{cases}$$

$$\check{\mu}'_k = \begin{cases} \mu'_k - \frac{\sigma(\mu)}{\sigma(\mu')}\epsilon & \text{if } k = b \\ \mu'_k + \frac{\sigma(\mu)}{\sigma(\mu')}\epsilon & \text{if } k = c \\ \mu'_k & \text{otherwise.} \end{cases}$$

By construction, $\bar{\mu} = \frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\mu')}\check{\mu} + \frac{\sigma(\mu')}{\sigma(\mu)+\sigma(\mu')}\check{\mu}'$. Note that v_a is still an optimal

price for $\breve{\mu}$. Indeed, for any $v_a \le v_k \le v_b$, the profit made by fixing price v_k is equal in markets μ and $\breve{\mu}$ and for any $v_b < v_k \le v_c$ the profit made by fixing price v_k is strictly lower in $\breve{\mu}$ than in μ. On the contrary, $\phi(\breve{\mu}') \ge \phi(\mu')$ and it is possible that the inequality holds strictly. In any case, it must be that $\phi(\breve{\mu}') = v_e$ for $b \le e \le d$. Denote $\alpha := \frac{\sigma(\mu)}{\sigma(\mu)+\sigma(\mu')}$, hence $\frac{\sigma(\mu)}{\sigma(\mu')} = \frac{\alpha}{1-\alpha}$.

$$\alpha W(\breve{\mu}) + (1-\alpha)W(\breve{\mu}') - \big(\alpha W(\mu) + (1-\alpha)W(\mu')\big) \tag{D.1}$$

$$= \alpha\big(W(\breve{\mu}) - W(\mu)\big) + (1-\alpha)\big(W(\breve{\mu}') - W(\mu')\big) \tag{D.2}$$

$$= \alpha\epsilon\big(\lambda_b(v_b - v_a) - \lambda_c(v_c - v_a)\big) \tag{D.3}$$

$$+ (1-\alpha)\Big(-\sum_{k>e} \lambda_k \mu'_k (v_e - v_b) - \sum_{b<k\le e} \lambda_k \mu'_k (v_k - v_b) + \lambda_c \frac{\alpha}{1-\alpha}\epsilon(v_c - v_e)\Big) \tag{D.4}$$

$$= \alpha\epsilon\lambda_b(v_b - v_a) - \alpha\epsilon\lambda_c(v_e - v_a) - (1-\alpha)\Big(\sum_{k>e} \lambda_k \mu'_k (v_e - v_b) + \sum_{b<k\le e} \lambda_k \mu'_k (v_k - v_b)\Big) \tag{D.5}$$

$$> \alpha\epsilon\lambda_b(v_b - v_a) - \alpha\epsilon\lambda_{b+1}(v_e - v_a) - (1-\alpha)\Big(\sum_{k>e} \lambda_{b+1} \mu'_k (v_e - v_b) + \sum_{b<k\le e} \lambda_{b+1} \mu'_k (v_k - v_b)\Big) \tag{D.6}$$

$$= \alpha\epsilon\lambda_b(v_b - v_a) - \lambda_{b+1}\Big[\alpha\epsilon(v_e - v_a) - (1-\alpha)\Big(\sum_{k>e} \mu'_k (v_e - v_b) + \sum_{b<k\le e} \mu'_k (v_k - v_b)\Big)\Big] \tag{D.7}$$

Finally,

$$(D.7) \ge 0 \iff \frac{\lambda_b}{\lambda_{b+1}} \ge \kappa$$

where

$$\kappa = \frac{\alpha\epsilon(v_e - v_a) - (1-\alpha)\big(\sum_{k>e} \mu'_k (v_e - v_b) + \sum_{b<k\le e} \mu'_k (v_k - v_b)\big)}{\alpha\epsilon(v_b - v_a)}$$

which ends the proof. □

D.3 Proofs of Section 4.4.3.

Proof of Proposition 17. We know from proposition 16 that all markets belonging to NRR_u must be optimally segmented by splitting μ^* between $\mu^s = (\frac{\mu_1^*}{\sigma}, \frac{\mu_2^*}{\sigma}, \dots, \mu_u^s, 0, \dots, 0)$ and $\mu^r = (0, 0, \dots, \mu_u^r, \frac{\mu_{u+1}^*}{1-\sigma}, \dots, \frac{\mu_K^*}{1-\sigma})$. Such a segmentation indeed gives no rents to the monopolist if v_u is an optimal price in both μ^s and μ^r. That is, if:

$$v_1 = v_u \mu_u^s \geq v_j \left(\sum_{i=j}^{u-1} \frac{\mu_i^*}{\sigma} + \mu_u^s\right) \qquad \forall\, 2 \leq j \leq u-1 \tag{NR-s}$$

$$v_u \geq v_j \left(\sum_{i=j}^{K} \frac{\mu_i^*}{1-\sigma}\right) \qquad \forall\, u+1 \leq j \leq K \tag{NR-r}$$

As such, any SRP-optimal segmentation that maximizes consumer surplus must have $\mu_u^s = \frac{v_1}{v_u}$, $\sigma = \frac{v_u}{v_u - v_1} \sum_{i=1}^{u-1} \mu_i^*$ and $\mu_u^r = \frac{\mu_u^* v_u - \sum_{i=1}^{u} \mu_i^* v_1}{\sum_{i=u}^{K} \mu_i^* v_u - v_1}$.

We can rearrange both conditions and get:

$$0 \geq -\alpha(j) \sum_{i=1}^{j-1} \mu_i^* + (1 - \alpha(j)) \sum_{i=j}^{u-1} \mu_i^* \qquad \forall\, 2 \leq j \leq u-1 \tag{NR-s}$$

$$-\frac{v_1}{v_j(v_u - v_1)} \geq -\beta(j) \sum_{i=u}^{j-1} \mu_i^* + \big(1 - \beta(j)\big) \sum_{i=j}^{K} \mu_i^* \qquad \forall\, u+1 \leq j \leq K \tag{NR-r}$$

for $\alpha(j) = \frac{v_1(v_u - v_j)}{v_j(v_u - v_1)}$ and $\beta(j) = \frac{v_u^2}{v_j(v_u - v_1)}$.

Conditions (NR-s) and (NR-r) expressed above define $K-2$ half-spaces in $\mathbb{R}^K$. The non-rent region in M_u is thus given by the closed polytope defined by the intersection of such half-spaces. □

Proof of Proposition 16. As explained in the core of the text, the structure is a direct consequence of corollary 4. The value of σ simply follows from simple algebra and Bayes-plausibility. The value of μ_u^s follows from the fact that both v_1 and v_u must be optimal prices on segment μ^s. The value of μ_u^r can easily be

computed from σ, μ_u^s and Bayes-plausibility. □

Bibliography

ABDELGADIR, A. AND V. FOUKA (2020): "Political Secularism and Muslim Integration in the West: Assessing the Effects of the French Headscarf Ban," *American Political Science Review*, 114, 707–723. 78

ABRAMOWITZ, A. I. AND K. L. SAUNDERS (2008): "Is Polarization a Myth?" *The Journal of Politics*, 70, 542–555. 38

ADIWENA, B. Y., M. SATYAJATI, AND W. HAPSARI (2020): "Psychological Reactance and Beliefs in Conspiracy Theories During the Covid-19 Pandemic: Overview of the Extended Parallel Process Model (EPPM)," *Buletin Psikologi*, 28, 182. 75

AKBARPOUR, M., P. DWORCZAK, AND S. D. KOMINERS (2020): "Redistributive Allocation Mechanisms," *Working Paper*. 90

AKERLOF, G. A. AND R. E. KRANTON (2000): "Economics and Identity," *The Quarterly Journal of Economics*, 115, 715–753. 22

ALESINA, A. F. AND B. REICH (2015): "Nation building," *Working Paper, Department of Economics, Harvard University*. 78, 81

ALI, S. N., G. LEWIS, AND S. VASSERMAN (2022): "Voluntary Disclosure and Personalized Pricing," *The Review of Economic Studies*, rdac033. 90

ALONSO, R. AND N. MATOUSCHEK (2008): "Optimal Delegation," *The Review of Economic Studies*, 75, 259–293. 55, 82

APESTEGUIA, J. AND M. BALLESTER (2015): "A Measure of Rationality and Welfare," *Journal of Political Economy*, 123, 1278 – 1310. 53

APESTEGUIA, J. AND M. A. BALLESTER (2013): "Choice by sequential procedures," *Games and Economic Behavior*, 77, 90–99. 70

ARAD, A. AND A. RUBINSTEIN (2018): "The People's Perspective on Libertarian-Paternalistic Policies," *The Journal of Law and Economics*, 61, 311–333. 50

AUMANN, R. J. (1976): "Agreeing to Disagree," *The annals of statistics*, 1236–1239. 38

BARRERA, O., S. GURIEV, E. HENRY, AND E. ZHURAVSKAYA (2020): "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics," *Journal of Public Economics*, 182, 104123. 17, 29

BAUJARD, A. (2007): "Conceptions of freedom and ranking opportunity sets. A typology," *Homo Oeconomicus*, 2, 1–24. 54, 72

BECKER, G. S. AND C. B. MULLIGAN (1997): "The Endogenous Determination of Time Preference," *The Quarterly Journal of Economics*, 112, 729–758. 22

BERGEMANN, D., B. BROOKS, AND S. MORRIS (2015): "The Limits of Price Discrimination," *American Economic Review*, 105, 921–57. 14, 87, 183, 184

BERGEMANN, D. AND S. MORRIS (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 44–95. 14, 182

BERNHEIM, B. D., L. BRAGHIERI, A. MARTÍNEZ-MARQUINA, AND D. ZUCKERMAN (2021): "A Theory of Chosen Preferences," *American Economic Review*, 111, 720–54. 19, 22

BERNHEIM, B. D. AND A. RANGEL (2007): "Toward Choice-Theoretic Foundations for Behavioral Welfare Economics," *American Economic Review*, 97, 464–470. 53

——— (2008): "Choice-theoretic foundations for behavioral welfare economics," *The foundations of positive and normative economics: A handbook*, 155–192. 53

——— (2009): "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics," *The Quarterly Journal of Economics*, 124, 51–104. 53, 65

BERTRAND, M., D. CHUGH, AND S. MULLAINATHAN (2005): "Implicit Discrimination," *American Economic Review*, 95, 94–98. 25

BERTRAND, M. AND E. DUFLO (2017): "Field Experiments on Discrimination," *Handbook of Economic Field Experiments*, 1, 309–393. 25

BISIN, A. AND T. VERDIER (2001): "The economics of cultural transmission and the dynamics of preferences," *Journal of Economic theory*, 97, 298–319. 78, 79, 81

BOISSONNET, N. (2019): "Rationalizing Preference Formation by Partial Deliberation," *PhD Thesis*. 21

BOISSONNET, N. AND A. GHERSENGORIN (2022): "Note on the Indeterminacy of Deliberate Preference Change," *Working Paper*. 32

BOISSONNET, N., A. GHERSENGORIN, AND S. GLEYZE (2022a): "Revealed Deliberate Preference Change," *Working Paper*. 42

——— (2022b): "Supplement to "Revealed Deliberate Preference Change"," *Working Paper*. 20

BOXELL, L., M. GENTZKOW, AND J. M. SHAPIRO (2020): "Cross-Country Trends in Affective Polarization," *NBER Working Paper*. 38

BREHM, J. W. (1966): *A theory of psychological reactance.*, Academic Press. 13, 52, 60, 180

BREHM, S. S. AND J. W. BREHM (2013): *Psychological reactance: A theory of freedom and control*, Academic Press. 52, 58, 60

BROCK, T. C. (1968): "Implications of Commodity Theory for Value Change1," in *Psychological Foundations of Attitudes*, ed. by A. G. Greenwald, T. C. Brock, and T. M. Ostrom, Academic Press, 243–275. 13, 52, 59, 180

BUSHMAN, B. J. (2006): "Effects of Warning and Information Labels on Attraction to Television Violence in Viewers of Different Ages," *Journal of Applied Social Psychology*, 36, 2073–2078. 50

CARNAP, R. (1923): "Über die Aufgabe der Physik," *Kant-Studien*, 28, 90–107. 7, 173

CHAMBERS, C. P. AND T. HAYASHI (2012): "Choice and individual welfare," *Journal of Economic Theory*, 147, 1818–1849. 53

CHE, Y.-K. AND K. MIERENDORFF (2019): "Optimal Dynamic Allocation of Attention," *American Economic Review*, 109, 2993–3029. 75, 77

CHEREPANOV, V., T. FEDDERSEN, AND A. SANDRONI (2013a): "Rationalization," *Theoretical Economics*, 8, 775–800. 21

——— (2013b): "Rationalization," *Theoretical Economics*, 8, 775–800. 70

CHERNOFF, H. (1954): "Rational Selection of Decision Functions," *Econometrica*, 22, 422. 13, 51, 181

CLARKE, C. (2016): "Preferences and Positivist Methodology in Economics," *Philosophy of Science*, 83, 192–212. 7, 8, 173, 174

CRIPPS, M. W. (2018): "Divisible Updating," *Working Paper*. 38

DE CLIPPEL, G. AND K. ELIAZ (2012): "Reason-Based Choice: A Bargaining Rationale for the Attraction and Compromise Effects," *Theoretical Economics*, 7, 125–162. 21

DEKEL, E., B. L. LIPMAN, AND A. RUSTICHINI (2009): "Temptation-Driven Preferences," *The Review of Economic Studies*, 76, 937–971. 21

DIETRICH, F. AND C. LIST (2011): "A Model of Non-Informational Preference Change," *Journal of Theoretical Politics*, 23, 145–164. 21

——— (2013): "A Reason-Based Theory of Rational Choice," *Nous*, 47, 104–134. 11, 20, 27, 30, 178

——— (2016a): "Mentalism Versus Behaviourism in Economics: A Philosophy-of-Science Perspective," *Economics and Philosophy*, forthcoming. 8, 174

——— (2016b): "Reason-Based Choice and Context-Dependence: An Explanatory Framework," *Economics & Philosophy*, 32, 175–229. 20, 21

DUBE, J.-P. AND S. MISRA (2022): "Personalized Pricing and Consumer Welfare," *Journal of Political Economy*, Forthcoming. 90

DWORCZAK, P., S. D. KOMINERS, AND M. AKBARPOUR (2021): "Redistribution Through Markets," *Econometrica*, 89, 1665–1698. 88, 90, 93, 184

EHLERS, L. AND Y. SPRUMONT (2008): "Weakened WARP and top-cycle choice rules," *Journal of Mathematical Economics*, 44, 87–94. 70

ELLIOTT, M., A. GALEOTTI, A. KOH, AND W. LI (2021): "Market Segmentation through Information," *Working Paper*. 90

FOUKA, V. (2020): "Backlash: The unintended effects of language prohibition in US schools after World War I," *The Review of Economic Studies*, 87, 204–239. 78

GILBOA, I., A. POSTLEWAITE, L. SAMUELSON, AND D. SCHMEIDLER (2019): "What are axiomatizations good for?" *Theory and Decision*, 86. 7, 9, 173, 176

GLOVER, D., A. PALLAIS, AND W. PARIENTE (2017): "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores," *The Quarterly Journal of Economics*, 132, 1219–1260. 25

GOSSELT, J., M. DE JONG, AND J. VAN HOOF (2012): "Effects of Media Ratings on Children and Adolescents: A Litmus Test of the Forbidden Fruit Effect," *Journal of Communication*, 62. 50

GRUNE-YANOFF, T. AND S. HANSSON (2009): "Preference change: an introduction," in *Preference Change: Approaches from Philosophy, Economics and Psychology*, ed. by Till Grune-Yanoff and Sven Ove Hansson, 1–26. 9, 175

GRÜNE-YANOFF, T. (2022): "What preferences for behavioral welfare economics?" *Journal of Economic Methodology*, 29, 153–165. 53

GUALA, F. (2019): "Preferences: neither behavioural nor mental," *Economics and Philosophy*, 35, 383–401. 7, 8, 173, 174

GUL, F. AND W. PESENDORFER (2001): "Temptation and Self-Control," *Econometrica*, 69, 1403–1435. 21

——— (2005): "The Revealed Preference Theory of Changing Tastes," *The Review of Economic Studies*, 72, 429–448. 21

HAGHPANAH, N. AND R. SIEGEL (2022a): "The Limits of Multi-Product Price Discrimination," *American Economic Review: Insights*, Forthcoming. 90

——— (2022b): "Pareto Improving Segmentation of Multi-Product Markets," *Journal of Political Economy*, Forthcoming. 90, 104

HANKIN, J. R., I. J. FIRESTONE, J. J. SLOAN, J. AGER, A. C. GOODMAN, R. J. SOKOL, AND S. S. MARTIER (1993): "The Impact of the Alcohol Warning Label on Drinking during Pregnancy," *Journal of Public Policy & Marketing*, 12, 10 – 18. 50

HELLER, Y. (2012): "Justifiable Choice," *Games and Economic Behavior*, 76, 375–390. 21

HIDIR, S. AND N. VELLODI (2020): "Privacy, Personalization, and Price Discrimination," *Journal of the European Economic Association*, 19, 1342–1363. 90

HOLMSTROM, B. (1980): "On The Theory of Delegation," Discussion papers, Northwestern University, Center for Mathematical Studies in Economics and Management Science. 82

HORAN, S. (2016): "A simple model of two-stage choice," *Journal of Economic Theory*, 162, 372–406. 54, 70, 71

HOVLAND, C. I., I. L. JANIS, AND H. H. KELLEY (1953): *Communication and persuasion; psychological studies of opinion change*, Yale University Press New Haven. 75

IYENGAR, S., Y. LELKES, M. LEVENDUSKY, N. MALHOTRA, AND S. J. WESTWOOD (2019): "The Origins and Consequences of Affective Polarization in the United States," *Annual Review of Political Science*, 22, 129–146. 38

IYENGAR, S. AND S. J. WESTWOOD (2015): "Fear and Loathing Across Party Lines: New Evidence on Group Polarization," *American Journal of Political Science*, 59, 690–707. 38

JANSEN, E., S. MULKENS, Y. EMOND, AND A. JANSEN (2008): "From the Garden of Eden to the land of plenty: Restriction of fruit and sweets intake leads to increased fruit and sweets consumption in children," *Appetite*, 51, 570–575. 50

JANSEN, E., S. MULKENS, AND A. JANSEN (2007): "Do not eat the red food!: Prohibition of snacks leads to their relatively higher consumption in children," *Appetite*, 49, 572–577. 50

JONAS, E., V. GRAUPMANN, D. N. KAYSER, M. ZANNA, E. TRAUT-MATTAUSCH, AND D. FREY (2009): "Culture, self, and the emergence of reactance: Is there a "universal" freedom?" *Journal of Experimental Social Psychology*, 45, 1068–1080. 81

JONES, P. AND R. SUGDEN (1982): "Evaluating choice," *International Review of Law and Economics*, 2, 47–65. 71

KALAI, G., A. RUBINSTEIN, AND R. SPIEGLER (2002): "Rationalizing Choice Functions by Multiple Rationales," *Econometrica*, 70, 2481–2488. 21

KAMENICA, E. (2019): "Bayesian Persuasion and Information Design," *Annual Review of Economics*, 11, 249–272. 95

KAMENICA, E. AND M. GENTZKOW (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615. 154

KREPS, D. M. (1979): "A representation theorem for "preference for flexibility"," *Econometrica*, 565–577. 73

LANCASTER, K. J. (1966): "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132–157. 20

LEVESQUE, R. J. R. (2018): *Forbidden Fruit*, Springer International Publishing, 1453–1456. 12, 50, 59, 180

LIPNOWSKI, E. AND L. MATHEVET (2017): "Simplifying Bayesian Persuasion," *Working Paper*. 154

LLERAS, J. S., Y. MASATLIOGLU, D. NAKAJIMA, AND E. Y. OZBAY (2017): "When more is less: Limited consideration," *Journal of Economic Theory*, 170, 70–85. 70

LYNN, M. (1991): "Scarcity effects on value: A quantitative review of the commodity theory literature," *Psychology & Marketing*, 8, 43–57. 12, 52, 180

MANZINI, P. AND M. MARIOTTI (2007): "Sequentially rationalizable choice," *American Economic Review*, 97, 1824–1839. 54, 65, 70, 71

——— (2012): "Categorize then Choose: Boundedly Rational Choice and Welfare," *Journal of the European Economic Association*, 10, 1141–1165. 53, 70

MAS-COLELL, A., M. D. WHINSTON, AND J. R. GREEN (1995): *Microeconomic Theory*, New York: Oxford University Press. 95

MASATLIOGLU, Y., D. NAKAJIMA, AND E. Y. OZBAY (2012): "Revealed attention," *American Economic Review*, 102, 2183–2205. 53, 54, 65, 70, 71

MAZIS, M. B., R. B. SETTLE, AND D. C. LESLIE (1973): "Elimination of phosphate detergents and psychological reactance," *Journal of Marketing Research*, 10, 390–395. 50, 51

NEHRING, K. (2006): "Self-Control Through Second-Order Preferences," *Working Paper*. 21

NEHRING, K. AND C. PUPPE (2002): "A theory of diversity," *Econometrica*, 70, 1155–1198. 54, 72

NYHAN, B. AND J. REIFLER (2010): "When corrections fail: The persistence of political misperceptions," *Political Behavior*, 32, 303–330. 75

OK, E. A. (2007): *Real Analysis with Economic Applications*, Princeton University Press. 118, 123

OK, E. A., P. ORTOLEVA, AND G. RIELLA (2015): "Revealed (p) reference theory," *American Economic Review*, 105, 299–321. 54, 70

PALACIOS-HUERTA, I. AND T. J. SANTOS (2004): "A Theory of Markets, Institutions, and Endogenous Preferences," *Journal of Public Economics*, 88, 601–627. 22

PATTANAIK, P. K. AND Y. XU (1990): "On Ranking Opportunity Sets in Terms of Freedom of Choice," *Recherches Économiques de Louvain / Louvain Economic Review*, 56, 383–390. 54, 71, 73

——— (1998): "On preference and freedom," *Theory and Decision*, 44, 173–198. 54, 72

——— (2000): "On diversity and freedom of choice," *Mathematical Social Sciences*, 40, 123–130. 54, 72

PECHMANN, C. AND C.-F. SHIH (1999): "Smoking Scenes in Movies and Antismoking Advertisements before Movies: Effects on Youth," *Journal of Marketing*, 63, 1–13. 50

PIGOU, A. C. (1920): *The Economics of Welfare*, London: Macmillan. 89

POPPER, K. R. (1934): *The Logic of Scientific Discovery*, London: Hutchinson. 7, 173

RAVID, D., A.-K. ROESLER, AND B. SZENTES (2022): "Learning before Trading: On the Inefficiency of Ignoring Free Information," *Journal of Political Economy*, 130, 346–387. 90

RIDOUT, S. (2021): "Choosing for the Right Reasons," *Unpublished manuscript*. 21, 70

ROBINSON, J. (1933): *The Economics of Imperfect Competition*, London: Macmillan. 89

ROESLER, A.-K. AND B. SZENTES (2017): "Buyer-Optimal Learning and Monopoly Pricing," *American Economic Review*, 107, 2072–80. 90

ROSENBERG, B. AND J. SIEGEL (2018): "A 50-Year Review of Psychological Reactance Theory: Do Not Read This Article." *Motivation Science*, 4, 281–300. 12, 52, 59, 180

RUBINSTEIN, A. AND Y. SALANT (2012): "Eliciting Welfare Preferences from Behavioural Data Sets," *The Review of Economic Studies*, 79, 375–387. 53

SAMUELSON, P. A. (1938): "A Note on the Pure Theory of Consumer's Behaviour," *Economica*, 5, 61–71. 7, 51, 174

SCHMALENSEE, R. (1981): "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, 71, 242–247. 89

SEN, A. (1971): "Choice Functions and Revealed Preference," *Review of Economic Studies*, 38. 13, 51, 60, 65, 181

——— (1973): "Behaviour and the Concept of Preference," *Economica*, 40, 241–259. 31

——— (1993): "Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms," *Oxford Economic Papers*, 45, 519–541. 72

SENSENIG, J. AND J. W. BREHM (1968): "Attitude change from an implied threat to attitudinal freedom." *Journal of Personality and Social Psychology*, 8, 324. 75

SHAFIR, E., I. SIMONSON, AND A. TVERSKY (1993): "Reason-Based Choice," *Cognition*, 49, 11–36. 20

SIMONSON, I. (1989): "Choice Based on Reasons: The Case of Attraction and Compromise Effects," *Journal of Consumer Research*, 16, 158–174. 20

SNEEGAS, J. E. AND T. A. PLANK (1998): "Gender differences in pre-adolescent reactance to age-categorized television advisory labels," *Journal of Broadcasting & Electronic Media*, 42, 423–434. 50

STROTZ, R. H. (1955): "Myopia and Inconsistency in Dynamic Utility Maximization," *The Review of Economic Studies*, 23, 165–180. 21

THOMA, J. (2021): "In defence of revealed preference theory," *Economics and Philosophy*, 37, 163–187. 7, 8, 174

TOPKIS, D. M. (1998): *Supermodularity and Complementarity*, Princeton University Press. 106

TVERSKY, A. AND I. SIMONSON (1993): "Context-Dependent Preferences," *Management Science*, 39, 1179–1189. 20

VARAVA, K. A. AND B. L. QUICK (2015): "Adolescents and Movie Ratings: Is Psychological Reactance a Theoretical Explanation for the Forbidden Fruit Effect?" *Journal of Broadcasting & Electronic Media*, 59, 149–168. 50

VARIAN, H. R. (1985): "Price Discrimination and Social Welfare," *American Economic Review*, 75, 870–875. 89

WOOD, T. AND E. PORTER (2019): "The elusive backfire effect: Mass attitudes' steadfast factual adherence," *Political Behavior*, 41, 135–163. 75

Résumé en français

Un titre si vague et général laisse suggérer que chercher un fil conducteur aux thèmes qui sont abordés dans cette thèse semble, sinon vain, pour le moins quelque peu artificiel. Les chapitres suivent cependant l'ordre de mes questionnements et l'évolution de mes interêts concernant la science économique. C'est donc par ce biais que je tenterai d'unifier un tant soit peu la réflexion qui sous-tend cette thèse. Dans ce but, un bref historique de l'étude du choix individuel en économie est retracé. L'accent est particulièrement mis sur les interrogations qui ont mené à la formulation de la théorie des préférences révélées (TPR). Comme nous le verrons, les deux premiers projets de cette thèse (regroupés dans les chapitres 1, 2 et 3) sont directement liés aux objectifs poursuivies par la TPR, à savoir: (1) trouver une contrepartie observable à chaque concept théorique utilisé dans les modèles économiques de choix individuel et (2) rendre ces modèles réfutables. Les chapitres 1 et 2 ont été initiés par l'interrogation épistémologique suivante : un modèle raisonnable de changement de préférence peut-il être réfutable ? Le chapitre 3 se situe également dans le domaine de la TPR mais aborde une question plus pratique. Il est né d'un souci d'intégrer des considérations sur la liberté dans la modélisation économique du choix individuel. Mes doutes liminaires quant aux fondements de la théorie microéconomique se trouvant progressivement dissipés, mon intérêt pour des questions économiques plus concrètes s'est accru, ce qui m'a conduit à mon troisième projet (le chapitre 4). Ce dernier est quelque peu à part dans la

thèse puisqu'il est le seul qui ne relève pas du domaine de la théorie de la décision mais d'un champ récemment foisonnant de la littérature théorique en économie : le design de l'information. Il étudie l'effet (re)distributif de la discrimination par les prix fondée sur les caractéristiques individuelles.

De la maximisation de l'utilité à la théorie des préférences révélées. L'économie suscitait en moi des sentiments mitigés ; en effet, une science sociale qui utilise les mathématiques comme principal outil et langage pour étudier les décisions et les interactions humaines était à la fois source de fascination et de doute. Le début de mon doctorat a donc été consacré à mieux comprendre les fondements de l'analyse économique moderne. Au cœur de ces fondements se trouve la modélisation de la prise de décision individuelle comme la maximisation d'une fonction d'utilité sur un ensemble d'alternatives possibles (par exemple, afin de choisir un panier de consommation étant donnée une contrainte budgétaire). Si l'information est incomplète, cette utilité est associée à une probabilité telle que l'individu cherche désormais à maximiser l'espérance des utilités obtenus dans chaque scénario possible. L'étude des décisions individuelles est devenue au fil des ans un sous-domaine indépendant de l'économie, communément appelé théorie de la décision (ou théorie du choix).

Sous l'influence du positivisme philosophique (Carnap, 1923) et de la notion poppérienne de réfutabilité (Popper, 1934), les économistes ont voulu : (1) relier leurs concepts théoriques, tels que l'utilité et la probabilité, à des comportements observables ; et (2) rendre leurs théories réfutables.[1] À cette fin, la théorie du choix a d'abord remplacé l'utilité (ou le plaisir) par la notion de "préférences". Bien qu'une préférence soit plus tangible que l'utilité (puisqu'elle ne nécessite que de classer les options entre elles et non de donner une mesure quantifiée du plaisir

[1] Pour l'héritage positiviste en économie, voir Clarke (2016); Guala (2019); Gilboa et al. (2019).

ou de la satisfaction), elle n'en reste pas moins un concept théorique qui n'est pas directement observables. La théorie des préférences révélées, inaugurée par Samuelson (1938), a finalisé le programme positiviste de la théorie du choix en donnant "une définition entièrement comportementale des préférences, par laquelle ces dernières sont assimilés à un comportement de choix réel ou hypothétique" (Thoma, 2021).[2] Ainsi, les théoriciens des préférences révélées énoncent des conditions de cohérence sur les comportements de choix (les *axiomes*) qui sont nécessaires et suffisantes pour représenter le choix d'un agent *comme si* il maximisait une préférence avec certaines caractéristiques spécifiques (par exemple, la transitivité). Par conséquent, la TPR fait d'une pierre deux coups : (1) elle dérive directement les préférences des choix observables et (2) elle fournit un cadre pour énoncer formellement les conditions de réfutabilité des modèles de décisions individuelles.

Le premier de ces deux objectifs remplis par la TPR — à savoir, le fondement comportemental des préférences — est souvent appelé "comportementalisme" et a fait l'objet de vives critiques dans la littérature philosophique, donnant naissance à la position antagoniste : le mentalisme.[3]. Ce débat s'articule autour de la question de savoir si les économistes doivent faire explicitement appel aux états mentaux conatifs et cognitifs des individus pour rationaliser leur comportement, ou fonder leur analyse uniquement sur des objets observables. Sans entrer dans les détails de cette controverse, disons simplement que, conformément à certains développements récents (Guala, 2019; Thoma, 2021), nous ne partageons pas le point de vue selon lequel embrasser la TPR implique d'adopter la posture comportementaliste radicale communément admise.[4] À l'appui de cette affirmation, il semble que les

[2]Traduction par l'auteur.

[3]Pour des références et un résumé du débat entre mentalisme et comportementalisme en économie, voir, entre autres, Clarke (2016); Dietrich and List (2016a); Guala (2019); Thoma (2021).

[4]Nous soulignons également que cela n'implique pas nécessairement que nous acceptions pleinement la position mentaliste. Je considère plutôt mon travail comme se situant quelque part entre les deux.

travaux récents de la TPR donnent fréquemment des interprétations et explications psychologiques possibles de leurs modèles comportementaux de choix.

Sous l'impulsion de l'économie expérimentale et comportementale, la TPR a récemment connu une vive expansion. Étant données les preuves croissantes de transgression du modèle canonique de choix (c'est-à-dire, la maximisation d'un ordre complet et transitif sur les options), de nouvelles procédures de choix sophistiquées tenant compte de ces violations ont reçu des fondations axiomatiques. Bien que ce ne soit pas toujours le cas, ces modèles sont souvent appuyés par des références à des explications psychologiques. Les trois premiers chapitres de cette thèse s'inscrivent dans ce courant de la littérature. Par conséquent, à mon sens, le principal apport de l'adoption du cadre de la TPR dans ces travaux concerne le deuxième objectif rempli par la TPR : la réfutabilité des modèles de choix individuels. Cette préoccupation épistémologique est ce qui a initialement motivé le projet des chapitres 1 et 2. En effet, les modèles économiques intégrant les changements de préférences ont généralement été critiqués pour leur manque de contenu empirique : en somme, tout phénomène social pourrait être expliqué en supposant des changements de préférences *ad hoc*.[5] Ces chapitres visent à combler cette lacune en proposant un modèle réfutable de changement de préférences.

Nous ajoutons que la méthode des préférences révélées sert un autre objectif, comme on peut le noter dans nombre des travaux récents : elle permet de déterminer si différentes explications ou théories psychologiques ont des conséquences comportementales distinctes. Du moins, elle offre un langage permettant de discuter des implications comportementales de différents concepts théoriques hérités de la psychologie ou de la philosophie. Le chapitre 3 en donne une illustration : nous proposons un modèle qui rationalise les réactions comportementales possibles des individus face aux restrictions de leur ensemble d'opportunités et soutenons que ce

[5] Voir Grune-Yanoff and Hansson (2009) pour une revue des raisons avancées contre l'intégration des changements de préférences dans la science économique.

modèle est compatible avec les principales explications données en psychologie.

Avant de passer au résumé détaillée de chaque chapitre, nous souhaitons souligner un autre objectif commun de l'approche axiomatique en théorie du choix. Bien qu'elle ne figurait pas explicitement parmi les objectifs initiaux de la TPR, l'approche axiomatique (qui, au-delà de la TPR, est la principale méthode adoptée en théorie de la décision) est souvent utilisée en vue de faire une étude normative d'un phénomène. Ceci explique d'ailleurs pourquoi le modèle canonique de la TPR est parfois considéré comme la formulation classique de la théorie du choix rationnel. À cet égard, les axiomes "peuvent aider l'économiste à convaincre les personnes auxquelles il ou elle s'adresse qu'il faudrait effectivement suivre sa recommandation, ou peuvent attirer l'attention sur les faiblesses d'un modèle" (Gilboa et al., 2019).[6] En accord avec cet objectif, notre modèle de changement de préférence dans les chapitres 1 et 2 est charactérisé par deux axiomes qui non seulement sont réfutables mais sont également la traduction de deux principes normatifs.

La thèse s'organise autour de trois projets qui sont détaillés ci-dessous. Dans les chapitres 1 et 2, nous proposons un modèle de changement de préférence qui est réfutable et normativement fondé. Le chapitre 3 est consacré à l'étude des réactions comportementales aux restrictions de l'ensemble d'opportunités: nous présentons un modèle de choix individuel qui traduit les principales théories psychologiques. Finalement, le chapitre 4 étudie la conceptualisation d'une politique de discrimination par les prix qui maximiserait le surplus des consommateurs avec un objectif de redistribution.

Réfutabilité et changement de préférence (chapitres 1 et 2). Comme nous l'avons souligné précédemment, l'absence de fondements réfutables constitue l'une

[6] Traduction par l'auteur.

des principales raisons pour lesquelles les économistes ont longtemps été réticents à invoquer les changements de préférence dans leur analyse. Historiquement, ils se sont contentés d'expliquer les changements de comportement, soit par des évolutions des contraintes auxquelles font face les individus (par exemple, leurs contraintes budgétaires), soit par l'arrivée de nouvelles informations, induisant les agent à mettre à jour leurs connaissances (ou croyances) sur leur environnement et à adapter leur comportement en conséquence. Néanmoins, un large éventail de phénomènes semble mieux s'expliquer par des changements de préférences, notamment lorsqu'ils impliquent des valeurs telles que l'équité, le progressisme, etc. Par exemple, l'expansion du droit à l'avortement dans les sociétés occidentales (ou sa remise en question actuelle dans certains pays) est plus vraisemblablement due à la diffusion (ou à la réévaluation) de valeurs telles que les droits des femmes qu'à un changement de croyances sur un état du monde sous-jacent.

La difficulté d'obtenir un modèle réfutable de changement des préférences est liée à l'absence de fondements normatifs apparents (par exemple, par rapport à la formule de Bayes en probabilité). Le défi est donc double : trouver un modèle de changement des préférences qui soit à la fois testable empiriquement et convaincant sur le plan normatif. Le chapitre 1, rédigé conjointement avec Niels Boissonnet et Simon Gleyze, est une humble tentative de relever ce défi. Pour ce faire, nous restreignons l'ensemble des changements de préférences que nous souhaitons décrire. Nous proposons ainsi un modèle dans lequel le changement de préférence est délibéré. Ce modèle est identifiable à partir de l'observation de choix successifs d'un individu et caractérisé par deux principes normatifs réfutables : le *principe des raisons suffisantes* et le *principe de délibération*.

Notre cadre est le suivant : il existe un ensemble fini d'alternatives, chacune d'entre elles étant définie par des attributs observables, et un analyste extérieur observe un agent (il) faire des choix parmi ces alternatives durant plusieurs périodes

consécutives. Les attributs sont pensés comme la traduction matérielle de la notion de *raisons* utilisée en philosophie. Selon cette littérature, les raisons seraient les motifs fondamentaux de l'action.[7] Dans ce contexte, la question que nous posons est la suivante: sous quelles conditions sur les choix de l'agent pouvons-nous rationaliser les modifications successives de ses choix comme étant induit par des changements délibérés de préférences?

En observant les choix et les attributs des alternatives, nous pouvons définir lesquels d'entre eux sont *pertinents* pour le choix de l'agent à chaque période. Intuitivement, si l'agent exprime une préférence stricte entre deux options qui ne diffèrent que sur un attribut, cela signifie que cet attribut doit être pris en compte par l'agent dans ce choix. Nous pouvons donc identifier, à chaque période t, l'ensemble M_t des *attributs pertinents* de l'agent.

Notre premier axiome, le principe des raisons suffisantes, stipule que l'agent doit classer de manière similaire, à l'intérieur de chaque période ainsi qu'entre elles, les options qui ont les mêmes attributs pertinents. De là, nous obtenons que l'agent change ses préférences si et seulement si cela peut être justifié par (au moins) un attribut qui est rendu (non) pertinent. Par exemple, si un employeur se rend compte que sa décision d'embauche est fondée sur l'attribut "genre", il pourrait rendre cet attribut non pertinent à l'avenir pour cesser d'être discriminant.[8] En d'autres termes, chaque ensemble d'attributs pertinents M_t détermine une unique préférence; par conséquent, un changement de préférence correspond à un changement de l'ensemble d'attributs pertinents.

Le principe de délibération stipule que l'agent ne doit pas commettre d'erreur (de son point de vue) lorsqu'il modifie ses préférences. Autrement dit, si l'agent modifie son ensemble d'attributs pertinents en t, il ne peut plus choisir dans les

[7] Voir Dietrich and List (2013) pour une discussion plus détaillée sur la fondation du concept de préférences sur celui de raisons et une revue de la littérature philosophique associée.

[8] La discrimination implicite impliquerait que l'attribut "sexe" soit *pertinent*. Par conséquent, un attribut peut être pertinent même si l'agent n'utilise pas consciemment cet attribut.

périodes futures aucun ensemble d'attributs pertinents qu'il aurait pu prendre en t. Cela se traduit par un axiom qui stipule que l'agent ne peut pas changer d'avis deux fois concernant un attribut si aucun événement supplémentaire ne s'est produit entre-temps. Sinon, cela indiquerait qu'il aurait échoué dans sa délibération.

Nous montrons que ces deux axiomes sont équivalents à une procédure que nous appelons *changement de préférence délibéré* : (i) les préférences à chaque période sont rationalisées par les attributs pertinents ainsi qu'un ordre de préférence stable dans le temps ; (ii) les changements de préférences sont induits par la *prise de conscience* de nouveaux attributs et une délibération sur l'ensemble des attributs pertinents à adopter pour les périodes futures — ceci est rationalisé par la maximisation d'un ordre sur les préférences elles-mêmes, la *méta-préférence*. Formellement, le point (ii) signifie qu'il existe à chaque période t un ensemble A_t d'attributs (c'est-à-dire de raisons, valeurs, etc.) dont l'individu prend conscience; ce sont les attributs qu'il peut modifier pour la prochaine période. Conjointement avec l'ensemble actuel d'attributs pertinents M_t, cela détermine la collection des ensembles atteignables d'attributs pertinents, notée $R(M_t, A_t)$. L'idée de changement délibéré est alors représentée par le fait que l'agent choisit, parmi cette collection $R(M_t, A_t)$, l'ensemble M_{t+1} qui maximise sa méta-préférence.

Notre interprétation est la suivante : chaque fois que l'agent devient *conscient* d'un attribut — par l'éducation, les interactions sociales, les médias ou l'introspection — il peut décider de le rendre pertinent ou non pour la période suivante, induisant un changement de préférence. Ces changements sont effectués délibérément et sont donc cohérents dans le temps ; ils peuvent résulter des valeurs morales de l'agent, d'un raisonnement motivé, d'objectifs sociaux, de normes, etc.

Le chapitre 2, écrit conjointement avec Niels Boissonnet, s'attaque simplement à un problème d'indétermination qui est laissé de côté dans le premier chapitre. Il caractérise une version plus générale du changement de préférence délibéré où

d'autres suites possibles d'attributs pertinents sont autorisées.

Choix sensibles aux restrictions (Chapitre 3). Lorsque l'accès à certaines opportunités nous est refusé, il a été observé dans différents contextes que ces alternatives interdites (ou leurs substituts) jouïssent d'un surcroît d'attrait. Ce phénomène est communément appelé l'*effet du fruit défendu*, une allusion au célèbre épisode biblique de la Genèse (Levesque, 2018). Le chapitre 3, écrit en collaboration avec Niels Boissonnet, propose un modèle de choix qui rationalise l'effet du fruit défendu. Ce modèle est parfaitement caractérisé et identifiable par l'observation des comportements de choix d'un individu.

Les deux principales explications de l'effet du fruit défendu en psychologie sont la théorie de la réactance et la théorie du produit rare, qui sont toutes deux cohérentes avec notre modèle.[9] La réactance établit un lien entre la réaction des gens aux restrictions ou aux interdictions et leur attitude à l'égard de la liberté de comportement. Lorsqu'une restriction est perçue comme une menace pour sa liberté de comportement, la personne éprouve un état émotionnel déplaisant qui la pousse à *restaurer* cette liberté menacée, ce que les psychologues ont appelé la réactance (Brehm, 1966). L'explication de l'effet du fruit défendu par la théorie du produit rare a un caractère plus "hédoniste" (Brock, 1968). Selon cette théorie, plus une marchandise est perçue comme indisponible ou nécessitant beaucoup d'efforts pour être obtenue, plus elle sera valorisée, augmentant ainsi le désir des agents pour cette alternative.

Nous utilisons l'environnement typique de la TPR, à savoir: nous observons les choix effectués par un agent (elle) dans chaque menu possible formé à partir d'un ensemble fini d'options. Dans ce cadre, l'effet du fruit défendu se manifeste par des

[9] Voir Rosenberg and Siegel (2018) pour une revue de la littérature sur la théorie de la réactance psychologique ; Lynn (1991) pour la théorie du produit rare (traduction par l'auteur de "commodity theory").

renversements de choix provoqués par le retrait d'une alternative apparemment non pertinente : par exemple, z est choisi dans $\{x,y,z\}$, mais une fois y supprimé, x est choisi dans $\{x,z\}$. La privation de y oriente le désir de l'agent vers un substitut potentiel (en l'occurrence, x). Formellement, analyser les choix sensibles aux restrictions revient à étudier les violations de l'axiome d'"Indépendance des alternatives non pertinentes" (Chernoff, 1954; Sen, 1971, propriété α) causées par la *suppression* d'alternatives.

Nous étudions une classe de procédures de choix, que nous appelons *restriction-sensitive choice* (RSC), qui rend compte de l'effet du fruit défendu. RSC peut être considéré comme un processus en quatre étapes. Premièrement, l'agent catégorise l'ensemble des options en *types* (par exemple, la différenciation horizontale). Deuxièmement, les options au sein des types sont classées en fonction d'une *fonction d'utilité* u, qui représente la satisfaction intrinsèque de l'agent, ou son bien-être matériel (par exemple, la différenciation verticale). Troisièmement, au sein de chaque type, elle détermine un *seuil* d'utilité, en dessous duquel les options sont évaluées par une *fonction de réaction* v (qui diffère de u). Quatrièmement, le choix s'effectue en choisissant parmi les meilleurs éléments disponibles de chaque type (selon u), où l'élément supérieur est évalué par v ou u selon qu'il se situe au-dessous ou au-dessus du seuil. Pour illustrer le modèle, considérons trois options x,y,z qui sont différenciées horizontalement et verticalement ; à savoir, x est le même produit que y, mais à un prix plus élevé ; z est un autre produit. L'agent a une préférence intrinsèque pour z, représentée par la fonction d'utilité u : $u(z) > u(y) > u(x)$. Par conséquent, z est choisi dans l'ensemble $\{x,y,z\}$. Cependant, lorsque l'accès aux options du premier type est limité à la mauvaise option (c'est-à-dire x), le désir de l'agent de choisir une option de ce type est augmenté (c'est-à-dire choisir x plutôt que z), ce qui génère l'effet du fruit défendu. Ceci est formellement obtenu par le seuil du premier type, qui se situe entre $u(y)$ et $u(x)$, et la *fonction de réaction* v,

qui est telle que $v(x) > u(z)$. Nous interprétons ainsi v comme la combinaison du bien-être et du désir supplémentaire provoqué par les restrictions.

En plus de la caractérisation axiomatique de notre modèle de choix, nous montrons comment ses ingrédients peuvent être identifiés à partir des réactions observées aux restrictions. Nous explorons ensuite les conclusions en termes de bien-être pour l'agent qui peuvent être tirées de notre modèle et nous démontrons la différence entre le critère de bien-être qui en découle et les principaux critères proposés dans la littérature. Nous dérivons également une mesure de la liberté de choix offerte par les différents ensembles possibles d'opportunités auxquels peut être confronté un agent, étant donné que ses choix finaux sont déterminés par notre modèle.

Enfin, nous étudions trois applications de notre modèle. Nous montrons tout d'abord comment il peut fournir une explication à l'émergence de théories conspirationnistes et à l'effet contre-productif des politiques d'intégration ciblant les minorités — deux phénomènes qui ont souvent été associés à la réactance. Nous étudions ensuite le problème d'un principal qui délègue ses décisions à un agent mieux informé mais ayant des intérêts divergents, et dont les choix suivent notre modèle. Nous montrons que l'effet sur le bien-être de l'agent est ambigu.

Les effets (re)distributifs de la discrimination par les prix (Chapitre 4). Le dernier chapitre, fruit d'un projet conjoint avec Daniel Barreto et Victor Augias, se distingue dans cette thèse car c'est le seul qui ne se situe pas dans le domaine de la théorie de la décision. Il s'inscrit dans la littérature récente sur le design de l'information (voir Bergemann and Morris, 2019). Notre objectif est de fournir une analyse normative des effets distributifs de la tarification personnalisée, une pratique qui suscite un intérêt croissant de la part des régulateurs étant donnée la quantité croissante de données de consommation collectées sur internet. En

effet, les consommateurs laissent continuellement des traces de leur identité, que ce soit par leur activité sur les réseaux sociaux, leur utilisation des moteurs de recherche, leurs achats en ligne, etc. Ces données de consommation sont très prisées par les acteurs de l'économie numérique.[10] Une pratique présentant un intérêt réglementaire particulier est celle qui consiste à facturer des prix personnalisés à différents consommateurs en fonction de l'estimation de leur propension à payer pour des produits — ce que l'on nomme parfois la discrimination par les prix de troisème type.

Comme l'a montré Bergemann et al. (2015), la tarification personnalisée peut non seulement être utilisée pour augmenter le surplus économique — en appliquant des prix qui permettent à chaque consommateur d'acheter —, mais elle peut également être réalisée de manière à garantir que tout le surplus créé revienne aux consommateurs. Cependant, un aspect important n'a pas été étudié par la littérature : puisque différents consommateurs paient des prix différents, la politique de discrimination par les prix définit la manière dont le surplus est distribué *entre* les consommateurs. Ceci soulève des questions sur la manière dont une telle politique peut bénéficier aux consommateurs les plus pauvres par rapport aux plus riches.

Notre objectif est d'étudier l'impact de la tarification personnalisée sur les différents consommateurs et la manière dont elle devrait être réalisée dans le but d'augmenter le surplus des consommateurs tout en donnant la priorité aux plus pauvres d'entre eux. Formellement, un producteur en situation de monopole vend un bien sur un marché composé de consommateurs hétérogènes, chacun d'entre eux peut consommer au plus une unité et est caractérisé par sa propension à payer pour ce bien. Un planificateur social peut transmettre de l'information au producteur

[10] Ceci peut être illustré par l'ascension rapide de la licorne française de l'analyse numérique Contentsquare, qui a levé plus de 1,1 milliard de dollars de fonds d'investissement entre mai 2021 et juin 2022 et dont les services permettent aux entreprises d'adapter des décisions, telles que la tarification et la publicité, à différents consommateurs.

sur la propension à payer des consommateurs. Une stratégie de transmission d'information détermine une *segmentation*, c'est-à-dire, une division du marché aggrégé des consommateurs en différents sous-marchés, nommés des *segments*, sur chacun desquels le producteur peut fixer un prix différent. L'objectif du planificateur est de maximiser la somme pondérée des surplus des consommateurs où les poids sont décroissants avec la richesse. La richesse relative des consommateurs est identifiée à partir de leur propension à payer, sous l'hypothèse simple que les individus qui sont prêts à payer des prix plus élevés sont en moyenne plus riches (Dworczak et al., 2021).

Nous établissons d'abord que, tant que les poids sont positifs, la maximisation de l'objectif de redistribution du planificateur ne se fait jamais au détriment de l'efficience, c'est-à-dire qu'elle ne sacrifie jamais de surplus social total réalisable. Par conséquent, l'objectif de redistribution n'implique jamais de perte sèche. Nous montrons ensuite que, pour certains marchés agrégés de consommateurs, un objectif de redistribution suffisamment fort ne peut pas être atteint tout en maximisant le surplus total des consommateurs. Dans ce cas, le surplus des consommateurs les plus pauvres est augmenté au sacrifice des plus riches, ce qui permet au producteur de dégager des profits supplémentaires. Nous caractérisons l'ensemble des marchés pour lesquels c'est le cas, que nous nommons *marchés de rente*. Pour les autres marchés, au contraire, *n'importe quel* objectif de redistribution peut être atteint tout en maximisant le surplus total du consommateur. Dans ce cas, notre analyse sélectionne une segmentation parmi l'infinité de segmentations maximisant le surplus du consommateur établies par Bergemann et al. (2015). Ces segmentations redistributives optimales présentent une forme étonnamment simple : elles divisent simplement les consommateurs en un segment de rabais et un segment résiduel. Le segment de rabais regroupe tous les consommateurs qui ne recevraient aucun surplus sous le prix uniforme (i.e., le prix que le producteur fixerait pour le

marché aggrégé), tandis que le segment résiduel regroupe tous les consommateurs restants.

Notre analyse fournit également des indications sur la manière de construire une segmentation optimale dans les marchés de rentes, lorsque les préférences redistributives sont suffisamment fortes. Dans ce cas, il faut diviser les consommateurs en segments contigus sur la base de leur propension à payer, les consommateurs ayant la même propension à payer appartenant à au plus deux segments différents. Ceci nous permet de construire une procédure qui génère une segmentation optimale.

Résumé

La thèse s'articule autour de trois parties. Les deux premiers chapitres traitent d'un modèle de changement de préférence délibéré. Selon ce dernier, un agent prend conscience de nouvelles dimensions du monde au cours du temps et peut décider en conséquence de modifier son système de valeur, induisant ainsi un changement de ses préférences. Ce modèle est entièrement caractérisé par deux fondements normatifs qui sont traduits en axiomes sur les comportements de choix d'un individu à travers plusieurs périodes consécutives. Une application montre certaines particularités du modèle et fournit une explication possible de la polarisation politique. Le troisième chapitre étudie l' « effet du fruit défendu », un phénomène régulièrement observé selon lequel restreindre l'ensemble d'opportunités d'un individu redirige son attrait vers les alternatives interdites ou leurs substituts. Nous proposons un modèle de choix qui tient compte de ce phénomène, dont les ingrédients sont identifiables par l'observation des réactions d'un agent à des restrictions, et qui est caractérisés par cinq axiomes sur les comportements de choix. Nous explorons les conséquences du modèle dans trois applications : la première traite de la formation des théories du complot ; la seconde de l'effet contre-productif des politiques d'intégration à destination des minorités ; la troisième reprend un problème classique de délégation dans un cadre de principal-agent. Nous dérivons également une mesure de la liberté de choix offerte par les différents ensembles d'opportunités possibles auquel l'agent puisse faire face. Le quatrième et dernier chapitre traite des effets redistributifs de la tarification personnalisée. Nous montrons que maximiser le surplus des consommateurs en priorisant les plus pauvres d'entre eux peut impliquer de donner un profit supplémentaire au producteur. Nous caractérisons les marchés pour lesquels c'est le cas. Nous des caractéristiques qualitatives de la politique optimale de discrimination par les prix et dérivons ainsi une procédure pour la construire.

Mots-clés

Théorie du choix – Théorie de la décision – Design d'information

Changements de préférences – Violations de WARP – Liberté de Choix

Discrimination par les prix – Segmentation

Summary

The thesis is organized into three parts. The first two chapters deal with a model of deliberate preference change. According to this model, an agent becomes aware of new dimensions of the world over time and may decide to change they value system as a result, thus inducing a preference change. This model is entirely characterized by two normative foundations that are translated into axioms on choice behaviors across several consecutive periods. An application shows some features of the model and provides a possible explanation of political polarization. The third chapter studies the "forbidden fruit effect," a regularly observed phenomenon in which restricting an individual's opportunity set redirects their attraction toward the forbidden options or their substitutes. We propose a model of choice that accounts for this phenomenon, the ingredients of which are identifiable by observing an agent's reactions to restrictions, and which is characterized by five axioms about choice behaviors. We explore the consequences of the model in three applications: the first deals with the formation of conspiracy theories; the second with the backlash effect of integration policies for minorities; the third takes up a classical delegation problem in a principal-agent framework. We also derive a measure of the freedom of choice offered by the different sets of opportunities that the agent can face. The fourth and last chapter deals with the redistributive effects of personalized pricing. We show that maximizing consumer surplus while prioritizing the poorest consumers may imply giving an additional profit to the producer. We characterize the markets for which this is the case. We derive qualitative characteristics of the optimal price discrimination policy and thus derive a procedure to construct it.

Keywords

Choice theory - Decision theory - Information design

Preference changes – Violations of WARP - Freedom of Choice

Price discrimination - Segmentation