<Your Thesis Title>

A thesis submitted for the degree of Doctor of Philosophy

by

Enrico Mattia Salonia

April 23, 2025

Contents

Chapter 1

Introduction

This thesis studies principles underlying individual behaviour, information processing, and resource allocation. It focuses on foundational issues at the core of behavioural economics, where the concepts investigated in this thesis are introduced. Behavioural economics—which examines variations in individual attitudes toward behaviour, belief revision, and normative stances—has historically been characterized by a reduced-form approach (Spiegler, 2019). This approach explains empirical observations through simple reinterpretations of classical theoretical constructs that intuitively capture relevant psychological mechanisms. In contrast, I argue that more nuanced treatments are necessary, requiring the development of novel theoretical tools rather than merely reinterpreting existing ones. As a result, I derive distinctions within the behavioural phenomena under study that are difficult to discern without explicit modelling.

The chapters appear in the chronological order of their conception. In Chapter 2, I study individuals who universalise their behaviour—that is, they consider what would happen if everyone behaved as they do, under different interpretations of this notion. In Chapter 3, I examine criteria for the allocation of resources in society that are responsive to considerations of merit and individual responsibility. In Chapter 4, I explore how individuals with preferences over their beliefs behave and interpret new information. Common to all three chapters is the use of the axiomatic method. This approach is natural given the aims of the thesis. In the words of Debreu: "*Allegiance to rigour dictates the axiomatic form of the analysis where the theory, in the strict sense, is logically entirely disconnected from its interpretations*" (Debreu, 1959, p. x). Given the conceptual complexity of the topics at hand, the axiomatic method ensures that the logical development of the theory is not contaminated by its interpretations, leading to "*a deeper understanding of the problem*" (Debreu, 1959, p. x).

In all three chapters, I first introduce a syntax that allows me to describe the objects of interest. The principles under study constitute particular instances of these objects. I then specify properties that elements of the syntax must satisfy to be equivalent to the object of interest. In particular, I characterise these objects as the unique elements satisfying the relevant properties. This method is useful because it allows one to express a theory of individual behaviour, belief revision or a principle for resource allocation using a small number of logically consistent conditions. These conditions serve a dual purpose: first, they can be normatively evaluated—whether they are conditions one might wish to respect when acting, processing information, or distributing goods. Second, they provide testable implications. If an individual's observed behaviour violates these conditions, then the theory is not a good description of that individual's behaviour. Each chapter

includes illustrative applications of the theory in canonical economic settings. I now turn to a brief description of the chapters.

Chapter 2: A Foundation for Universalisation in Games. In the first chapter, I study individuals with preferences for universalisation—that is, they consider what would happen if everyone were to act as they do. Universalisation has been shown to have evolutionary foundations, to align with experimentally observed behaviour, and to lead to desirable allocations under various normative frameworks. Existing models, such as Homo Moralis preferences (Alger & Weibull, 2013) and Kantian equilibrium (Roemer, 2019), lack choice-theoretic foundations, limiting their generalisability. To address this, I develop an axiomatic model characterising preferences for universalisation. The main challenge is that universalisation reflects a non-consequentialist attitude, which is difficult to capture using standard choice-theoretic tools. A key behavioural prediction of my model is that the independence axiom holds only among actions that are universalised in equivalent ways. My framework unifies previous models, introduces a broader class of universalisation preferences, and offers guidance for empirical investigation.

Chapter 3: Meritocracy as an End and as a Means. The second chapter studies the concept of meritocracy, widely discussed both publicly and in the economics and philosophy literature. An allocation is meritocratic if more meritorious individuals receive better rewards. Each form of meritocracy is characterised by two components: a merit criterion, which determines what counts as meritorious behaviour, and a reward criterion, which specifies how merit is translated into outcomes. By examining whether the allocation choices of impartial spectators align with particular merit and reward criteria, one can test the extent to which individuals adhere to different meritocratic principles. I consider two motivations for supporting meritocracy: rewarding merit as intrinsically fair—interpreting meritocracy as an end—and using meritocracy as an instrument to achieve other goals, such as efficiency—thus treating it as a means. I show that these two justifications are equivalent in terms of the rules they imply. Different assumptions about the merit and reward criteria accommodate various interpretations of meritocracy. I characterise and examine two meritocratic principles found in the literature: Pareto meritocracy, in which merit derives from generating a Pareto improvement, and proportional meritocracy, in which consumption increases proportionally with effort. I conclude by distinguishing my model from responsibility-sensitive egalitarianism, as in (Fleurbaey, 2008).

Chapter 4: Identifying Belief-dependent Preferences. The third chapter investigates individuals whose well-being—that is, their preferences over outcomes—is directly shaped by their beliefs. Such belief-dependent preferences explain a range of behaviours that deviate from expected utility theory. A growing body of evidence suggests that individuals selectively avoid or distort information, consistent with a preference for holding particular beliefs (Golman et al., 2017). When beliefs influence preferences over outcomes, belief formation itself may be endogenously shaped by those preferences. This interdependence complicates the task of inferring tastes and beliefs from choice data. The main con-

tribution of the chapter is to present a model of belief-dependent preferences combined with non-Bayesian updating, and to provide choice data sufficient to test and identify the model's components. I introduce a novel form of choice data, which generalises the notion of a menu in the menu-choice literature, and introduce axioms over preferences on such menus. To clarify the contrast with existing approaches, consider Brunnermeier & Parker (2005), where individuals choose their beliefs, balancing belief-based utility with material payoffs. While technically equivalent to standard models, this framework departs from prior decision-theoretic foundations by assuming endogenous belief choice. As noted by Eliaz & Spiegler (2006) and Spiegler (2019), this assumption complicates the interpretation and testability of the model. In contrast, my model uses standard tools from choice theory to identify the behavioural implications of belief dependence and provides clear conditions under which the theory can be falsified.

References

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302. 2

Brunnermeier, M. K., & Parker, J. A. (2005). Optimal expectations. *American Economic Review*, *95*(4), 1092–1118. 3

Debreu, G. (1959). *Theory of value: An axiomatic analysis of economic equilibrium* (Vol. 17). Yale University Press. 1

Eliaz, K., & Spiegler, R. (2006). Can anticipatory feelings explain anomalous choices of information sources? *Games and Economic Behavior*, *56*(1), 87–104. 3

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford University Press. 2

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of economic literature*, *55*(1), 96–135. 2

Roemer, J. E. (2019). *How we cooperate*. Yale University Press. 2

Spiegler, R. (2019). Behavioral economics and the atheoretical style. *American Economic Journal: Microeconomics*, *11*(2), 173–194. 1, 3

Chapter 2

A Foundation for Universalisation in Games

Abstract

I study the behaviour of individuals who have preferences for universalisation. When considering a course of action, they evaluate the consequence that would occur if everyone else acted equivalently, according to some criterion of equivalence. That is, they universalise their behaviour. I develop and axiomatise a model for individuals who value their choices in light of the consequences they induce when their action is universalised. The key behavioural prediction is that the independence axiom is satisfied only among actions that are universalised equivalently. I impose conditions to single out the most prominent models of universalisation, compare them, highlight and arguably overcome their limitations. I propose a unifying model of universalisation inspired by the equal sacrifice principle.

2.1 Introduction

What would I get if everyone acted as I do? An individual who acts based on the answer to this question exhibits universalisation reasoning. In group interactions, universalisation reasoning prescribes that individuals consider what would happen if everyone chooses the same action as them. Universalisation has been shown to have evolutionary foundations (Alger & Weibull, 2013) and aligns with behaviour observed in experiments (Levine et al., 2020; Miettinen et al., 2020; van Leeuwen & Alger, 2024). Furthermore, it leads to desirable allocations under several normative criteria (Roemer, 2010).

Universalisation appears in the literature in various forms, with two prominent formulations being Homo Moralis preferences (Alger & Weibull, 2013) and the Kantian equilibrium concept (Roemer, 2019). Nevertheless, these models lack choice-theoretic foundations, complicating their unification and empirical testing. Without such foundations, extending the models beyond symmetric settings becomes challenging. It is unclear what "behaving in the same way" means in asymmetric contexts. Furthermore, the conceptual relationship between universalisation and other pro-social preferences remains unexplored. More worryingly, the models' predictions depend on the labels assigned to the primitive objects of choice, namely, actions in games. Universalisation prescribes considering what happens when everyone chooses the same action, therefore, changing

the names of actions alters the predictions of these models. I show that developing choice-theoretic foundations for universalisation allows resolution of these issues.

I develop a model and introduce axioms that characterise preferences for universalisation. This characterisation enables the unification of previous models, rationalises existing empirical identification practices, and provides new testable predictions. I also introduce a new class of preferences for universalisation that are applicable to asymmetric settings. These preferences generalise the symmetric models, and their predictions are independent of the labelling of actions in games.

The main difficulty in modelling universalisation is that it is a non-consequentialist motivation. Preferences over actions do not depend on the material consequences these induce. Therefore, it is not straightforward to identify preferences for universalisation from choices over material consequences.[1] Economics is often resistant to considering non-consequentialist motivations (Fleurbaey, 2019). The classical models of Anscombe & Aumann (1963) and Savage (1972) illustrate this resistance. In these models, individuals rank mappings from uncertain states to consequences, usually referred to as "acts". Preferences for an act inducing a sure consequence are equivalent to preferences for that consequence. It is impossible to rank acts according to a criterion that does not depend on their induced consequences without trivializing such a notion, for example, by including the chosen act in the description of consequences. Thus, the question is whether universalisation, as a form of non-consequentialism, can be reconciled with the consequentialist approach of choice theory without resorting to ad hoc solutions.

I show that it is fruitful to study non-consequentialist decision criteria by taking a ranking over actions in a game, not consequences, as the primitive. An example in Section 2.1.1 illustrates that behaviour consistent with preferences for universalisation in a game cannot be rationalised by a preference ranking over material consequences. This motivates the use of Luce & Raiffa (1957)'s model, where the object of choice is an element of an action set. In a two-player game, an action induces an act, a mapping between the opponent's action and a consequence of the game. The novelty is that preferences over actions are not equivalent to preferences over the induced acts. In particular, the individual cares about the consequence that would obtain if his action is universalised. For instance, in a symmetric game, the individual considers what would happen if his opponent chooses the same action as he does. To capture more general criteria for universalisation, I introduce a universalisation function that maps an individual's action to an opponent's action, given a reference profile of actions. As an example, in Multiplicative Kantian Equilibrium (Roemer, 2019), individual actions that deviate from the reference profile by a specific proportion are universalised to opponent's action that deviates by the same proportion.

The main result, Theorem 2.1, provides a representation of preferences over mixed actions. The representation is a convex combination of two components. The first is the

[1] Sen (1973) suggested that non-consequentialism poses a challenge for revealed preference theory.

usual subjective expected utility. The second is the expected utility over the distribution of consequences obtained if the action is universalised. Therefore, preferences are a generalisation of subjective expected utility. The theorem allows to identify preferences over sure consequences up to the usual affine transformations and beliefs uniquely. The key deviation from expected utility is a violation of the independence axiom. In particular, independence holds only among actions that are universalised equivalently. Such violation of independence also constitutes a novel behavioural prediction. A direct test of a specific model of universalisation requires observing a violation of independence among actions that are universalised equivalently.

Because the theorem is silent on the shape of preferences over sure consequences, it reveals that universalisation and pro-social preferences are distinct assumptions; it is possible for an individual to exhibit both, consistent with empirical evidence (van Leeuwen & Alger, 2024). Moreover, the theorem implies that welfare analysis for individuals with preferences for universalisation cannot use material consequences as a currency, contrary to the standard practice in Kantian Equilibrium models (Roemer, 2019). An individual with consequentialist preferences can always be compensated with material payoff, such as money, to refrain from taking a specific action. This is not true for individuals with preferences for universalisation, as they desire to induce a specific consequence as a result of their action being universalised. Non-consequentialist individuals thus suffer when they cannot choose the action they prefer, regardless of any material compensation. I thus argue that welfare criteria for non-consequentialist preferences may encompass a form of freedom of choice.[2]

By specifying the universalisation function, I provide a choice-theoretic foundation for Homo Moralis preferences for universalisation à la Alger & Weibull (2013) and the various definitions of Kantian Equilibrium by Roemer (2019), both of which constitute a generalisation of the model by Laffont (1975). I comment on the difference between my foundation for Kantian Equilibrium and that of Roemer. He suggests that his model does not assume preferences that deviate from selfish material satisfaction, but rather a different "optimisation protocol". I argue that his model's properties can be preserved by abandoning the distinction between the optimisation protocol and preferences, resulting in a more parsimonious framework in line with classical choice theory.

I develop a novel concept of universalisation inspired by the equal sacrifice principle (Mill, 1885; Young, 1988). Consider an individual with any given aim. Given a profile of actions in a game, the individual evaluates a deviation by considering the consequence that would occur if their opponents also deviated to induce an equivalent difference in aim satisfaction, that is, an equal sacrifice. I show that this form of universalisation is equivalent to that of Homo Moralis and Kantian Equilibrium in symmetric games. Moreover, its predictions do not depend on the labelling of actions, nor does its definition require the veil of ignorance construct used to define Homo Moralis in asymmetric contexts.

[2]See, for example, Fleurbaey (2008, Ch. 10).

The paper is organized as follows: In Section 2.2, I introduce the primitives of the model and the axioms. The main theorem is presented in Section 2.3. In Section 2.4, I show how assumptions on the universalisation function allow me to derive various models of universalisation. Equal sacrifice universalisation is introduced in Section 2.5. Section 2.6 concludes the paper. A literature review and illustrative example follow.

Related literature. In this paper, I study a decision problem as modelled in Luce & Raiffa (1957). The analysis builds on results by Battigalli et al. (2017), who study Luce & Raiffa decision problems and connect them to the approach in Anscombe & Aumann (1963).

The model here is reminiscent of context-dependent preferences by Gilboa & Schmeidler (2003). They study collections of individuals' preferences, one for each possible belief, over their actions and an uncertain state. As in this paper, the state is interpreted as opponents' choices. They also start from a primitive ranking over individuals' actions and obtain an expected utility representation in games. However, I here study a subjective beliefs setting, where these are derived from behaviour.

The intuition that non-consequentialist individuals do not care about an act because of its consequences has been highlighted by Chen & Schonger (2022), who develop a choice-theoretic model to guide an experiment testing for the presence of non-consequentialist preferences. They argue that, to identify non-consequentialism from choice, individuals must face the possibility that their actions will not be implemented or observed by the experimenter. Their model has a different interpretation compared to mine. In their experiment, subjects knew that there was a chance that their action would not have been implemented, whereas here there is no such possibility.

The model in this paper allows me to distinguish universalisation from the related concept of magical thinking, studied from a choice-theoretic perspective by Daley & Sadowski (2017). An individual exhibits magical thinking if he expects the probability the opponent selects a specific action to increase if he chooses that action. They provide axioms on behaviour in symmetric games that characterize magical thinking. I show that magical thinking and universalisation are different from a choice-theoretic perspective. An individual with preferences for universalisation does not believe he affects opponents' choice.

In this paper, I provide a choice-theoretic foundation for various models of universalisation. The two main alternatives are Homo Moralis preferences by Alger & Weibull (2013, 2016); Alger et al. (2020) and Kantian Equilibrium by Roemer (2010, 2015, 2019). In two-player games, Homo Moralis maximises a convex combination of his payoff and the payoff he would obtain if his opponent behaved as he does. The authors show that, among the set of continuous preferences, Homo Moralis is the only one that is evolutionary stable for all the game protocols their model covers, when interactions take place under incomplete information, and there is assortativity in the process. The result is generalised to multiplayer games and structured populations by Alger & Weibull (2016) and Alger et al. (2020). Roemer (2019) introduces a new solution concept, Kantian Equi-

librium. He argues that, if individuals are Kantian rather than Nash optimisers, when considering deviating from an action profile they assume other players will deviate in an equivalent manner, where "equivalent" is defined in various ways. Alger & Weibull derive novel preferences from evolutionary analysis and Roemer changes the equilibrium concept, when compared with selfish/Nash individuals. I comment on the relation between these two models in the body of the paper.

This paper relates to studies of universalisation and other non-consequentialist motivations in various settings. Some of these study moral attitudes or their relation with pro-social preferences, as Dewatripont & Tirole (2024), Ellingsen & Mohlin (2024), Fleurbaey et al. (2024) and Laslier (2022). Others are applications in economic environments, including bargaining (Dizarlar & Karagözoğlu, 2023; Juan-Bartroli & Karagözoğlu, 2024), contract theory (Sarkisian, 2017, 2021a,b), public goods (Brekke et al., 2003), social norms (Juan-Bartroli, 2024), taxation (Sobrado, 2022), vaccination (De Donder et al., 2023) and voting (Alger & Laslier, 2022; Dierks et al., 2024; Grillo, 2022). Finally, there is interest in choice-theoretic models of individual moral attitudes. For example, Ponthiere (2023), Ponthiere (2024) and Shi (2024) study, respectively, Epictetianism, Stoicism and preference for a social minimum consumption level.

2.1.1 Illustrative Example

I briefly illustrate the contribution of this paper through an example. I show that preferences for universalisation cannot be rationalised by a preference ranking over material consequences, and that predictions of previous models depends on the labeling. I then discuss the solution I propose and how it relates to the existing literature.

Two individuals play the following game. They can go left (ℓ), middle (m) or right (r). The numbers in the table are monetary rewards.

μ'		$1/2$	$1/2$
μ	$1/2$	$1/2$	
	ℓ	m	r
ℓ	$1,1$	$0,0$	$0,0$
m	$0,0$	$0,0$	$1,1$
r	$0,0$	$1,1$	$0,0$

Table 2.1: Preference reversal.

Assume the row player has beliefs μ in Table 2.1 and thus conjectures his opponent will play ℓ or m, each with probability $\frac{1}{2}$. By choosing a mixed action, the row player can induce any distribution over consequences that mixes between $(0,0)$ for sure and $(1,1)$ or $(0,0)$ with equal probability. If the row player has preferences for universalisation, he will choose ℓ, since it is the action that, if implemented by everyone in this game, max-

imises his monetary payoff. From a revealed preference perspective, it is inferred that he prefers the lottery $\frac{1}{2}(1,1)+\frac{1}{2}(0,0)$ to the sure consequence $(0,0)$. Now, consider a second scenario where the same individual has beliefs μ' in Table 2.1, according to which his opponent plays m or r with probability $\frac{1}{2}$. The feasible set of lotteries over consequences is the same as before. Actions m and r induce the midpoint between $(0,0)$ and $(1,1)$ whereas ℓ induces the sure consequence $(0,0)$. The row player still chooses ℓ, as it is again the action that maximises his payoff if implemented by everyone. When $(0,0)$ was available, he revealed to prefer $\frac{1}{2}(1,1)+\frac{1}{2}(0,0)$. Nevertheless, he exhibits a preference reversal in the second scenario, thus violating the weak axiom of revealed preference. There is no complete and transitive preference relation on lotteries consistent with this choice pattern. This impossibility does not occur for consequentialist preferences defined on distributions of material consequences, such as selfishness, altruism, inequity aversion, or maximin. Therefore, functional forms for preferences for universalisation in the literature represent orderings over objects that are different from distributions over material consequences. This implies that preferences over material consequences should not be the relevant measure for welfare analysis of an individual exhibiting universalisation reasoning, contrary to what Roemer (2019) proposes.

This example also shows that the predictions of models of universalisation depend on the labelling of actions. To avoid the preference reversal, it would suffice to swap the labels of one individual's actions, changing m to r and vice versa. Indeed, Roemer (2019) discusses in multiple instances how to change the label of actions to define and employ universalisation. In Section 2.5, I present a novel definition of universalisation, relying on the general theory, that is equivalent under any redescription of actions.

2.2 Model

In this section, I introduce the primitives of the model and the axioms I consider. For any set Y, I denote with $\Delta(Y)$ the set of finite probability distributions over Y.

Primitives. I focus on two-player games, defined as follows.

Definition 2.1. *A **two-player game** is a list $G=(\{1,2\},(A_i,\succsim_i)_{i\in\{1,2\}},X,\rho)$, featuring:*[3]

- *a finite set of actions A_i for each player i;*
- *a common set of consequences X;*
- *a consequence function $\rho\colon A_i\times A_{-i}\to X$;*
- *player i's preferences over mixed actions $\succsim_i$, for each player i.*

[3]The textbook by Bonanno (2018) discusses games whose primitives are ordinal preferences.

Each pair of pure actions (a_i, a_{-i}) induces a consequences $x = \rho_{a_i,a_{-i}}$ where $x \in X$. Any mixed action $\alpha_i \in \Delta(A_i)$ induces an Anscombe & Aumann act denoted with $\rho_{\alpha_i} \colon A_{-i} \to \Delta(X)$ leading to consequence x under opponent's action a_{-i} with probability $\rho_{\alpha_i,a_{-i}}(x) = \alpha_i(\{a_i \in A_i \mid \rho_{a_i,a_{-i}} = x\})$. Each pair of mixed actions (α_i, α_{-i}) induce the constant act $\rho_{\alpha_i,\alpha_{-i}} \in \Delta(X)$. I say that a mixed action α_i induces a constant act if $\rho_{\alpha_i,a_{-i}}(x) = \rho_{\alpha_i,a'_{-i}}(x)$ for each pair a_{-i}, a'_{-i} and x. I assume there exist mixed actions that, under various opponent's actions, can induce every possible distribution of consequences. A sufficient condition for this to hold is that for each consequence x there exists an action a_i such that $\rho_{a_i,a_{-i}} = x$ for each opponent's action a_{-i}. I need such richness assumption to identify preferences, as usual in decision theory. However, in examples of games in this paper, usually only a subset of A_i of feasible actions is available.

I now introduce a primitive instrumental to capture universalisation reasoning. The idea of universalisation is that, when individual i is evaluating mixed action α_i, he considers the distribution over consequences that would occur if his opponent plays equivalently, under some notion of equivalence. As an example, when the game is symmetric and the action set is the same for both players, he might consider the distribution of consequences induced when his opponent also plays α_i. Fix a reference mixed action profile $(\alpha^*_i, \alpha^*_{-i})$. A universalisation function $T_{(\alpha^*_i,\alpha^*_{-i})} \colon \Delta(A_i) \to \Delta(A_{-i})$, maps individual i's mixed action to an opponent's mixed action, given a reference profile. For each mixed action α_i, the corresponding $-i$ universalised action is $T_{\alpha^*_i,\alpha^*_{-i}}[\alpha_i]$.

Consider two pure actions inducing the same act. If the individual is consequentialist, he should be indifferent between these two actions, as they induce the same consequences under each opponent's action. Under consequentialism, it would be without loss of generality to study a game in which two actions inducing the same act are identified as the same action. However, preferences for universalisation are not consequentialist, they cannot be reduced to preferences over acts. Therefore, I impose a different notion of equivalence between actions.

I refer to two actions a_i and a'_i as realisation equivalent if

$$\left(\rho_{a_i}, \rho_{a_i,T_{\alpha^*_i,\alpha^*_{-i}}[a_i]}\right) = \left(\rho_{a'_i}, \rho_{a'_i,T_{\alpha^*_i,\alpha^*_{-i}}[a'_i]}\right).$$

Namely, two actions are realisation equivalent when they induce the same act and the same distribution over consequences when they are universalised. Consider an individual who cares only about the act he induces and the constant act induced under universalisation reasoning. Then, he would be indifferent between two realisation equivalent actions. A game is **reduced** if realisation equivalent actions are the same action.

I study preferences over mixed actions $\succsim_i$ of a generic individual i. I introduce axioms on $\succsim_i$ that characterise the following functional representation.

Definition 2.2. *A ranking $\succsim_i$ is a **Universalisation Preference** (UP) with respect to the universalisation function $T_{\alpha^*_i,\alpha^*_{-i}}$ if it is represented by*

$$U_i(\alpha_i) = (1-\kappa) \sum_{a_i, a_{-i}} \alpha_i(a_i)\mu_i(a_{-i})u_i(\rho_{a_i,a_{-i}}) + \kappa \sum_{a_i, a_{-i}} \alpha_i(a_i)T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i](a_{-i})u_i(\rho_{a_i,a_{-i}}), \tag{2.1}$$

for some utility function $u_i \colon X \to \mathbb{R}$ *and belief* $\mu_i \in \Delta(A_{-i})$.

A *UP* is a linear combination of two components. The first component, weighted by $1-\kappa$, is a standard subjective expected utility. The individual computes the probability the action profile (a_i, a_{-i}) realises, which depends on his mixed action α_i and his belief over opponent's actions μ_{-i}. Then, he evaluates the consequence $\rho_{a_i,a_{-i}}$ obtained according to the utility u_i. The second component, weighted by κ, is the result of universalisation reasoning. Instead of evaluating the probability an opponent's action realises according to the belief μ_i, the individual considers the opponent's mixed action that results from universalising his action via the universalisation function $T_{\alpha_i^*,\alpha_{-i}^*}$. As an example, when the game is symmetric, and then $A_i = A_{-i}$, one can define the identity universalisation function as $T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i] = \alpha_i$ for each α_i, regardless of the reference profile. I show in Section 2.4 that the identity universalisation function singles out Homo Moralis preferences from Equation (2.1).

Axioms. I now introduce the axioms that characterise *UP*. I start with standard axioms allowing me to obtain a utility representation of preferences over mixed actions.

Axiom 2.1. *(**Weak order**) Preferences* $\succsim_i$ *are a continuous weak order.*

Axiom 2.2. *(**Non-triviality**) There exist* α_i, α_i' *such that* $\alpha_i \succ \alpha_i'$.

I now proceed with axioms that characterise preferences for universalisation. First, the individual only satisfies independence among actions that are universalised equivalently. The intuition is as follows. If the individual was a consequentialist, he would satisfy independence, which would result in the standard independence condition among acts. However, the individual is also interested in the distribution of consequences induced when his action is universalised. A mixture of two actions induce a mixture in their corresponding universalised action. Therefore, when mixing, the distribution over consequences induced by the action and its universalised counterpart is not guaranteed to change linearly. Linearity is guaranteed only if the two actions are universalised equivalently.

Axiom 2.3. *(**Universalisation Independence**) If*

$$T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i] = T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i'] = T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i''],$$

then, for all $\lambda \in (0,1)$,

$$\alpha \succsim_i \alpha_i' \implies \lambda\alpha_i + (1-\lambda)\alpha_i'' \succsim_i \lambda\alpha_i' + (1-\lambda)\alpha_i''.$$

The next axiom states that, when two actions induce the same act, then their ranking depends on the distribution of consequences they induce when universalised. It restricts attention to preferences over actions that not only depend on the induced act, but also on the distribution of consequences induced by the universalised action.

Axiom 2.4. *(**Universalisation evaluation**) If $\rho_{\alpha_i} = \rho_{\alpha_i'}$, then*

$$\alpha_i \succsim_i \alpha_i' \text{ if and only if } \rho_{\alpha_i, T_{\alpha_i^*, \alpha_{-i}^*}[\alpha_i]} \succsim_i \rho_{\alpha_i', T_{\alpha_i^*, \alpha_{-i}^*}[\alpha_i']} .$$

Lastly, I assume the individual satisfies independence among actions inducing constant acts. The reason is the following. A constant act induces the same distribution of consequences regardless of the opponent's action. Therefore, regardless of how the action is universalised, the opponent is not able to affect the distribution of consequences. The reason for violating independence therefore decades when considering constant acts.

Axiom 2.5. *(**Lotteries independence**) If α_i, α_i' and α_i'' induce constant acts, then for all $\lambda \in (0,1)$,*

$$\alpha \succsim_i \alpha_i' \implies \lambda\alpha_i + (1-\lambda)\alpha_i'' \succsim_i \lambda\alpha_i + (1-\lambda)\alpha_i''.$$

As an alternative, one could dispense from Lotteries independence and assume that all actions inducing constant acts are universalised equivalently. Then, Universalisation Independence would imply Lotteries independence. The next section studies the implication of imposing these axioms on preferences over mixed actions.

2.3 Functional Representation

The main result of this paper shows that the axioms in the previous section are necessary and sufficient to characterise *UP*.[4]

Theorem 2.1. *A ranking $\succsim_i$ satisfies Weak order, Non-triviality, Universalisation Independence, Universalisation evaluation and Lotteries independence in a reduced game if and only if it is a UP. Moreover, the utility function u_i is unique up to affine transformations and beliefs μ_i are unique.*

Theorem 2.1 states that choices of mixed actions satisfying the axioms are consistent with the following utility function: when choosing the mixed action α_i, the individual evaluates the probability that each opponent's action a_{-i} realises according to his subjective belief μ_i. However, he also considers the distribution of consequences induced by his mixed action α_i and the universalised action according to the universalisation function $T_{\alpha_i^*, \alpha_{-i}^*}$. The two components are aggregate linearly.

[4]All proofs are in Appendix A.2.

I do not derive the form of u_i, the individual may have any preferences over consequences. This fact clarifies the difference between my exercise and, as an example, that of Rohde (2010). Rohde (2010) establishes conditions on a ranking over collective monetary rewards that characterise inequity aversion. In the language of the present paper, she studies the shape of u_i. The axioms here imply nothing about such shape. The representation allows the individual, as an example, to both exhibit preferences for universalisation and, say, inequity aversion, as captured by u_i. Then, in a game, the individual would choose the action that, if universalised, satisfies his inequity averse preference. Theorem 2.1 thus clarifies that pro-social and non-consequentialist preferences are not exclusive. On the contrary, these two can coexist.

Lastly, Theorem 2.1 allows marking the difference between universalisation and magical thinking. An individual exhibiting magical thinking believes he affects the opponent's probability to choose an action by choosing it himself. An individual with preferences for universalisation, instead, develops standard subjective beliefs about opponents' actions, and his behaviour does not affect them. Since beliefs are standard, their updating should be consistent with Bayes rule in dynamic settings. The first component of the utility function, representing preferences over the induced act, is standard, and therefore results on Bayesian updating holds.[5] In the next section, I study different forms of the universalisation function corresponding to particular preferences in the literature.

2.4 Preferences for Universalisation

In this section, I study conditions on the universalisation function under which *UP* preferences are equivalent to various notions of universalisation in games. I start with Simple Kantian Equilibrium by Roemer (2019), to later proceed with Homo Moralis by Alger & Weibull (2013) and conclude with Multiplicative Kantian Equilibrium by Roemer (2019). I supplement results with discussions on the interpretation of these concepts and the relation between them.

2.4.1 Homo Kantiensis and Simple Kantian equilibrium

In this section, I restrict attention to games with common action sets, where $A_1 = A_2 = A$. Simple Kantian Equilibrium is defined as follows.

Definition 2.3. *An action profile (α, α) constitutes a **Simple Kantian Equilibrium** (SKE) of a game with common action sets if, for all players i and actions α'*

$$\sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i})u_i(\rho_{a_i,a_{-i}}) \geq \sum_{a_i, a_{-i}} \alpha'(a_i)\alpha'(a_{-i})u_i(\rho_{a_i,a_{-i}}).$$

[5] See e.g. Epstein & Schneider (2003); Ghirardato (2002).

A symmetric mixed action profile constitutes a *SKE* if it induces the best distribution over consequences over all symmetric mixed action profiles. I show that a *SKE* can be interpreted as a Nash Equilibrium in a game between two players with *Homo Kantiensis* preferences.

Definition 2.4. *A ranking $\succsim_i$ is a **Homo Kantiensis** (HK) preference if it is represented by*

$$U_i(\alpha) = \sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i})u_i(\rho_{a_i,a_{-i}}), \tag{2.2}$$

for some utility function $u_i \colon X \to \mathbb{R}$.

When evaluating any mixed action α, a *HK*, first introduced in Laffont (1975), only considers the distribution over consequences induced when his opponent chooses α as well. A *HK* is a particular case of a *UP* preference. If $\kappa = 1$ and $T_{\alpha_i^*, \alpha_{-i}^*}[\alpha] = \alpha$ for each α, then Equation (2.1) reduces to Equation (2.2). In other words, a *HK* satisfies the axioms in Theorem 2.1 with respect to the identity universalisation function.

Proposition 2.1. *An action profile (α, α) constitutes a SKE in a game with common action sets if and only if it constitutes a Nash Equilibrium between two HK.*

Proposition 2.1 thus establishes that *SKE* is a Nash Equilibrium in a game between two players with preferences over mixed actions satisfying the axioms in Theorem 2.1 with respect to the identity universalisation function. The result allows me to compare the foundation I offer for *SKE* with that of Roemer (2019). He argues that, contrary to other models in economics, he does not assume exotic preferences, but classical self-regarding attitudes.[6] What he varies, instead, is individuals' "optimisation protocol", as he refers to it. He contrasts Nash optimisation with Kantian optimisation. Nash optimisation, he maintains, relies on the counterfactual "what would happen were I to change my action alone?". Instead, Kantian optimisation induces the counterfactual "what would happen were I and all others to deviate equally?" This argument is echoed in the papers employing various declinations of Kantian Equilibrium.[7]

In the following, I argue that, although appealing, such reasoning cannot be backed up by classical choice theory. I do not take any stance on this point. It is legitimate to employ concepts that diverge from standard theory. Nevertheless, this incompatibility is particularly relevant here, as Roemer relies on his distinction between preferences and optimisation protocol to derive welfare statements.

Roemer's description of the Nash counterfactual refers to the logic employed to check whether an action profile constitutes a Nash Equilibrium. Nevertheless, this is only vaguely related to the foundation of the concept.[8] Outside contexts of long repeated

[6] See, among many others, Roemer (2019, p. 69).

[7] See the papers in the literature review in Section 2.1.

[8] Battigalli et al. (2023) offer a thorough discussion on the interpretation of Nash Equilibrium.

interactions and adaptive dynamics, an action in a Nash Equilibrium profile is played by an individual holding correct conjectures about opponents' behaviour.[9] However, players cannot perform the Nash counterfactual exercise, because they do not know what opponents will do, and are unable to evaluate the gain obtained from a unilateral deviation. An individual in a game selects the action that he considers the best one according to his beliefs about what his opponents will do. In turn, the definition of "best" is, in economics, his preference. In choice theory, observed behaviour is interpreted as revealing a preference for an object compared with others available, actions in this case. Optimisation is a mathematical technique employed to compute what the maximal element is given a primitive ranking over the objects of choice, it is not a feature of the individual or of an equilibrium concept. There is no empirical observation able to tell that two individuals have the same preferences but different optimisation protocols. If they choose differently in the same problem, by definition they have different preferences.

I show with Proposition 2.1 that there is no need to rely on informal arguments regarding how individuals optimise. Behaviour consistent with *SKE* can be interpreted as Nash Equilibrium behaviour in a game between two *HK*. Therefore, Roemer is correct in arguing that assuming individuals behave according to *SKE* is different from saying that they are pro-social. Nevertheless, this does not mean that they optimise differently.

The critique above has implications for welfare analysis. Roemer's argument according to which, in *SKE*, individuals have selfish preferences over material consequences but the optimisation protocol is different from Nash, generates confusion. As I showed in the motivating example, it is possible that an individual who plays according to *SKE* does not have a complete and transitive preference, and hence a utility representation, over material consequences. I believe the closest reformulation of Roemer's point is that one can have preferences for universalisation even if the utility index in Theorem 2.1 for material consequences u_i is the same as that for consequences induced by universalised actions, as in *UP* preferences in Equation (2.1). Nevertheless, this equivalence does not imply the individual would be indifferent between receiving a monetary amount and acting to induce it as a consequence of universalisation reasoning. Great care must be devoted to make welfare statements for non-consequentialist preferences over actions. Given that universalisation is a preference over actions, one interesting avenue is to consider that welfare should be evaluated in terms of the freedom the individual has in choosing an action he prefers.[10]

Proposition 2.1 also offers a novel rationale for using mixed actions. Under expected utility, there is always a pure action in the set of best replies to probabilistic conjectures regarding opponents' behaviour. The Nash equilibrium mixed action of player i can be interpreted as strategic uncertainty from player $-i$'s perspective. Nevertheless, a *HK* who plays a mixed action in a *SKE* profile has no interest in being difficult to be predicted by his opponents. In his best reply set, there may be no pure actions. A rationale for

[9] See Perea (2012) or Dekel & Siniscalchi (2015) and references therein.

[10] Laslier et al. (1998) offer a review of approaches on how to conceptualise freedom in economics.

employing mixed actions is therefore the adherence to a non-consequentialist attitude.

2.4.2 Homo Moralis

In this section, I exploit the representation in Theorem 2.1 to derive Homo Moralis preferences as a special case of Equation (2.1). I again restrict attention to games with common action sets, to define Homo Moralis preferences as follows.

Definition 2.5. *A ranking $\succsim_i$ is a **Homo Moralis** (HM) preference if it is represented by*

$$\begin{aligned} U_i(\alpha) = (1-\kappa) \sum_{a_i, a_{-i}} \alpha(a_i)\mu_i(a_{-i})u_i(\rho_{a_i,a_{-i}}) \\ + \kappa \sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i})u_i(\rho_{a_i,a_{-i}}), \end{aligned} \tag{2.3}$$

for some utility function $u_i \colon X \to \mathbb{R}$ and belief $\mu_i \in \Delta(A_{-i})$.

A Homo Moralis maximises a convex combination between subjective expected utility and expected payoff when both individuals play his action. Contrary to *SKE*, *HM* is a preference, not a property of action profiles. A *HM* with $\kappa = 1$ is a *HK* and his preferences are represented by Equation (2.2). A *HM* is a *UP* where the universalisation function is the identity, as in *HK*. If $T_{\alpha_i^*, \alpha_{-i}^*}[\alpha] = \alpha$ for each α, then Equation (2.1) coincides with Equation (2.3).

A *HM* is not only interested in the consequence obtained when his action is universalised, but trades off consequentialist and non-consequentialist motives. Since *HM* is partially strategic, he also cares about his opponent's action and thus his beliefs matter. However, contrary to magical thinking, a *HM* exhibits standard beliefs that do not depend on his action. Such distinction rationalises current experimental practices relying on identifying beliefs of individuals with preferences for universalisation (van Leeuwen & Alger, 2024). Moreover, my analysis offers a new behavioural prediction to test for the presence of *HM*, and therefore also of *HK*: they both violate independence between mixed actions that are not universalised equivalently, according to the identity universalisation function.

Both *SKE* and *HM* are well-defined only in games with common action sets. Alger & Weibull (2013) suggest a way to employ *HM* preferences in asymmetric games. They propose to consider an incomplete information expansion of the basic game where players are not aware of their role, reminiscent of the veil of ignorance of Harsanyi (1955) and Rawls (1971). Such incomplete information game is a symmetric interaction where a strategy is a map between role and action. Universalisation can then be defined as strategies are common across players. The authors refer to this preference as *Ex-ante Homo Moralis*. Another definition of universalisation in asymmetric games is Multiplicative Kantian Equilibrium by Roemer (2019). In the next section, I discuss Multiplicative Kantian Equilibrium, postponing observations on *Ex-ante HM* to Section 2.5.

2.4.3 Multiplicative Kantian equilibrium

In this section, I discuss the relationship between Multiplicative Kantian Equilibrium and *UP*.[11] The solution concept is defined in games where the action space has a linear structure. It is employed when players can choose a number from the real line. For simplicity, I follow the recommendation in Roemer (2019, p. 42) and consider the mixed extension of two-player two-actions games, though developing a generalisation of Multiplicative Kantian Equilibrium to multiple actions games is not trivial.

I remove the restriction to games with common action sets and assume players only have two pure actions available. For each α_i and $r \geq 0$, I denote with $r \cdot \alpha_i$ an operation that affects α_i by the multiplicative factor r and $1 - \alpha_i$ by the complementary weight to obtain a probability distribution on pure actions, that is

$$r \cdot \alpha_i := \frac{r\alpha_i}{r\alpha_i + (1 - \alpha_i)} .$$

Definition 2.6. *An action profile (α_i, α_{-i}) constitutes a **Multiplicative Kantian Equilibrium** (MKE) of a game if, for all players i and real numbers $r \geq 0$*

$$\sum_{a_i, a_{-i}} \alpha_i(a_i)\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}) \geq \sum_{a_i, a_{-i}} r \cdot \alpha_i(a_i) r \cdot \alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}).$$

A mixed action profile constitutes a *MKE* if it induces the best distribution over consequences when compared with all mixed action profiles that can be obtained by multiplying both actions by r.[12] The notion is well defined as I am restricting attention to two actions. A multiplicative deviation is equivalent to moving weight from one action to the other.

Paralleling the analysis for *SKE*, I show that *MKE* can be interpreted as a Nash Equilibrium in a game between two individuals with Multiplicative Homo Kantiensis preferences, defined as follows.

Definition 2.7. *A ranking $\succsim_i$ is a **Multiplicative Homo Kantiensis (MHK)** preference relative to the profile (α_i, α_{-i}) if it is represented by*

$$U_i(r \cdot \alpha_i) = \sum_{a_i, a_{-i}} r \cdot \alpha_i(a_i) r \cdot \alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}), \tag{2.4}$$

for some utility function $u_i \colon X \to \mathbb{R}$.

The difference between *MHK* and *HK* is how actions are universalised, as illustrated in Figure 2.1. A *HK* only conceives both players choosing the same action. In two-player symmetric games with two actions, these profiles correspond to the diagonal of the square representing mixed actions. Instead, *MHK* considers actions as multiplicative deviations

[11] An equivalent analysis delivers similar results for Additive Kantian Equilibrium (Roemer, 2019).
[12] According to this definition, $(0, 0)$ is always a *MKE* if it is available.

from a specific profile. Universalised actions lie on the line connecting the origin and the reference profile, i.e., all the pairs in which the ratio between the two actions is preserved. A profile (α_i, α_{-i}) constitutes a *MKE* if it is the preferred one for both players compared with any other on the line joining the origin and (α_i, α_{-i}). If (α_i, α_{-i}) lays on the 45° line, the two universalisation criteria are identical. In the next section, I discuss a novel version under which the universalisation criterion is endogenous and depends on the game at hand.

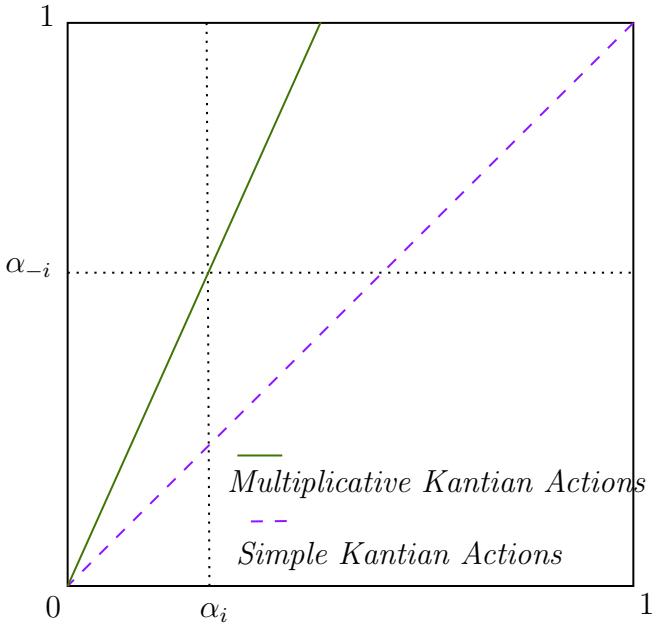

Figure 2.1: Universalised action profiles of Simple and Multiplicative Kantian Equilibria.

A *MHK* is a particular case of *UP* in which the actions are universalised through a multiplicative logic. If $\kappa = 1$ and $T_{\alpha_i,\alpha_{-i}}[r \cdot \alpha_i] = r \cdot \alpha_{-i}$ for each r, then Equation (2.1) reduces to Equation (2.4).

Proposition 2.2. *An action profile (α_i, α_{-i}) constitutes a MKE in a game if and only if it constitutes a Nash Equilibrium between two MHK relative to the profile (α_i, α_{-i}).*

Proposition 2.2 has an interpretation on the lines of its parallel result for *SKE*. It establishes that *MKE* is a Nash Equilibrium in a game between two players with preferences over mixed actions satisfying the axioms in Theorem 2.1 with respect to the multiplicative universalisation function. Contrary to *SKE*, *MKE* can be defined when action sets are not common and allows individuals to choose heterogeneous actions, as *Ex-ante HM* does. In the next section, I develop a new concept that takes a different route to define universalisation in games with general action sets.

2.5 Equal Sacrifice Universalisation

In this section, I elaborate on the concept of universalisation and present a new notion, inspired by the equal sacrifice principle (Mill, 1885; Young, 1988). The model I propose

has several features: its definition does not depend on the label of actions; it can be defined in asymmetric games; in symmetric games it is equivalent to *HM*.

Universalisation requires the definition of two objects. First, it must be transparent what "doing the same thing" is. Second, it must be equally clear what "deviating in the same manner" means. For these two concepts to be defined, a common currency must exist for the adjective "same" to have meaning. Previous ideas employed the label of actions in games and a notion of distance between them when the action space is structured. Such an approach, I argue, is partially lacking. In most economic models, the label of actions bears no conceptual relevance and might be misleading to use it as the main ingredient of a model of universalisation. In fact, in many applications where *MKE* gives intuitive results, the action labels have clear conceptual significance, as they represent effort, contribution to a public good, or use of a common resource.

I propose to use the relevant consequence of the game as a currency. In game theory, this is usually players' utility, but it can be any other index of well-being. Then "doing the same thing" and "deviating in the same manner" are interpreted as "inducing the same utility" and "inducing the same difference in utility". The following example illustrates the idea. Consider two individuals playing the prisoners' dilemma. Numbers are Bernoulli utilities for consequences.

$1 \setminus 2$	a_2	a_2'
a_1	$2, 2$	$0, 3$
a_1'	$3, 0$	$1, 1$

Table 2.2: Prisoners' dilemma.

Row player 1 attains his highest Bernoulli utility in $(3, 0)$, induced by the profile (a_1', a_2). I define α_1^s as a mixed action which, compared to a_1', given a_2, induces a reduction in expected Bernoulli utility by s, i.e., $3 - [2\alpha_1^s + 3(1 - \alpha_1^s)] = s$. The profile leading to the highest Bernoulli utility for column player 2 is instead (a_1, a_2'). As for player 2, his mixed actions α_2^s induces the difference $3 - [2\alpha_2^s + 3(1 - \alpha_2^s)] = s$. Assume that player 1, when picking any action α_1^s, considers a scenario where 2 chooses α_2^s. He envisions the consequence that would obtain if his opponent chooses the action that, if considered a unilateral deviation from (a_1, a_2'), generates the same difference in Bernoulli utility. In this game, $\alpha_1^s = \alpha_2^s = s$ for all s. Whenever they deviate from the action profile that yields their preferred material outcome, both players consider a scenario in which their opponents also deviate by choosing the same action. Therefore, the universalisation is identical to the one of *HK* and *HM* in this example. The actions that lead to the highest Bernoulli utility under this universalisation reasoning are a_1 for 1 and a_2 for 2, i.e., $\alpha_1^s = \alpha_2^s = 1$, the same optimal actions of *HK* under proper re-labelling of actions.

Evaluating differences from the maximum attainable payoff is reminiscent of the equal sacrifice principle of Mill (1885) in the context of taxation. Hence, I dub this concept

equal sacrifice universalisation (*ESU*). An individual with *ESU* preferences first identifies the profile of actions inducing his preferred consequence. Second, he evaluates each action considering the induced difference in payoff compared with the optimal action computed previously. Third, he individuates the collection of opponents' deviations that, compared with their maximal action profiles in the material dimension, lead to obtain the same absolute difference.

To ease the exposition, I here focus on equal absolute sacrifice (Young, 1988). In Appendix A.1, I consider general equal sacrifice rules. The results and arguments in this section hold for any equal sacrifice rule. Under Weak order, Non-triviality and Lotteries independence, preferences over actions inducing constant acts satisfy the standard Anscombe & Aumann conditions. These axioms guarantee the existence of two distributions over consequence $\overline{\gamma}, \underline{\gamma} \in \Delta(X)$ such that $\overline{\gamma} \succsim_i \gamma \succsim_i \underline{\gamma}$ for all γ. Then, one could define a distribution over consequences inducing sacrifice s as[13]

$$\gamma^s := \lambda_s \overline{\gamma} + (1 - \lambda_s)\underline{\gamma}.$$

Then, define $\alpha_i^*, \alpha_{-i}^*$ as any two actions such that $\rho_{\alpha_i^*, \alpha_{-i}^*} = \overline{\gamma}$, and α_i^s as any action such that $\rho_{\alpha_i^s, \alpha_{-i}^*} = \gamma^s$. Actions α_i^s leading to absolute sacrifice s by definition satisfy

$$\sum_{a_{-i}} \alpha_{-i}^*(a_{-i}) \sum_x \rho_{\alpha_i^*, a_{-i}}(x) u_i(x) - \sum_{a_{-i}} \alpha_{-i}^*(a_{-i}) \sum_x \rho_{\alpha_i^s, a_{-i}}(x) u_i(x) = s. \tag{2.5}$$

Neither $(\alpha_i^*, \alpha_{-i}^*)$ nor any α_i^s are guaranteed to be unique. For the sake of the following definition, I assume they are.[14]

Definition 2.8. *A ranking $\succsim_i$ is an **Equal Sacrifice Universalisation** preference if it is represented by*

$$\begin{aligned} U_i(\alpha_i^s) = (1 - \kappa) &\sum_{a_i, a_{-i}} \alpha_i^s(a_i) \mu_i(a_{-i}) u_i(\rho_{a_i, a_{-i}}) \\ + \kappa &\sum_{a_i, a_{-i}} \alpha_i^s(a_i) \alpha_{-i}^s(a_{-i}) u_i(\rho_{a_i, a_{-i}}), \end{aligned} \tag{2.6}$$

for some utility function $u_i \colon X \to \mathbb{R}$ and belief $\mu_i \in \Delta(A_{-i})$.

[13]The weight λ_s depends on the normalisation of utility. As an example, if

$$\sum_x \overline{\gamma}(x) u_i(x) = S \text{ and } \sum_x \underline{\gamma}(x) u_i(x) = 0,$$

then $\lambda_s = 1 - \frac{s}{S}$. In fact, by linearity $\sum_x \gamma^s(x) u_i(x) = \lambda_s S$, and I would like that $\sum_x \gamma^s(x) u_i(x) = S - s$, so that

$$\lambda_s S = S - s \;\Rightarrow\; \lambda_s = 1 - \frac{s}{S}.$$

[14]Therefore, I am implicitly considering a restriction of preferences or games.

When choosing α_i^s, the individual evaluates the scenario where his opponent deviates from the action in the profile leading to the highest Bernoulli utility to induce the same sacrifice. *ESU* is a *UP* in which $T_{\alpha_i^*,\alpha_{-i}^*}[\alpha_i^s] = \alpha_{-i}^s$ for all s, so that Equations (2.1) and (2.6) coincide.

The key difference between *ESU* and previous concepts is that universalisation reasoning depends on the game at hand. I illustrate this point in the battle of the sexes. Numbers are again Bernoulli utilities for consequences.

1 \ 2	a_2	a_2'
a_1	$2,1$	$0,0$
a_1'	$0,0$	$1,2$

Table 2.3: Asymmetry.

1 \ 2	a	a'
a	$2,1$	$0,0$
a'	$0,0$	$1,2$

Table 2.4: Same Actions.

1 \ 2	a	a'
a	$0,0$	$2,1$
a'	$1,2$	$0,0$

Table 2.5: Symmetry.

Consider the game in Table 2.3 on the left and assume throughout that $\kappa = 0$ for simplicity. The table represents a standard battle of the sexes, in which player 1 would like to coordinate in the top-left corner, while player 2 would like to coordinate on the bottom-right corner. The greatest achievable payoff of both players is 2, in (a_1, a_2) and (a_1', a_2'). The action α_1^s of player 1 inducing a sacrifice of s solves $2 - 2\alpha_1^s = s$ and hence $\alpha_1^s = \frac{2-s}{2}$. The equivalent for player 2 is $2 - (2 - 2\alpha_2^s) = s$ which implies $\alpha_2^s = \frac{s}{2} = 1 - \alpha_1^s$. The optimum for *ESU* is reached at $s = \frac{2}{3}$ with $\alpha_1^s = \alpha_2^s = \frac{1}{2}$ which, if picked by both players, leads to a common expected payoff of $\frac{3}{2}$.

This simple example allows me to discuss important differences between *ESU* and previous concepts. First, even if we were to relabel the actions (a_2, a_2') to (a, a') for employing *SKE*, as in the table in the middle, one would not exist anyway. The optimal action is not common, as it is a for 1 and a' for 2. Nevertheless, I argue that the problem here is not existence. It is possible to define universalisation from an individual perspective and obtain the profile composed by subjectively optimal actions (a, a'). This is indeed what would happen assuming both players are *HK*. The issue is that it is meaningless to define "the same thing" as "the same action" in this scenario. The relabelling of actions from Table 2.3 to 2.4 is arbitrary as any other, it is not surprising that it does not lead to intuitive results.

As a solution, Roemer (2019, p. 26) suggests to relabel the game as in Table 2.5 on the right, to make it symmetric. Now actions are interpreted as "do the favourite thing" and "do the least favourite thing". The *SKE* of this reformulation of the game is $(\frac{1}{2}, \frac{1}{2})$, i.e., the optimal actions of *ESU*. Not only the optimal profile coincides, but also the set of profiles considered in the universalisation evaluation is identical. The re-labelling of actions from the first to the third table amounts to changing any mixed action α_2^s to $1 - \alpha_1^s$, which leads to $a_2 = a_1'$ and $a_2' = a_1$ and switching columns. This is exactly the *ESU* universalisation reasoning.

Now consider the difference between *ESU* and *Ex-Ante HM*. *Ex-Ante HM* is defined in

an incomplete information expansion of the game in which players do not know whether they will be the row player or the column player. When $\kappa = 1$, it prescribes players to choose the strategy, in this case mapping between identity and action, that ex-ante, before identities are revealed, maximises expected utility over material consequences. The optimal strategies according to such criterion are (a_1, a_2) or (a_1', a_2'). Contrary to what is implemented if both players exhibit *ESU*, these two profiles are Pareto-Efficient. It is already known that *Ex-Ante HM* is related to utilitarian altruism (Laslier, 2022). Hence, it is possible that *ESU* delivers an inefficient allocation in terms of material payoff. By contrast, *Ex-ante HM* is always efficient, but is indifferent to inequality.

The following result establishes that optimal actions under *ESU* are always optimal actions under *HK* in symmetric games and therefore the first is a generalisation of the second. The result holds for any equal sacrifice rule, not only absolute sacrifice, as shown in the proof in Appendix A.2.

Proposition 2.3. *Assume the game is symmetric. Then, if an action is optimal under ESU, it is also optimal under HM.*

The result may be interpreted as a conceptual robustness check. In games where "same action" has meaning, because of symmetry, *ESU* delivers the intuitive counterfactual evaluation of previous concepts. In asymmetric games, the counterfactual depends on the equal sacrifice conception of the individual.

I conclude by addressing possible critiques to *ESU*. First, it relies on interpersonal comparisons of utility, and thus is less parsimonious compared with previous concepts. I acknowledge the issue, but I argue that universalisation always relies on some form of interpersonal comparison and hence the problem is not idiosyncratic to *ESU*. *Ex-ante HM* also relies on the same informational requirement, as it employs the veil of ignorance construct, and thus relies on the same interpersonal comparisons of Harsanyi's utilitarianism. As for the various forms of Kantian Equilibrium, these rely on interpersonal comparisons of actions, as argued by Sher (2020), as actions need to have a cardinal interpretation common to all players. Some form of interpersonal comparisons is therefore needed also in previous conceptions.

The issue is deeper. It is not that universalisation needs some form of interpersonal comparison outside symmetric environments. It always does, but under symmetry, both concepts of "same action" and "same utility" have meaning, so comparisons of actions and utility are easy to deal with. Universalisation becomes problematic without symmetry not because of labels, but because of heterogeneity among players. The implicit suggestion of *Ex-ante HM* is to solve such heterogeneity by aggregating preferences in the utilitarian fashion. MKE, instead, suggests to give actions a cardinal meaning. *ESU* offers a third way.[15]

A second issue is that *ESU* might lead to corner solutions. The problem is related to the previous one. It is possible that utility indexes across players have different scales

[15]Incidentally, *ESU* is reminiscent of Kalai & Smorodinsky (1975) bargaining solution.

and range and this makes it hard for equal sacrifice of utility to be feasible. A partial solution is to perform a proper rescaling of utility.[16] When this is not enough, constrained versions of equal sacrifice, developed by Stovall (2020), can be employed.

2.6 Conclusion

I developed and axiomatically characterised a model to account for non-consequentialist preferences for universalisation. The main behavioural prediction is that independence is satisfied only among actions that are universalised equivalently. I showed that the general model unifies the two most prominent models of universalisation, namely Homo Moralis and Kantian Equilibrium. Lastly, inspired by the equal sacrifice principle, I proposed a novel concept of universalisation that does not rely on the labelling of actions, is equivalent to the previous models under symmetry, and can be defined in asymmetric games. I showed how the results shed light on the conceptual underpinnings of universalisation, guide empirical work, and inform the evaluation of welfare statements. In the last paragraphs, I discuss two points regarding the methodology and implications of this paper.

I am not the first to propose changing the set of consequences to account for apparent paradoxes. Baccelli & Mongin (2021), among others, have criticised this practice, as a redescription of the problem might solve technical but not conceptual issues. They argue that it is more reasonable to capture non-material determinants of utility in the evaluation of consequences, without affecting their definition. In this paper, I adhere to this principle. I do not need to alter the set of consequences by including other features in the game. The key is to introduce a link between actions and consequences without changing these two primitives. As my introductory example shows, universalisation cannot be rationalised without assuming that the individual cares about something unrelated to the material consequences of the game. An expansion of the consequence domain is necessary. A second possibility is to include the chosen action in the description of the consequence. It would then be easy to formalise a trade-off between selecting the preferred action and maximising material payoff. This has been done in empirical work on moral preferences, notably by Cappelen et al. (2007). By contrast, my universalisation theory does not rely on assuming that an action is optimal but explains why, i.e., because it induces the preferred universalised consequence.

The final point concerns the nature of preferences for universalisation. I have denoted these as non-consequentialist, and the literature refers to them as moral. Nevertheless, I show that universalisation satisfies consequentialism under an appropriate redefinition of consequences which consider what is induced by the universalised action. What, then, is the difference between universalisation and consequentialist pro-social attitudes? John

[16] Interpersonal comparisons of utility are widely discussed in social choice theory. Binmore (1994, Ch. 4) and Sen (2017, Ch. 7) offer critical overviews of approaches to perform this exercise.

Broome argues, in Bradley & Fleurbaey (2021, p. 120), that "*a very specific version of consequentialism is a view I call distribution (it is often called welfarism), which is the view that the goodness of an act is determined by the goodness of the distribution of well-being that results from it*". Universalisation is, strictly speaking, not a welfarist attitude, as the optimal action is unrelated to the distribution of well-being it induces.

References

Alger, I., & Laslier, J.-F. (2022). Homo moralis goes to the voting booth: Coordination and information aggregation. *Journal of Theoretical Politics*, *34*(2), 280–312. 8

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302. 4, 6, 7, 8, 13, 16

Alger, I., & Weibull, J. W. (2016). Evolution and kantian morality. *Games and Economic Behavior*, *98*, 56–67. 7

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory*, *185*, 104951. 7

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199–205. 5, 7, 10, 20

Baccelli, J., & Mongin, P. (2021). Can redescriptions of outcomes salvage the axioms of decision theory? *Philosophical Studies*, 1–28. 23

Battigalli, P., Catonini, E., & De Vito, N. (2023). *Game theory: Analysis of strategic thinking.* 14

Battigalli, P., Cerreia-Vioglio, S., Maccheroni, F., & Marinacci, M. (2017). Mixed extensions of decision problems under uncertainty. *Economic Theory*, *63*(4), 827–866. 7

Binmore, K. (1994). *Game Theory and the Social Contract: Playing fair*. MIT Press. 23

Bonanno, G. (2018). *Game theory* (2nd ed.). Kindle Direct Publishing. 9

Bradley, R., & Fleurbaey, M. (2021). John Broome. *Conversations on Social Choice and Welfare Theory-Vol. 1*, 115–127. 24

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of public economics*, *87*(9-10), 1967–1983. 8

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*(3), 818–827. 23

Chen, D. L., & Schonger, M. (2022). Social preferences or sacred values? Theory and evidence of deontological motivations. *Science Advances*, *8*(19). 7

Daley, B., & Sadowski, P. (2017). Magical thinking: A representation result. *Theoretical Economics*, *12*(2), 909–956. 7

De Donder, P., Llavador, H., Penczynski, S., Roemer, J. E., & Vélez, R. (2023). A game-theoretic analysis of childhood vaccination behavior: Nash versus Kant. *Working Paper*. 8

Dekel, E., & Siniscalchi, M. (2015). Epistemic game theory. In *Handbook of Game Theory with Economic Applications* (Vol. 4, pp. 619–702). Elsevier. 15

Dewatripont, M., & Tirole, J. (2024). The Morality of Markets. *Journal of Political Economy*, 000–000. 8

Dierks, K., Alger, I., & Laslier, J.-F. (2024). Does universalization ethics justify participation in large elections? *TSE Working Paper*. 8

Dizarlar, A., & Karagözoğlu, E. (2023). Kantian equilibria of a class of Nash bargaining games. *Journal of Public Economic Theory*, *25*(4), 867–891. 8

Ellingsen, T., & Mohlin, E. (2024). A model of social duties. *Working Paper*. 8

Epstein, L. G., & Schneider, M. (2003). Recursive multiple-priors. *Journal of Economic Theory*, *113*(1), 1–31. 13

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford University Press. 6

Fleurbaey, M. (2019). Economic theories of justice. *Annual Review of Economics*, *11*, 665–684. 5

Fleurbaey, M., Kanbur, R., & Snower, D. J. (2024). An Analysis of Moral Motives in Economic and Social Decisions. *Working Paper*. 8

Ghirardato, P. (2002). Revisiting Savage in a conditional world. *Economic theory*, *20*, 83–92. 13

Gilboa, I., & Schmeidler, D. (2003). A derivation of expected utility maximization in the context of a game. *Games and Economic Behavior*, *44*(1), 172–182. 7

Grillo, A. (2022). Ethical Voting in Heterogenous Groups. *Working Paper*. 8

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, *61*(5), 434–435. 22

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, *63*(4), 309–321. 16

Juan-Bartroli, P. (2024). On Injunctive Norms: Theory and Experiment. *TSE Working Paper, n. 24-1515*. 8

Juan-Bartroli, P., & Karagözoğlu, E. (2024). Moral preferences in bargaining. *Economic Theory*, 1–24. 8

Kalai, E., & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica: Journal of the Econometric Society*, 513–518. 22

Laffont, J.-J. (1975). Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, *42*(168), 430–437. 6, 14

Laslier, J.-F. (2022). Universalization and altruism. *Social Choice and Welfare*, 1–16. 8, 22

Laslier, J.-F., Fleurbaey, M., Gravel, N., & Trannoy, A. (1998). Freedom in Economics. *New Perspectives in Normative Analysis*. 15

Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences*, *117*(42), 26158–26169. 4

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley. 5, 7

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior & Organization*, *173*, 1–25. 4

Mill, J. S. (1885). *Principles of political economy*. D. Appleton. 6, 18, 19

Perea, A. (2012). *Epistemic game theory: Reasoning and choice*. Cambridge University Press. 15

Ponthiere, G. (2023). Epictetusian rationality. *Economic Theory*, 1–44. 8

Ponthiere, G. (2024). Stoicism and the Tragedy of the Commons. *GLO Discussion Paper*. 8

Rawls, J. (1971). *A theory of justice*. Harvard university press. 16

Roemer, J. E. (2010). Kantian equilibrium. *Scandinavian Journal of Economics*, *112*(1), 1–24. 4, 7

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, *127*, 45–57. 7, 8

Roemer, J. E. (2019). *How we cooperate*. Yale University Press. 4, 5, 6, 7, 9, 13, 14, 15, 16, 17, 21

Rohde, K. I. (2010). A preference foundation for Fehr and Schmidt's model of inequity aversion. *Social Choice and Welfare*, *34*(4), 537–547. 13

Sarkisian, R. (2017). Team Incentives under Moral and Altruistic Preferences: Which Team to Choose? *Games*, *8*(3), 37. 8

Sarkisian, R. (2021a). Optimal Incentives Schemes under Homo Moralis Preferences. *Games*, *12*(1), 28. 8

Sarkisian, R. (2021b). Screening Teams of Moral and Altruistic Agents. *Games*, *12*(4), 77. 8

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications. 5

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica*, *40*(159), 241–259. 5

Sen, A. (2017). *Collective choice and social welfare*. Harvard University Press. 23

Sher, I. (2020). Normative aspects of kantian equilibrium. *Erasmus Journal for Philosophy and Economics*, *13*(2), 43–84. 22

Shi, Y. (2024). Endogenous Social Minimum. *Working Paper*. 8

Sobrado, E. M. (2022). Taxing moral agents. *CESifo working paper series 9867*. 8

Stovall, J. E. (2020). Equal sacrifice taxation. *Games and Economic Behavior*, *121*, 55–75. 23

van Leeuwen, B., & Alger, I. (2024). Estimating social preferences and Kantian morality in strategic interactions. *Journal of Political Economy Microeconomics*. 4, 6, 16

Young, H. P. (1988). Distributive justice in taxation. *Journal of Economic Theory*, *44*(2), 321–335. 6, 18, 20

Chapter 3

Meritocracy as an End and as a Means

Abstract

I introduce a framework for studying different interpretations of meritocracy and testing whether individuals adhere to them. Each meritocracy has two components: a merit criterion, determining when one individual is more meritorious than another, and a reward criterion for each individual, determining when one outcome constitutes a better reward than another for that individual. An allocation is meritocratic if more meritorious individuals are better rewarded. I distinguish between two conceptions of meritocracy. Meritocracy as an end holds it intrinsically valuable that individuals are rewarded according to their merit. Meritocracy as a means views rewarding merit as instrumental in achieving desirable outcomes according to other standards, such as efficiency. I show that these two conceptions are equivalent: each instance of meritocracy as a means can be associated with a corresponding meritocracy as an end. Finally, I examine two specific meritocracies present in the literature. Pareto meritocracy defines an action as more meritorious if it leads to a Pareto improvement in welfare, whereas proportional meritocracy requires that an individual's consumption be proportional to the amount of labour he provides. By observing whether allocation choices of impartial spectators align with specific merit criteria, one can test whether individuals adhere to these meritocracies.

3.1 Introduction

Meritocracy has recently attracted considerable attention in economics and political philosophy (Markovits, 2019; Mulligan, 2018; Sandel, 2020; Tirole, 2022). However, as Sen (2000, p. 5) observes, "*the idea of meritocracy may have many virtues, but clarity is not one of them.*" This paper attempts to bring some clarity. I introduce a framework that allows me to explicitly distinguish between multiple interpretations of meritocracy across different literatures and clarify their relationships. To showcase the use of the framework, I characterise, through assumptions on primitive elements, a particular form of meritocracy referenced in behavioural and experimental economics: proportional meritocracy,

according to which an individual's consumption should be proportional to his provided labour.

The setting is as follows. Individuals have preferences over a set of outcomes. A social choice function maps each preference profile to an outcome. A social choice function is meritocratic if it assigns better outcomes to more meritorious individuals. To express this condition, I consider two criteria: first, the conditions under which an individual or an action is deemed more meritorious; and second, when an outcome qualifies as a better reward for merit. Accordingly, the framework relies on two fundamental primitives: the **merit criterion** and the **reward criterion**. The merit criterion is a binary relation over preferences determining when a preference is more meritorious than another. By interpreting preferences as representations of individual choices, the merit criterion identifies who is more meritorious based on their behavioural attitudes. For example, one might be deemed more meritorious if he requires less consumption as compensation for performing more productive work. The reward criterion for an individual is a binary relation over outcomes determining when an outcome is a better reward than another for that individual. For instance, an individual may consider an outcome a better reward if it is preferred to another. A social choice function is meritocratic if, for each individual, having more meritorious preferences according to the merit criterion leads to better outcomes according to the reward criterion.

I refer to the notion of meritocracy represented by a meritocratic social choice function as **meritocracy as an end**. One might desire a social choice function to be meritocratic *per se*, independently of any other properties it has. That is, it may be considered intrinsically valuable that an individual is better rewarded if he is more meritorious. This view aligns with desert-based theories of justice in the philosophical literature: "*It is a good thing, morally speaking, if people are getting what they deserve*" (Kagan, 2014, p. 5).

I also consider a second notion, **meritocracy as a means**, in which meritocracy is a tool to achieve outcomes desirable according to other criteria. For example, meritocracy as a means could be employed to induce efficient outcomes: "*the creed of meritocracy: the belief that in the rat race of life rewards should go to the best performers, thereby unleashing society's full potential*" (Morgan et al., 2022, p. 1). To represent meritocracy as a means, I introduce mechanisms, which consist of a collection of action sets, one for each individual, and an outcome function, mapping each action profile to an outcome. Within a mechanism, a merit criterion is a binary relation over action profiles, determining when an individual's action, given the actions of others, is considered more meritorious. The reward criterion for an individual remains a binary relation over outcomes determining when an outcome is a better reward than another for that individual. A mechanism is meritocratic if, for each individual, more meritorious actions lead to better outcomes under the respective criteria.

I derive three results. The first result, Proposition 3.1, details the relationship between meritocracy as an end and meritocracy as a means. I characterise the conditions

under which a mechanism implements a meritocratic social choice function. This requires that the mechanism is meritocratic relative to an action-based merit criterion that agrees with the merit criterion over preferences, where agreement means that more meritorious preferences induce more meritorious actions in the mechanism. Therefore, each conception of meritocracy as a means has a meritocracy as an end counterpart. This characterization yields two main implications. First, if one commits to a merit criterion over preferences, the merit criterion over actions of a mechanism is meaningless by itself, as its interpretation relies on the mechanism's outcome function. Common merit criteria in the literature and public discourse are binary relations of the type "exerting more effort is more meritorious". However, the result indicates that such criteria are meaningless unless explicitly linked to an outcome function clarifying the implications of effort. Second, if one commits to a merit criterion over actions, independent of the outcome function, he must also accept that such ranking is independent from the outcome the meritocratic mechanism induces. However, this viewpoint is in contrast with meritocracy as a means, as the outcome of the mechanism is thus irrelevant. In a nutshell, one is forced to abandon the idea of meritocracy as a means to stick to a merit criterion over actions as a primitive object.

I showcase the use of the framework and the equivalence between meritocracy as an end and as a means by analysing assumptions on merit and reward criteria that lead to two distinct forms of meritocracy. I start by studying a conception of meritocracy in which an individual's action is more meritorious than another if it leads to a Pareto improvement. The second result, Proposition 3.2, shows that such merit criterion, if not complemented with other assumptions, is vacuous, in the sense that it ranks as more meritorious individuals who prefer a Pareto improvement to a Pareto worsening, but nothing more. Since such Pareto ordering is weak, I proceed to study assumptions on the merit criterion leading to a stronger ordering of individuals.

The third and main result of this paper, Proposition 3.3, is a characterisation of a widely referenced version of meritocracy within a more structured environment, a private goods economy. Individuals provide a labour input to produce a consumption output. They have preferences over pairs of labour and consumption allocations. Several settings in which meritocracy is explored in the literature are particular cases of such economy. I characterise a **proportional meritocracy**, where each individual's consumption is proportional to his labour input. Proportional meritocracy is characterised by two conditions on the merit criterion and one on the reward criterion. First, the merit criterion is monotonic in the labour input, if one works more, he is more meritorious. Second, the merit criterion is scale-invariant, if the labour input of everyone is multiplied by a constant, the ordering of labour input profiles according to the merit criterion does not change. Third, the reward criterion is welfarist, the reward for merit is individual welfare. I then show that the merit criteria over preferences agreeing with the merit criteria of actions in a proportional meritocracy ranks preferences according to the marginal rate of substitution between labour and consumption. I suggest that such criterion over preferences is

more fundamental than one over the labour input, as it can be formulated independently from any specific mechanism. The characterisation of proportional meritocracy can guide empirical studies. Observing allocation choices of an impartial spectator, one could test whether individuals adhere to proportional meritocracy by checking whether they abide by the requirements on the merit criterion.[1] I further discuss how the present paper paves the way for further empirical evidence on meritocracy in Section 3.4.1. A brief literature review follows.

Related Literature. This paper contributes to the literature on responsibility sensitive social choice theory, surveyed by Fleurbaey (2008) and Roemer & Trannoy (2016). The literature develops allocation criteria that are sensitive to factors for which individuals are held responsible, with meritocracy being an instance.[2] Equality of opportunity is the main focus of this literature. An allocation rule satisfies equality of opportunity if it neutralises the effects of circumstances beyond individuals' control on their outcomes. One example is allowing students access to education regardless of their parents' income. A key element of this literature is the distinction between compensation for unequal circumstances and reward for choices under control of the individual. A relevant insight is that a compensation criterion for unequal circumstances does not specify a rewarding rule for free actions, these two are distinct concepts. In general, fair compensation is logically independent from fair reward (Moulin, 2004). Both theoretical and empirical work studying meritocracy often neglect such distinction. Compensation for unequal circumstances is frequently considered a necessary condition for meritocracy, but if the latter is interpreted as a reward criterion for choices, there are no links between the two concepts. This results in meritocracy being conflated with equality of opportunity. In this paper, I propose to view meritocracy as an allocation rule that depends on a merit criterion, regardless of individuals' unequal circumstances. Characterising meritocracy separately from equality of opportunity allows for studying the scope of their intersection. Section 3.5 further discussed the relationship between meritocracy and equality of opportunity.

The need for a framework to explicitly discuss meritocracy is revealed by the different meanings different authors give to the same word. I briefly review the theoretical and experimental literature on meritocracy to showcase the lack of common language and conceptual overlap with equal opportunity. I also discuss the relation between past literature and the present work in the body of the text. Moisson & Tirole (2024) study co-optation into an organisation. Candidates have a quality and a personal trait, such as race, gender, or taste. A higher quality trait benefits all members of the organisation. A specific personal trait benefits members with the same trait. The organisation is meritocratic if it selects candidates based on quality traits. Therefore, in this paper, meritocracy is rewarding a trait benefiting more individuals. I study a general version of such assumption in Section 3.4.1. The only distinction between the two traits is the preferences individuals have for them. The underlying idea is that only quality should

[1] See e.g. Cappelen et al. (2024).

[2] Carroll (2025) discusses the distinction between responsibility sensitive criteria and standard welfarism.

matter because it is the trait that impacts the firm's output, but equality of opportunity is necessary for individuals with the same productivity not to be discriminated against due to their personal traits. However, such distinction between the two traits is outside the model, as candidates do not choose them, and there is no reason to distinguish them except for employers' preferences.

In Morgan et al. (2022), individuals exert effort in a contest and are rewarded according to a ranking of their performance. There is no meritocracy when the ranking is random, while there is full meritocracy when the ranking precisely reflects the amount of effort exerted. Individuals exert effort to win the contest and obtain a prize. The paper focuses on the relationship between the precision of the ranking and the effort exerted. Interpreting meritocracy as a more or less precise ranking of effort is common, but the reason effort should be rewarded and to which extent is not addressed. Interpreting meritocracy as the precision of effort-based rankings effectively collapses the concept into that of equal opportunity, as randomness can be interpreted as a circumstance outside the individual's control. Moisson (2024) adheres to a similar definition: in his paper, meritocracy is the relative weight given to effort, ability, and unequal circumstances in determining outcomes.

Similar understandings of meritocracy are also present in the experimental literature. In Cappelen et al. (2023) and related work by the same authors, subjects in an experiment are divided into workers and spectators.[3] Workers must complete a real effort task, and spectators have to distribute an amount of money among two workers. The experimental variation is the degree to which the outcome of the effort task is due to the workers' effort or randomness introduced by the experimenter. The idea in Cappelen et al. (2023) is thus similar to Morgan et al. (2022). Meritocracy is defined as rewarding workers increasingly in their effort. Andre (2024) studies how circumstances of which individuals are not in control of shape how meritorious others consider them. One individual has to reward others after observing their choices. He shows that rewards are insensitive to unequal circumstances which shape choices. Both Andre (2024) and Cappelen et al. (2024) more or less explicitly refer to proportional meritocracy, which I discuss in Section 3.4.2. Moreover, they find empirical evidence that individuals adhere to such a particular form of meritocracy.

This brief review shows that papers mentioning meritocracy interpret it as a criterion for rewarding effort under equal opportunity, adapted to different settings. Surprisingly, in these papers there is often no reference to the relevant literature in social choice theory on the topic. I will further discuss differences between previous papers and this one in the body of the text.

Lastly, this paper draws from the philosophical literature studying desert-based theories of justice, according to which the allocation of goods in society should reflect individuals' deservingness.[4]

[3] Cappelen et al. (2024, 2020, 2022).

[4] See Arneson (2007); Kagan (2014); Mulligan (2023) for recent developments relevant to the current

3.2 Framework

Let I be a finite set of n individuals. Each individual i has preferences $R_i \in \mathcal{R}$ over a set X of outcomes. A social choice function $f\colon \mathcal{R}^n \to X$ assigns an outcome to each profile of individual preferences. I introduce two primitives that allow me to define a meritocratic social choice function. First, a binary relation $\geq_{\mathcal{R}}$ over the set of preferences, determining whether a preference is more meritorious than another, the **merit criterion**. For each individual i, preference R_i over outcomes is more meritorious than R_i' if $R_i \geq_{\mathcal{R}} R_i'$. The relation $\geq_{\mathcal{R}}$ over preferences is common for all individuals. Second, a collection of binary relations $\geq_i$ over outcomes, one for each individual i, determining whether an outcome is a better reward for individual i than another, the **reward criterion**. An outcome x is a better reward for individual i than x' if $x \geq_i x'$. In general, the reward criterion does not coincide with preferences and need not be identical across individuals.

Example 3.1. (**Private Goods Economy**) Several papers studying meritocracy (Andre, 2024; Cappelen et al., 2024, 2023, 2022; Fleurbaey, 2008) consider particular cases of the following setting. There is an economy with two private goods, labour serves as an input in the production of a consumable output. The production of y units of consumption requires $\ell = c(y)$ units of labour, where the cost function c is strictly increasing and invertible. Each individual i has preferences R_i over his labour and consumption allocation represented by the utility function $u_i(\ell_i, y_i)$, which is strictly decreasing in labour ℓ_i, strictly increasing in consumption y_i and differentiable in both its arguments. An allocation (ℓ, y) is feasible if the total labour input suffices to produce the total consumption $\sum_i \ell_i = c(\sum_i y_i)$. A social choice function maps each preference profile to a feasible allocation.[5]

Denote the marginal rate of substitution between labour and consumption for individual i in each point by $\mathrm{MRS}_{R_i}(\ell_i, y_i)$, measuring how many units of consumption the individual needs to be compensated for one additional unit of labour. A possible merit criterion states that a preference is more meritorious if, for each level of labour and consumption, it requires less consumption as compensation for one more unit of labour, that is

$$R_i \geq_{\mathcal{R}} R_i' \quad \Longleftrightarrow \quad \mathrm{MRS}_{R_i}(\ell_i, y_i) \leq \mathrm{MRS}_{R_i'}(\ell_i, y_i) \text{ for each } (\ell_i, y_i)\,.$$

One plausible reward criterion holds that an individual is better rewarded if he receives a preferred allocation, and therefore that $\geq_i = R_i$ for each individual i. I explore the implications of these merit and reward criteria in Section 3.4.2.

With these primitives at hand, I define a meritocratic social choice function.

---

paper.

[5] More structure would allow me to consider different productivities. Say that $\ell_i = b_i e_i$ measures unit of effective labour, where b_i is productivity. Then preferences are on pairs of $(\ell_i/b_i, y_i) = (e_i, y_i)$.

Definition 3.1. *A social choice function f is* ***meritocratic*** *if, for each individual i, for all preferences R_i, R_i' and for each preference profile R_{-i},*

$$R_i \geq_{\mathcal{R}} R_i' \implies f(R_i, R_{-i}) \geq_i f(R_i', R_{-i}) .$$

A social choice function is meritocratic if it assigns a better reward, according to the reward criterion, to a more meritorious preference, according to the merit criterion. Therefore, different notions of meritocracy correspond to different pairs of criteria. In the words of Sen (2000): "*the notion of merit is fundamentally derivative, and thus cannot but be qualified and contingent.*" Strictly speaking, one should say "meritocratic with respect to these merit and reward criteria", but I avoid the qualification when no confusion should arise. Arguably, such modelling choice is convenient as it allows one to trace the underlying sources of disagreement about meritocracy. As an example, some critique meritocracy because it does not take into account unequal circumstances outside individuals' control. I argue that such critique could be phrased as disagreement about what the merit criterion $\geq_{\mathcal{R}}$ should trace. One might want it to be independent of features individuals cannot control. Definition 3.1 represents **meritocracy as an end**. One might desire the social choice function to be meritocratic regardless of other properties it has: "*a just distribution is merit-based*" (Mulligan, 2023, p. 2).

This notion of meritocracy as an end aligns with desert-based theories of justice as studied in Kagan (2014). The difference is that, in the present setting, I only consider qualitative notions of merit and reward. Kagan (2014) instead assumes that these two are measurable quantities, and is therefore able to describe how far an allocation is from being meritocratic. Definition 3.1 is instead either satisfied or not. In Section 3.4.2, where I introduce proportional meritocracy, I show how assumptions on the merit criterion allow me to derive a quantitative representation of individual merit.

However, meritocracy is sometimes conceived as a mechanism for inducing outcomes satisfying other desirable properties, such as Pareto efficient outcomes. To represent such instrumental understanding of meritocracy, I introduce mechanisms. A mechanism $M = ((A_i)_{i \in I}, g)$ is a collection of action sets, one for each individual i, and an outcome function $g \colon A \to X$ mapping each action profile to an outcome. To define a meritocratic mechanism, I introduce a merit criterion over action profiles of the mechanism. Consider a collection of binary relations $\geq_{A_i}$, one for each individual i, such that, fixing a profile of actions a_{-i} of individuals other than i, action a_i of individual i is more meritorious than a_i' if $(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i})$. The binary relation $\geq_{A_i}$ is distinct from $\geq_{\mathcal{R}}$, as the first ranks action profiles in a mechanism, while the second ranks preferences. I discuss a notion of agreement between the two merit criteria in Section 3.3. The reward criterion is again defined as a collection of binary relations $\geq_i$ over outcomes, one for each individual i.

Definition 3.2. *A mechanism M is* ***meritocratic*** *if, for each individual i, for all actions a_i, a_i' and for each action profile a_{-i},*

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \implies g(a_i, a_{-i}) \geq_i g(a_i', a_{-i}).$$

A mechanism is meritocratic if it assigns a better reward as a consequence of an individual choosing a more meritorious action. Definition 3.2 differs from Definition 3.1 of meritocratic social choice functions, which assigned a better reward for individuals who are more meritorious according to their preferences. Definition 3.2 represents **meritocracy as a means**. One might ask whether a meritocratic mechanism implements a social choice function satisfying desirable properties. As an example, it is often explored in the literature, and discussed in popular debate, whether meritocratic mechanisms implement efficient outcomes. For instance, Moisson (2024) examines various definitions of "merit" based on different weightings assigned to talent, effort, and head start in a specific game. The author investigates which weight distributions achieve desirable outcomes, such as maximizing efficiency in preference satisfaction.

In the next section, I study the relationship between meritocracy as an end and meritocracy as a means, i.e., meritocratic social choice functions and mechanisms.

3.3 When Means and End Coincide

In this section, I characterise the conditions for a mechanism to implement a meritocratic social choice function. I show that the mechanism must be meritocratic according to a merit criterion on actions of the mechanism that agrees with the merit criterion on individuals' preferences. I say that $\geq_{\mathcal{R}}$ and $\geq_{A_i}$ agree if, whenever, for a fixed profile a_{-i}, individual i's actions a_i, a_i' produce outcomes with opposite rankings under two distinct preferences R_i and R_i', then a_i is deemed more meritorious than a_i' according to $\geq_{A_i}$ if and only if R_i is deemed more meritorious than R_i' according to $\geq_{\mathcal{R}}$.

Definition 3.3. *Fix a mechanism M. Two merit criteria $\geq_{\mathcal{R}}$ and $\geq_{A_i}$ **agree** if,*

- *for all actions a_i, a_i' and action profiles a_{-i},*
- *for all preferences R_i, R_i' such that $g(a_i, a_{-i}) R_i g(a_i', a_{-i})$ and $g(a_i', a_{-i}) R_i' g(a_i, a_{-i})$,*

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \iff R_i \geq_{\mathcal{R}} R_i'\,.$$

Crucially, agreement between the two merit criteria depends on the outcome function g. That is, fixing action sets and a merit criterion over preferences, two different outcome functions induce two different criteria of merit over actions agreeing with the original criterion over preferences. In fact, if one wants to adhere to a merit criterion that depends on preferences, it is meaningless to assign a merit criterion to actions, as these might be ranked differently by preferences under different outcome functions.

Instead, if one wants to stick to an understanding of merit that is related to a particular interpretation of actions, such as effort or labour, regardless of the outcome function, one

should also accept that such ranking is independent from the outcome the meritocratic mechanism induces. However, this is in contrast with the view of meritocracy as a means, as the outcome of the mechanism is irrelevant for defining a merit criterion over actions. Therefore, supporting both meritocracy as a means and an action-based merit criterion leads to logical inconsistency. In fact, I suggest that it would be more transparent to consider a merit criterion over preferences as a primitive object, and then derive the merit criterion over actions agreeing with it. Moreover, as I show in my next result, any meritocratic mechanism has a meritocratic social choice function counterpart. Therefore, adherence to a meritocratic mechanisms is equivalent to adherence to a meritocratic social choice function.

I now define the notion of implementation I consider. With a slight abuse of notation, I denote a strategy of individual i in a mechanism with $a_i \colon \mathcal{R} \to A_i$ mapping each preference to an action. I say that a mechanism M implements a social choice function f if, for each preference profile R, each Nash equilibrium of the game induced by M results in the outcome $f(R)$.[6]

Definition 3.4. *A mechanism M **implements** a social choice function f if, for each preference profile R, for each Nash equilibrium strategy profile a^*, it holds that $g(a^*(R)) = f(R)$.*

I now state my first result, a characterisation of the mechanisms implementing meritocratic social choice functions.[7] In the statement of the proposition, I do not mention agreement between the two reward criteria, which are assumed to coincide.

Proposition 3.1. *A mechanism implements a meritocratic social choice function if and only if it is meritocratic with respect to merit criteria $\geq_{A_i}$ which, for each i, agree with $\geq_{\mathcal{R}}$.*

Proposition 3.1 shows that, whenever a meritocratic social choice function is implemented by a mechanism, such mechanism must be meritocratic with respect to the same criteria. The result remarks that, once one is committed to meritocracy as an end, the merit criterion on actions depends on the outcome function of the mechanism, and therefore ranking actions by merit independent of the outcome function is unjustified. Such a discussion raises doubts about common criteria of merit of the type "exerting more effort is more meritorious." Conversely, every meritocratic mechanism implements a meritocratic social welfare function according to the same criteria. Therefore, every meritocratic mechanism has a social choice function counterpart.

This equivalence is closely related to the "control principle" (Arneson, 2007; Fleurbaey, 2008), which asserts that, under desert-based theories of justice, individuals should be held accountable only for what lies within their power to control. This principle is elusive

[6] See, e.g., Mas-Colell et al. (1995, p. 913).
[7] All proofs are in Appendix B.1.

to capture within the economic revealed preference framework. If an action causally follows from preferences, it seems odd to hold an individual responsible for them. However, within a mechanism, the action an individual ultimately chooses is shaped by the mechanism itself, via the outcome function. This dependence highlights the arbitrariness of assigning merit to actions, as any action may be part of an equilibrium in a suitable mechanism, thereby violating the control principle. I therefore argue that considering merit criteria over preferences, rather than actions, is more appropriate, despite the criticisms that have been raised against such practice.[8]

In the next section, I examine two specific notions of meritocracy that have been employed in the literature. These notions, often implicitly, arise within the context of mechanisms. I use Proposition 3.1 to identify their corresponding meritocratic social choice functions.

3.4 Two Versions of Meritocracy

This section explores commonly assumed properties of the merit and reward criteria that give rise to distinct meritocracies. In particular, I study what I call a Pareto meritocracy and (labour) proportional meritocracy. In the first, an action is considered more meritorious if it induces a Pareto improvement. In the second, defined in the private goods economy of Example 3.1, the more an individual works, the more meritorious he is.

3.4.1 Pareto Meritocracy

Here, I study a merit criterion according to which an action in a mechanism is more meritorious than another if it induces a Pareto improvement in welfare. The definition follows.

Definition 3.5. *A merit criterion $\geq_{A_i}$ satisfies **Pareto merit** if, for all actions a_i, a_i' and for each action profile a_{-i},*

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \iff g(a_i, a_{-i}) R_j g(a_i', a_{-i}) \text{ for all } j\,,$$

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \iff g(a_i, a_{-i}) P_j g(a_i', a_{-i}) \text{ for some } j\,.$$

In words, fixing an opponents' action profile a_{-i}, an action a_i scores higher in the merit criterion than action a_i' if the outcome resulting from a_i is weakly preferred to the outcome resulting from a_i' by all individuals, and it is strictly preferred by at least one individual. This assumption is frequently invoked in the economic literature and popular debate. As an example, Moisson & Tirole (2024) refer to the employment of a candidate with a talent trait that benefits all members of an organisation as meritocratic. In the

[8] Fleurbaey (2008, ch. 10) offers a comprehensive discussion of this point.

philosophical literature, it is argued that a meritocratic governance should be in the hands of capable individuals, able to make people better-off (Mulligan, 2023).

In what follows, I show that the Pareto merit criterion is vacuous in that it only identifies individuals who prefer Pareto improvements as more meritorious—an uninformative condition in standard settings.

Proposition 3.2. *Assume that for each preference R_i there is a unique maximal element $x^*(R_i)$. Fix a mechanism M. Merit criteria over preferences $\geq_{\mathcal{R}}$ agreeing with a merit criterion over actions $\geq_{A_i}$ satisfying Pareto Merit must satisfy the following:*

$$R_i \geq_{\mathcal{R}} R_i' \iff x^*(R_i) R_j x^*(R_i') \text{ for all } j\,,$$

$$R_i \geq_{\mathcal{R}} R_i' \iff x^*(R_i) P_j x^*(R_i') \text{ for some } j\,.$$

To evaluate Proposition 3.2, consider the implications of the Pareto Merit criterion in the private economy of Example 3.1. In that setting, each individual prefers a Pareto improvement to a Pareto worsening in that setting. Therefore, the Pareto Merit criterion does not rank any individual as more meritorious than another. In other words, all individuals are equally meritorious according to Pareto Merit. If not supplemented with other assumptions, the Pareto Merit criterion does not provide any information about the relative merits of individuals.

It is often argued in philosophy that merit is a contextual phenomenon, suggesting that more structured settings are necessary to meaningfully define it. In the next subsection, I consider a second meritocracy, defined in the more structured setting of Example 3.1, which yields more substantive implications.

3.4.2 Proportional Meritocracy

In this section, I study a meritocratic mechanism in the setting of Example 3.1. Each individual i chooses his labour supply ℓ_i, and the outcome function g maps each labour input profile to a feasible consumption profile. In this mechanism, a merit criterion is a binary relation over profiles of labour inputs, while each reward criterion is a binary relation over pairs of labour input and consumption. I study conditions on merit and reward criteria that, under the assumption that the mechanism is meritocratic, uniquely characterise a proportional outcome function, where each individual's consumption is proportional to the labour input he provides.

Definition 3.6. *A mechanism is a **Proportional Meritocracy** if, for each individual i, labour input ℓ_i, ℓ_i' and labour input profile ℓ_{-i},*[9]

[9]With a slight abuse of notation, I do not include the allocation of labour inputs g induces, as each individual is always assigned his chosen ℓ_i.

$$g(\ell_i, \ell_{-i}) = \alpha_i(\ell) c^{-1}\left(\sum_j \ell_j\right),$$

where the numbers $\alpha_i(\ell)$ satisfy the following conditions:

1. *sum to a unit $\sum_i \alpha_i(\ell) = 1$;*
2. *are monotonic in labour inputs $\ell_i \geq \ell_i' \implies \alpha_i(\ell_i, \ell_{-i}) \geq \alpha_i(\ell_i', \ell_{-i})$;*
3. *are homogeneous of degree zero, $\alpha_i(\lambda\ell_1, \ldots, \lambda\ell_n) = \alpha_i(\ell_1, \ldots, \ell_n)$ for any $\lambda > 0$.*

Allocation rules similar to Proportional Meritocracy have been characterised in various settings, but I follow a slightly different conceptual path.[10] In a proportional meritocracy, the shares α_i are interpreted as quantitative measures of merit of individual i. The individual consumes a proportion of the total output, and such proportion coincides with his merit, as measured by α_i. There is a relationship between such interpretation of the allocation rule and a claim problem, where proportional allocations are typically defined. The distinction is that, in claim problems, claims are exogenous, while here the labour input is chosen by the individual once the mechanism is specified. Therefore, a proportional meritocracy might be viewed as a novel rationale for implementing proportional allocation rules, one related to desert-based views of justice.

A proportional conception of meritocracy traces back to Aristotle, who proposed that the ratio of merits and rewards should be equal among individuals (Mulligan, 2023). Taken literally, Aristotle's condition characterises proportional meritocracies in the current setting. For each individual i, his merit is α_i, while his reward is his consumption level

$$\alpha_i c^{-1}\left(\sum_j \ell_j\right),$$

and the ratio between these two quantities is

$$c^{-1}\left(\sum_j \ell_j\right),$$

equal among individuals. Therefore, Aristotle's condition is satisfied in a Proportional Meritocracy.

I now introduce the assumptions characterising Proportional Meritocracy. I start by restricting attention to merit criteria according to which a higher labour input is more meritorious, fixing the labour input profile of all other individuals.

Definition 3.7. *A merit criterion $\geq_{A_i}$ satisfies **Conditional Labour Monotonicity** if, for each individual i, labour input ℓ_i, ℓ_i' and labour input profile ℓ_{-i},*

[10] See the monograph on claim problems by Thomson (2019).

$$\ell_i \geq \ell_i' \implies (\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i}) .$$

Conditional Labour Monotonicity states that, fixing a profile of others' labour input ℓ_{-i}, a labour input ℓ_i is more meritorious than ℓ_i' if it is bigger. Such definition relies on the structured setting, which allows to measure individual actions. Notice that Conditional Labour Monotonicity does not state that if $\ell_i \geq \ell_i'$ then the first is more meritorious, which would be a stronger requirement, as captured in the following definition.

Definition 3.8. *A merit criterion $\geq_{A_i}$ satisfies **Labour Monotonicity** if, for each individual i, labour input ℓ_i, ℓ_i' and labour input profile ℓ_{-i},*

$$\ell_i \geq \ell_i' \iff (\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i}) .$$

Labour Monotonicity states that a higher labour input is more meritorious, regardless of what others' do. Under Conditional Labour Monotonicity, how an individual's labour input scores in merit is conditional on what others' do, which is not the case under Labour Monotonicity. This distinction is especially relevant under desert-based theories of justice, because Conditional Labour Monotonicity arguably violates the control principle, stating that individuals should be held accountable only for what is in their control. I study the implications of assuming Conditional Labour Monotonicity or Labour Monotonicity, and will show that Labour Monotonicity implies a quantitative measure of merit, captured by a number α_i, that only depends on the individual action, while Conditional Labour Monotonicity does not.

Next, I consider a condition on merit criteria establishing that, whenever each individual labour input is multiplied by a positive constant, the merit ranking is preserved, as captured by the following definition.

Definition 3.9. *A merit criterion $\geq_{A_i}$ satisfies **Scale-Invariance** if, for each individual i, labour input ℓ_i, ℓ_i', labour input profile ℓ_{-i} and $\lambda > 0$,*

$$(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i}) \implies (\lambda\ell_i, \lambda\ell_{-i}) \geq_{A_i} (\lambda\ell_i', \lambda\ell_{-i}) .$$

Scale-Invariance is a typical requirement to characterise proportional allocation rules. It states that the merit of an individual depends only on the ratio between his and others' labour input, which is preserved when each of them is multiplied by a constant. Other invariance properties—such as additive invariance—might instead support alternative characterizations of meritocracy in which merit is monotonic in labour, but the consumption allocation depends on merit in other ways than proportional.

Lastly, I introduce the only restriction imposed on the reward criterion: Welfarist Reward. Under Welfarist Reward, an outcome constitutes a better reward for an individual if it is preferred over another outcome.

Definition 3.10. *A reward criterion satisfies **Welfarist Reward** if $\geq_i = R_i$.*

It is often argued in philosophy (Arneson, 2007; Kagan, 2014) and tacitly assumed in economics that the appropriate reward criterion should be welfare. In other words, if an individual is more meritorious, he should be better off. Assuming that welfare corresponds to preference satisfaction, Welfarist Reward represents this assumption.

I now characterise proportional meritocracy.

Proposition 3.3. *1. A meritocratic mechanism is a Proportional Meritocracy if and only if Welfarist Reward, Conditional Labour Monotonicity, and Scale-Invariance hold for each i.*

2. Assume $n \geq 3$. A meritocratic mechanism is a Proportional Meritocracy where, for each i,

$$\alpha_i(\ell) = \frac{\ell_i}{\sum_j \ell_j}\,,$$

if and only if Welfarist Reward, Labour Monotonicity, and Scale-Invariance hold for each i. In this case, the only merit criterion on preferences agreeing with merit criteria on actions satisfies the following

$$R_i \geq_{\mathcal{R}} R_i' \quad \Longleftrightarrow \quad \mathrm{MRS}_{R_i}(\ell_i, y_i) \leq \mathrm{MRS}_{R_i'}(\ell_i, y_i) \text{ for each } (\ell_i, y_i)\,.$$

Proposition 3.3 establishes that the stated assumptions are equivalent to a quantitative representation of individual merit via $\alpha_i(\ell)$, which depends on the profile of labour input. Each individual's consumption corresponds to a share of the total output, where this share reflects their relative merit. If Conditional Labour Monotonicity is strengthened to Labour Monotonicity, the measure of merit of each individual is just his labour supply, and thus does not depend on others' actions.

The second item of Proposition 3.3 establishes that, under Labour Monotonicity, the merit criterion on preferences agreeing with the merit criterion on actions is a ranking of marginal rates of substitution between labour and consumption. Such equivalence gives a more fundamental interpretation of the merit criterion underlying Proportional Meritocracy. Under different outcome functions, an individual might choose different levels of labour inputs, rendering the interpretation of the merit criterion as a measure of effort less meaningful. Instead, the marginal rate of substitution is independent of the outcome function.

There is empirical evidence suggesting that individuals adhere to a form of Proportional Meritocracy (Andre, 2024; Cappelen et al., 2024). However, these empirical studies primarily test whether impartial observers' choices satisfy Labour Monotonicity, without discussing other conditions, such as Conditional Labour Monotonicity or Scale-Invariance. Proposition 3.3 shows that a test of Labour Monotonicity is not enough to conclude one adheres to Proportional Meritocracy, and should be complemented with other empirical results. The result hopefully showcases how the framework in this paper could be employed to advance empirical studies of meritocracy. By introducing other assumptions,

such as the additive invariance discussed before, other forms of meritocracies could be characterised, and empirical studies could be designed to test whether impartial observers adhere to them.

3.5 Meritocracy and Equality of Opportunity

Before concluding, I briefly examine the relationship between meritocracy, as defined in this paper, and the concept of equality of opportunity, terms that are often used interchangeably in the economic literature. For this purpose, I compare the model here with the responsibility-sensitive allocation model of Fleurbaey (2008). I discuss a simplified version of his general model, which is also a particular case of the private good economy of the previous section. I show that meritocracy, understood as an allocation rule rewarding merit, is distinct from equality of opportunity, i.e., guaranteeing that each individual "starts on the same line", a point that has already been made elsewhere (Fleurbaey, 2008; Moulin, 2004), but is often neglected in more recent literature.

There are four individuals, each of whom has a level of bequest $b_i \in \{1, 3\}$ and dedication $a_i \in \{1, 3\}$, determining their production in monetary amounts. The set of outcomes is the set of monetary allocations such that $\sum_i x_i = \sum_i b_i a_i$. Each individual i prefers to have more money. An allocation rule maps bequest and action profiles to monetary allocations. Consider an allocation rule under which each individual consumes what he produces and there are no transfers. Such allocation rule is illustrated in the following table.

Table 3.1: No transfers

$x_i = b_i a_i$	low dedication $a_i = 1$	high dedication $a_i = 3$
low bequest $b_i = 1$	1	3
high bequest $b_i = 3$	3	9

Under this allocation rule, individuals with high bequest obtain a better outcome, thus violating equality of opportunity, since bequests are outside their control. The next table illustrates a second allocation rule neutralising the effect of bequests, transferring from individuals with high bequest to individuals with low bequest.

Table 3.2: Neutralising bequests

$x_i = (b_i + t_i)a_i$	low dedication $a_i = 1$	high dedication $a_i = 3$
low bequest $b_i = 1$	2 (transfer = +1)	6 (transfer = +1)
high bequest $b_i = 3$	2 (transfer = -1)	6 (transfer = -1)

This allocation rule may be described as one satisfying equality of opportunity, as it neutralises the effect of unequal bequests in determining the outcome. However, the allocation rule in Table 3.2 is not the only one satisfying equality of opportunity. One might want to reward high dedication, and therefore induce a better outcome for individuals with high dedication. The following table illustrates such an allocation rule.

Table 3.3: Neutralising bequests and rewarding dedication

$x_i = (b_i + t_i)a_i$	low dedication $a_i = 1$	high dedication $a_i = 3$
low bequest $b_i = 1$	1 (transfer = +0)	9 (transfer = +2)
high bequest $b_i = 3$	1 (transfer = -2)	9 (transfer = +0)

A third allocation rule which rewards dedication independently of bequests is the following.

Table 3.4: Rewarding dedication

$x_i = (b_i + t_i)a_i$	low dedication $a_i = 1$	high dedication $a_i = 3$
low bequest $b_i = 1$	0 (transfer = -1)	6 (transfer = +1)
high bequest $b_i = 3$	2 (transfer = -1)	12 (transfer = +1)

Tables 3.2 and 3.3 both represent allocation rules consistent with equality of opportunity, though they differ in how they reward dedication. Instead, allocations in Tables 3.3 and 3.4 both reward dedication, but the first satisfies equality of opportunity and the second does not.

Since they are logically independent, meritocracy and equality of opportunity can be combined. For instance, in the private good economy of Example 3.1, one could put more structure to distinguish between labour input and productivity of each individual. Then, the merit criterion could be monotonic in labour input, rather than in productive units of labour, under the assumption that productivity comes from sources the individual cannot control, such as bequests. These requirements would complement the meritocratic idea of rewarding more effort and the equality of opportunity idea of neutralising the effect of bequests. Distinguishing these two concepts clarifies the rationale behind various allocation rules, and gives guidance on how to identify the two components separately in experiments, where these ideas are often conflated.[11]

[11] As an example, Andre (2024) defines "Shallow Meritocracy" as an allocation rule that does not take into account unequal circumstances. I argue that "Shallow Meritocracy" is more accurately described as a lack of equality of opportunity, rather than a failure of meritocracy.

3.6 Conclusion

This paper develops a unifying framework for analysing meritocracy, distinguishing between two conceptually distinct notions: meritocracy as an end, represented by meritocratic social choice functions, and meritocracy as a means, represented by meritocratic mechanisms. The core innovation lies in introducing two primitives: a merit criterion that identifies when one individual is more meritorious and a reward criterion that determines which outcomes constitute superior rewards. I thus suggest that common disagreements on meritocracy can often be traced to different assumptions about the merit and reward criteria.

I showed that meritocracy as an end and as a means are equivalent, in the sense that a mechanism implementing a meritocratic social choice function is itself meritocratic. Therefore, adhering to a notion of meritocracy as a means is equivalent to adhering to a notion of meritocracy as an end. This equivalence permits merit criteria to be consistently expressed over either individuals' preferences or their actions within a mechanism.

I examined two illustrative conceptions of meritocracy from the literature: Pareto meritocracy and proportional meritocracy. I showed that Pareto meritocracy is vacuous, as it only ranks individuals who prefer a Pareto improvement to a Pareto worsening. Proportional meritocracy, defined in a private goods economy in which individuals supply labour and share their production, is more structured and allows to define a quantitative measure of merit. Assumptions on the merit and reward criteria characterise a mechanism in which each individual consumes a proportion of the total output, and such proportion coincides with his merit.

The study of meritocracy as an allocation principle remains in its early stages. Given the empirical evidence supporting the conclusion that individuals adhere to various forms of meritocracy, theoretical study is much needed. Future work could explore the implications of alternative meritocracies in other settings, or empirically test whether individuals adhere to specific meritocracies. Additionally, the relationship between meritocracy and other social justice principles, such as equality of opportunity, warrants further investigation.

References

Andre, P. (2024). Shallow meritocracy. *Review of Economic Studies*, rdae040. 32, 33, 41, 43

Arneson, R. (2007). Desert and equality. *Egalitarianism: New essays on the nature and value of equality*, 262–293. 32, 36, 41

Cappelen, A. W., De Haan, T., & Tungodden, B. (2024). Fairness and limited information: Are people Bayesian meritocrats? *Journal of Public Economics*, *233*, 105097. 31, 32, 33, 41

Cappelen, A. W., Falch, R., & Tungodden, B. (2020). Fair and unfair income inequality. *Handbook of Labor, Human Resources and Population Economics*, 1–25. 32

Cappelen, A. W., Moene, K. O., Skjelbred, S.-E., & Tungodden, B. (2023). The merit primacy effect. *The Economic Journal*, *133*(651), 951–970. 32, 33

Cappelen, A. W., Mollerstrom, J., Reme, B.-A., & Tungodden, B. (2022). A meritocratic origin of egalitarian behaviour. *The Economic Journal*, *132*(646), 2101–2117. 32, 33

Carroll, G. (2025). Is Equal Opportunity Different from Welfarism? *Working Paper*. 31

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford University Press. 31, 33, 36, 37, 42

Kagan, S. (2014). *The geometry of desert*. Oxford University Press. 29, 32, 34, 41

Markovits, D. (2019). *The meritocracy trap*. Penguin UK. 28

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. Oxford university press New York. 36

Moisson, P.-H. (2024). Meritocracy and inequality. *TSE Working Paper*, *24*(1518). 32, 35

Moisson, P.-H., & Tirole, J. (2024). Cooptation: Meritocracy vs. Homophily in Organizations. *University of Toulouse*. 31, 37

Morgan, J., Tumlinson, J., & Várdy, F. (2022). The limits of meritocracy. *Journal of Economic Theory*, *201*, 105414. 29, 32

Moulin, H. (2004). *Fair division and collective welfare*. MIT press. 31, 42

Mulligan, T. (2018). *Justice and the meritocratic state*. Taylor & Francis. 28

Mulligan, T. (2023). Meritocracy. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023 ed.). Metaphysics Research Lab, Stanford University. 32, 34, 38, 39

Roemer, J. E., & Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, *54*(4), 1288–1332. 31

Sandel, M. J. (2020). *The tyranny of merit: What's become of the common good?* Penguin UK. 28

Sen, A. (2000). Merit and justice. *Meritocracy and economic inequality*, 5–16. 28, 34

Thomson, W. (2019). *How to divide when there isn't enough.* Cambridge University Press. 39

Tirole, J. (2022). *Meritocracy and Social Justice.* Prague University of Economics and Business. 28

Chapter 4

Identifying Belief-Dependent Preferences

Abstract

Why are investors overconfident and trade excessively? Why do patients at health risk avoid testing? Why are voters polarised? Possibly because their beliefs directly influence their well-being, i.e., they have belief-dependent preferences. However, existing theories of belief-dependent preferences struggle to generate testable predictions or to identify simultaneously beliefs and preferences. This paper addresses these issues by providing an axiomatic characterization of a class of preferences and belief-updating rules that deviate from Bayesian updating. Preferences, beliefs, and updating rules are identified from choices over contingent menus, each entailing a menu of acts available at a later time contingent on an uncertain state of the world. The results provide a theory-based approach to experimental designs to test information avoidance, distortion, and other behaviours consistent with belief-dependent preferences.

4.1 Introduction

People often hold some beliefs dear, even when faced with evidence against them. Do they change their views, or do they continue believing what they want? Indeed, research documents two common attitudes towards new information: avoidance and distortion. Investors avoid obtaining information when they expect the market to be in a bad state (Karlsson et al., 2009; Sicherman et al., 2016). Donors do not learn about the impact of their donation (Andreoni et al., 2017; Chan et al., 2024). Voters shun evidence that undermines their view (Bakshy et al., 2015; Gentzkow & Shapiro, 2010). Despite being at risk, patients refrain from getting tested. (Oster et al., 2013; Thornton, 2008). If information is obtained, it is distorted when unwelcome but correctly processed if welcomed. Negotiators interpret favourably information that supports their case and dismiss the quality of contrary evidence (Babcock et al., 1995). Individuals are less likely to accept bad news about their attractiveness or intelligence but correctly process good news (Eil & Rao, 2011). They also exaggerate good news about their ability or IQ (Drobner & Goerg, 2024; Möbius et al., 2022). Those with differing opinions on climate change consider

scientists experts only when they support their views (Kahan et al., 2012).[1]

These observations conflict with expected utility theory, where more information is always desirable, and information should be processed via Bayes rule. However, if an individual derives pleasure or pain from specific beliefs, she might avoid information or process it differently. Theories of belief-dependent preferences (hereafter BDP) assume that individuals hold beliefs to enhance their well-being, which explains why they might avoid or distort information. Examples include psychological expected utility (Caplin & Leahy, 2001), optimal expectations (Brunnermeier & Parker, 2005), ego concerns (Köszegi, 2006), preferences for anticipation (Köszegi, 2010), and motivated reasoning (Bénabou, 2015; Bénabou & Tirole, 2016). These theories share a common feature: they do not provide observable data to separately identify the individual's tastes over outcomes and beliefs. This lack of identification complicates the interpretation of empirical findings, renders distinguishing between different theories difficult, and limits policy recommendations. For accurate predictions and effective policy, economists need models that are identifiable and refutable. In this paper, I offer such a model. My model implies conditions on observed choices that exhaustively characterise BDP. If choices abide by these conditions, then tastes and beliefs can be identified separately. If they do not, the theory is rejected. Such "if and only if" characterisation is absent in previous theories.

In theories of BDP, how an outcome is evaluated depends on the individual's beliefs. The lack of separation between these two concepts complicates their interpretation and identification because the way beliefs are formed may depend on tastes. Inferring tastes and beliefs from choices thus becomes a challenging task. The implication is that it is harder to obtain clear predictions. In addition to the lack of identification, current approaches make different assumptions about information processing, ranging from using Bayes rule (Köszegi, 2006) to assuming that individuals forget information (Bénabou, 2015) or even choose their beliefs at will (Brunnermeier & Parker, 2005). These assumptions are difficult to test. Therefore, the extant theories of BDP have two related limitations. They fail to make explicit the link between tastes and belief revision and tastes and beliefs are not uniquely identified.

In this paper, I propose a model of BDP that addresses these limitations. I introduce axioms on choices that characterise individuals as having tastes over their beliefs and belief revision rules related to their tastes. The theory identifies prior beliefs, tastes, and their related belief updating rules from observable choices.

I study the preferences over contingent menus. Contingent menus are collections of acts available at a later time that are conditional on the realisation of an uncertain state of the world. Consider the following illustrative example. An investor can access her banking app to receive information about the market. If she accesses the app, she observes information such as prices or the available balance. She then chooses how much

[1]For additional examples of information avoidance and distortion, see the surveys by Daniel & Hirshleifer (2015), Bénabou & Tirole (2016), and Golman et al. (2017).

to invest or withdraw. Accessing the app or not represent two *contingent menus*, each associated to a probability distribution over menus of available acts. The realisation of the menu of acts that is finally available to the individual is conditional on an uncertain state of the world. Thus, observing an available menu of acts is informative for the investor. For example, if the investor opens the app and observes a high value of her portfolio, she receives information not only about the amount of money she can withdraw but also about the state of the market. If she chooses not to access the app, she receives no information about the market and her only available act is to do nothing.

The choice between contingent menus thus reflects intertemporal preferences over acts and information. As suggested by this illustrative example, in the general model of an individual's decision problem, there are three time periods. The individual first chooses one of the available *contingent menus*, each of which delivers a distribution of menus of acts conditional on the state of the world. Then, a menu of acts available to the individual realises, and the individual updates her beliefs on the basis of the observation of the realised menu. Finally, the individual chooses an act from the realised menu.

The reason why contingent menus allow for identification is as follows. Consider an individual who has to choose between Blackwell experiments. A Bayesian individual always prefers the most informative experiment. Instead, intuitively, individuals with different preferences over beliefs should prefer different experiments. When an individual with BDP chooses between experiments, she takes expectations over her belief-dependent expected utility, particularly over the posteriors that the experiments can induce. Under Bayesian updating, the mean of the posteriors induced by any experiment is equal to the prior. Therefore, if the individual maximises her belief-dependent expected utility under Bayesian updating, her choices over experiments cannot reveal preferences over distinct posteriors, as their expectations are the same. The intuition that different belief-dependent preferences correspond to different choices of experiments fails, as shown by Eliaz & Spiegler (2006) and Liang (2017). An immediate implication is that non-Bayesian updating overcomes this problem, as the mean of posteriors that are not the Bayesian update of a prior is not necessarily equal to the prior. However, non-Bayesian updating induces dynamic inconsistency. Choices over contingent menus allow me to identify how the individual manages such dynamic inconsistency, that is, how she deviates from Bayesian updating. In turn, non-Bayesian updating allows for the identification of preferences over beliefs.

Two crucial axioms characterise BDP preferences over contingent menus. First, I need an axiom similar to the independence of von Neumann & Morgenstern (2007), but weakened to obtain BDP. Here individuals are sensitive to the informational content of a realised menu of acts. Therefore, I formulate a weak version of independence that is satisfied only among contingent menus inducing the same inference about the uncertain state (Liang, 2017; Rommeswinkel et al., 2023). This weak version of independence induces dynamic inconsistency. After receiving information, the individual is tempted to deviate from her plan. This is because, since individual tastes depend on beliefs, the

individual might distort them after receiving information. The second axiom states that the individual faces no temptation when the realised menu of acts comprises her preferred choice under both the Bayesian posterior and according to her preferred posterior. This axiom, which I call *strategic rationality for best likelihood*, ascribes the source of temptation to distorted belief updating due to BDP, and constitutes the main departure from the literature. The axiom is intuitive: it implies that the individual is not tempted when the information she receives is the one she prefers.

The main result is Theorem 4.1, which provides a functional representation of preferences over contingent menus. The choices that are consistent with the representation can be interpreted as follows. The individual chooses a contingent menu according to her preferences over acts in menu realisations and over the information it provides. She anticipates that once a menu of acts realises, she will distort her posterior beliefs away from the Bayesian update to satisfy her BDP.[2] She would like to choose according to the Bayesian posterior but is tempted to choose according to the distorted one. She solves the conflict by maximising a weighted average of two expected utilities, one under the Bayesian posterior and the other with the distorted one. The second result, Corollary 4.1, is an identification one. The choices over contingent menus allow me to uniquely pin down the individual's prior beliefs, tastes over outcomes and beliefs, distorted posterior beliefs, and the weight on the distorted utility. I illustrate the use of the representation with an application in Section 4.6, where I show how BDP can lead to belief polarisation.

Resorting to contingent menus as a primitive makes it possible to infer tastes over both beliefs and acts. This, in turn, overcomes the two issues that plague BDP highlighted above. First, it allows to test BDP by observing choices alone, in contrast to other extant theories. Second, BDP typically lead to dynamic inconsistency of behaviour in risky settings (Battigalli & Dufwenberg, 2022, p. 863). Dynamic inconsistency means that the choice the individual plans to take before receiving information is different from the one she wants to take after receiving information when she is tempted to act differently. Previous theories have relied on multiple-selves models. However, these models do not provide clear recommendations for welfare analysis, as a choice must be made regarding which of the selves' preferences are relevant. The current setting instead identifies the individual as the unit of choice, and her preferences can be subject to welfare analysis.

In sum, my model has two distinctive features compared with previous theories: it details the link between tastes over beliefs and belief updating, rather than assuming specific revision rules, and it relies only on observable choices to identify preferences, prior beliefs and revision rules, moving one step towards the empirical testing of BDP and non-Bayesian updating.

I compare the paper with the literature in the following paragraphs. Section 4.2 illustrates the primitives of the theory with two examples and provides instances of pref-

[2]This interpretation relies on the sophistication of the individual at the ex ante stage. Cobb-Clark et al. (2022) provide evidence that a majority of time-inconsistent individuals exhibit at least partial sophistication.

erences satisfying the axioms. Section 4.3 details the general model and axioms. Section 4.4 presents the results, a functional representation of preferences satisfying the axioms and the uniqueness of the components of the functional form. Section 4.5 discusses the relationships with previous models. Section 4.6 presents a simple application showing how non-Bayesian updating arising from BDP can lead to polarisation. Section 4.7 concludes by discussing how the analysis sheds light on the interpretation of BDP.

Related literature. Unlike previous papers in decision theory, I allow tastes over beliefs to interact with tastes over outcomes and identify both prior beliefs and departures from Bayes rule. To the best of my knowledge, there have been three attempts to provide an axiomatic revealed preference foundation for BDP. Dillenberger & Raymond (2020) propose a model in which an individual has preferences over the probability of realisation of compound objective lotteries. In Liang (2017), the individual has preferences over the inference her choices induce in Anscombe & Aumann (1963)'s setting. Rommeswinkel et al. (2023) is similar to Liang (2017), except that the setting is that of Savage (1972). Given that it is an objective probability framework, Dillenberger & Raymond (2020) do not cover belief identification and updating. Moreover, despite working in a dynamic setting, they do not address inconsistency. Liang (2017) identifies beliefs when tastes over these are independent of tastes over outcomes. However, his setting is static and thus silent about belief revision. Rommeswinkel et al. (2023)'s model is instead dynamically consistent and identifies prior beliefs under the same separability assumption of Liang (2017), but belief updating is not addressed.

A novelty of the present paper is the study of contingent menus delivering different inferences about the state. Variation in the inference provided by contingent menus is the key to identifying preferences and belief revision rules. Therefore, the paper is related to the literature that studies menu choice to identify departures from subjective expected utility. Ozdenoren (2002) considers contingent menus of objective lotteries as primitives. Epstein (2006) and Epstein & Kopylov (2007) instead study contingent menus of Anscombe-Aumann acts. In all these papers, the state giving rise to the menu realisation is revealed; thus, preferences for information cannot be identified. The closest paper is Epstein (2006), who provides a model of non-Bayesian updating without considering BDP but does not study choice of information.

This paper is also related to the literature explaining empirical observations with BDP. Bénabou & Tirole (2016), Golman et al. (2017) and Battigalli & Dufwenberg (2022) are three surveys on motivated beliefs, information avoidance and psychological game theory, which are related to BDP. The distinction between the previous applied work and the present paper is testability. I follow the axiomatic approach and characterise behaviourally a class of preferences and revision rules. Previous papers have "if" results that can rationalise evidence but do not allow to distinguish between different theories. Moreover, I do not tailor the model to an application, and I do not commit to a specific psychological mechanism, taste or belief revision rule. Unlike psychological games, I consider a single individual and focus on identification rather than equilibrium in games.

Finally, I address dynamic inconsistency via temptation and self-control, dispensing from previously employed multiple-selves or intrapersonal equilibrium approaches (Brunnermeier & Parker, 2005; Köszegi, 2010), which lead to problematic welfare analyses (Siniscalchi, 2011, p. 404) and are difficult to test, limiting policy recommendations.

4.2 Illustrative Examples

In this section, I develop two examples. The aim is both to illustrate the primitives of the model and to show how the theory explains some empirical observations.

Information distortion to reduce donations. A dictator game is an interaction between two individuals. One of them, the dictator, decides how much of a given amount to keep for herself and how much to transfer to the other individual, the recipient. It has been observed in the laboratory that dictators avoid receiving free information about how much of their transfer will arrive at the recipient to justify acting in a self-interested manner (Dana et al., 2007). Moreover, conditional on receiving favourable information, they transfer more (Grossman & van der Weele, 2013; Van der Weele, 2014). Similar instances of information avoidance are common.[3] I show how such attitudes towards information can be explained by non-Bayesian updating induced by BDP, and introduce a sketch of an experimental design to test the theory.

I consider a stochastic version of a dictator game. The dictator chooses how much to transfer to a recipient with whom she is coupled. The transfer is inefficient, and the receiver only receives a stochastic proportion of it. The dictator chooses whether to observe a signal on the efficiency of the transfer or not. If she observes the signal, she learns the likelihood $\ell(e)$ of the efficiency level having value e.

The dictator derives warm-glow from her expectation of the receiver outcome.[4] She would like to believe that the efficiency of the transfer is high to increase her warm-glow feelings. However, the higher the expected efficiency is, the higher the optimal transfer, which is a cost for the dictator. The dictator's tastes over transfers $x \in [0, 10)$ at likelihood ℓ are

$$u(x; \ell) = log(10 - x) + \sum_{e} p_{\ell}(e)\, ex, \tag{4.1}$$

where p_{ℓ} is the posterior belief over efficiency e after observing likelihood ℓ.[5] If the dictator chooses not to observe the signal, she chooses according to her prior beliefs, i.e., under the uninformative likelihood. The dictator maximises the sum of her material payoff, logarithmic in money, and her expectation of what the receiver receives. This expression deviates from expected utility because the information received in the form of

[3] See Section 3 in Golman et al. (2017) for multiple references.

[4] Niehaus (2014) proposed a similar preference in the context of charitable giving.

[5] The Bayesian posterior of the prior p after observing likelihood ℓ is $p_{\ell}(s) = \frac{\ell(s)p(s)}{\sum_{s'} \ell(s')p(s')}$.

the likelihood ℓ affects the taste over outcomes, not only the beliefs about their realisation. The optimal transfer at likelihood ℓ is

$$x^*(\ell) = 10 - \frac{1}{\sum_e p_\ell(e)\, e},$$

which is increasing in the expected efficiency. The function $u(x^*(\ell), \ell)$ is convex; therefore, its expectation over signal realisation is always greater than the prior. An individual computing expected utility with her Bayesian posterior and tastes in Equation (4.1) would always prefer to observe the signal and not avoid information.

Now consider a case in which the dictator distorts her beliefs after receiving the signal to maximise u. The likelihood ℓ^* of inducing the preferred beliefs satisfies the following

$$\ell^* \in \arg\max_{\ell} \max_{x} u(x; \ell).$$

In this case, the preferred beliefs give probability one to the highest level of efficiency and therefore induce a high transfer. If the dictator expects to distort her beliefs, she will avoid information, contrary to what she would do if she was Bayesian, to justify transferring less. Such extreme distortion is a particular case of the general model in the body of the paper.

The experimenter can allow the dictator to choose whether to observe the signal and to commit to a transfer conditional on the signal realisation before receiving it. If the dictator prefers to receive the signal and commit to a transfer conditional on it, it means that she wants to be informed but anticipates that she will distort the signal and act according to distorted beliefs if she has the chance.

Ostrich effect and excessive trading. This example examines how an investor's decision to seek information about the market is influenced by her tastes over the beliefs she holds. An investor chooses whether to check her financial portfolio. If she checks, she observes a signal about the state of the market and can invest or withdraw money. If she does not check, she receives no information and cannot invest or withdraw.

The investor enjoys believing that the market is in a good state and suffers when she does not. If she receives a bad signal, she suffers from negative news. Instead, if she receives a good signal, she overweights the evidence and develops overly optimistic beliefs. These distorted beliefs lead her to invest more than she would do on the basis of the Bayesian update of her prior beliefs. When choosing whether to check the portfolio, she weights the following factors: receiving bad news and suffering from it or receiving good news and acting on distorted beliefs.

If the investor has a low prior belief that the market is in a good state, she prefers not to check the portfolio to avoid unpleasant information. Instead, if she expects the market to be in good state, she may choose to check the portfolio to update her beliefs in a favourable direction rather than remaining uninformed. These behavioural patterns are well-documented (Daniel & Hirshleifer, 2015; Golman et al., 2017). The first pattern,

known as the "ostrich effect" in finance, involves avoiding unpleasant information. The second pattern involves excessive motivation from overconfidence in belief formation, in this case leading to excessive trading.

I now introduce a utility function that consistent with the investor's choices and constitutes a particular case of the theory developed in this paper. The individual values both the monetary outcome of her decision and the information she receives, with these factors being separable. She values each unit of net monetary gain independently of her beliefs. She makes a choice at two time periods: first, she chooses whether to check the portfolio; and second, conditional on her previous choice and the signal she receives, she chooses how much to invest or withdraw.

The feasible distributions of net financial gains depend on the signal received. As an example, the investor can implement various investment strategies on the basis of the prices and available balance she observes. Therefore, a signal corresponds to a menu of feasible acts the individual can choose from, denoted by M. After observing the menu of feasible acts M as a signal, the individual updates her beliefs by combining her prior $p(s)$ with the likelihood $\ell_M(s)$ of state s being true given the observed menu M. The Bayesian posterior is $p_{\ell_M}(s)$. The individual then chooses an act f, an investment strategy, from the feasible set M, whose outcome f_s depends on the realisation of the state. Assume that there are three possible states: good (g), normal (n), and bad (b). The individual's tastes over monetary outcomes are $v(f_s)$. Her tastes over outcomes and information are represented by the following

$$u(f_s; \ell_M) = v(f_s) + \omega_g \ell_M(g) + \omega_n \ell_M(n),$$

where $\omega_g > \omega_n > 0$ are numbers representing how much the individual values observing greater likelihoods that the state is good and normal. The investor values a greater likelihood of the state being good than normal, which in turn is more valuable than observing a greater likelihood of the state being bad. A Bayesian individual with these tastes maximises the expectation of u computed via the Bayesian update

$$\sum_s p_{\ell_M}(s)\, u(f_s; \ell_M).$$

The most desirable information for the investor is a likelihood vector ℓ satisfying

$$\ell^* \in \arg\max_{\ell} \max_{x} \left[v(x) + \omega_g \ell(g) + \omega_n \ell(n) \right].$$

which is $\ell^*(g) = 1$ and $\ell^*(n) = \ell^*(n) = 0$.

If the investor checks her portfolio, she can receive one of two signals: a precise signal indicating conclusively that the market is in a bad state or an imprecise signal that rules out the bad state but does not allow her to distinguish between the good and normal states. The investor anticipates that upon receiving the imprecise signal, she will distort the likelihood to $\ell^*_M(g) = 1$, her preferred one according to her tastes u. Instead, the

likelihood induced by the precise signal cannot be distorted, as there is strong conclusive evidence.

The optimal acts under the true and distorted likelihood are different because the second one is more optimistic and leads to greater investment. Theorem 4.1 from the general model establishes that, under some assumptions, the investor maximises a weighted sum of her expected utility under the true likelihood and under the distorted likelihood. Additionally, she incurs a cost depending on the utility difference between her choice and the optimal choice under the distorted posterior. The utility of choosing from the menu M at likelihood ℓ_M can be represented as follows:

$$\max_{f \in M} \left[\sum_s p_{\ell_M}(s)\, u(f_s; \ell_M) + \alpha_{\ell_M} \sum_s p_{\ell^*}(s)\, u(f_s; \ell^*_M) \right] - \max_{f' \in M} \alpha_{\ell_M} \sum_s p_{\ell^*_M}(s)\, u(f'_s; \ell^*_M), \tag{4.2}$$

where ℓ^*_M denotes the distorted vector of likelihoods after observing menu M as a signal, with $\ell^*_M(g) = 1$. The number $\alpha_{\ell_M} \geq 0$ represents the rate at which the cost of resisting temptation increases when the optimal choices under the Bayesian and distorted posteriors differ. The better an option is according to the distorted posterior, the higher the cost. It is uniquely identified in the model. When choosing whether to check her portfolio, the investor considers the probability of receiving each menu of feasible acts M as a signal and the corresponding indirect utility of choosing from it, given by the utility function above.

Consider the scenario where the investor can check her portfolio and commit to an investment strategy, for example by delegating to a financial advisor or an investment algorithm. In this case, she can receive information without the temptation to act on distorted beliefs. If the menu M observed as a signal is a singleton, the second and third terms in the indirect utility function in Equation (4.2) cancel out. As a result, preferences are represented by the expectation of tastes over outcomes and information computed with Bayes rule. Observing such preferences for commitment is a crucial test of the theory presented in the paper.[6]

These examples suggest how the model components are identified. Choosing from singleton menus does not give rise to temptation to distort beliefs. Therefore, choosing from singleton menus is a standard choice problem, and the results of Anscombe & Aumann (1963) helps to identify tastes over outcomes and beliefs after each signal. Observing choices over information and menus of acts feasible at a later time, transfers in the first example and investment strategies in the second, allow identifying a demand for commitment due to distortion of beliefs in a specific direction. Therefore, the data needed to test the model are choices over information and sets of available acts conditional on the information received. I refer to such an object of choice as a contingent menu, introduced

[6]Relatedly, Derksen et al. (2024) show that medical appointments are effective commitment devices that significantly increase the probability of individuals at health risk getting tested.

in the next section.

4.3 Model

An individual faces a dynamic decision problem in which she makes two choices. First, she chooses an act that maps states to menus of acts from which she can later select. Then, she chooses an act from the realised menu, which maps states to outcomes. Upon observing the menu from which she can choose, the individual infers information about the state. This information is relevant for her choice of an act from the realised menu.

The decision problem involves three time periods $t = 0, 1, 2$. There is a finite set of uncertain states $\mathcal{S}$. At time 0, the individual chooses a contingent menu, a mapping from uncertain states to finite probability distributions over menus of acts she will choose from at a later time. A generic menu realisation is $M \in \mathcal{M}$. A contingent menu is $F : \mathcal{S} \to \Delta^{\circ}(\mathcal{M})$, where $\Delta^{\circ}(\mathcal{M})$ is the set of probability distributions over $\mathcal{M}$ with finite support. At time 1, a menu M realises. At time 2, the individual chooses an act from the menu M. An act is a mapping between states and outcomes $f : \mathcal{S} \to \Delta(X)$, where X is a compact metric set, and $\Delta(X)$ is the set of probability distributions over X.[7] A menu M is therefore a closed nonempty set of acts.[8] The outcome induced by act f when state s realises is $f_s \in \Delta(X)$.

Observing the menu realisation M from the contingent menu F at time 1 is informative for the individual. In fact, a contingent menu is a Blackwell experiment with menus as signals. Denote the probability that menu M is realised from contingent menu F in state s with $F_s(M)$. To capture the informational content of a contingent menu F, I define the normalised likelihood (henceforth likelihood) of state s after realisation of menu M

$$\ell_{M,F}(s) := \frac{F_s(M)}{\sum_{s'} F_{s'}(M)}. \tag{4.3}$$

After having observed likelihood ℓ, the individual knows that the state is in the event

$$S_{\ell} := \{s \in \mathcal{S} \mid \ell(s) > 0\}.$$

Throughout, I assume that only choices of contingent menus are observable. The main result of this paper identifies all the components of the following model relying on these choices. In particular, I introduce the likelihood because I assume that the beliefs of the individual are not observable and will therefore be inferred from her choices. I denote preferences over contingent menus with $\succsim$. Theorem 4.1 in Section 4.4 provides the conditions on choices over contingent menus that yield the following representation of preferences.[9]

[7] The set of lotteries $\Delta(X)$ is compact metric under the weak convergence topology.

[8] The set of menus $\mathcal{M}$ is compact metric under the Hausdorff metric (Aliprantis & Border, 2006, Theorem 3.85).

[9] Notation is collected in a table in Appendix C.4.

Representation. The main result of the paper, Theorem 4.1, is a characterisation of the following functional form. The individual has belief-dependent tastes over outcomes and likelihoods represented by a utility function $u(x;\ell)$, linear in mixtures of outcomes $x \in \Delta(X)$, jointly continuous, bounded, and nonconstant for each ℓ. The dependency of u on the likelihood ℓ is the main departure from expected utility. The individual evaluates outcomes differently depending on the information she receives.

The individual acts as if she has a full support prior over states p. The Bayesian posterior of the prior p after observing the likelihood ℓ is p_ℓ, where for each state s

$$p_\ell(s) := \frac{\ell(s)\, p(s)}{\sum_{s'} \ell(s')\, p(s')}.$$

The time 2 expected utility of act f computed with the Bayesian posterior at likelihood ℓ is

$$\sum_s p_\ell(s)\, u(f_s;\ell). \tag{4.4}$$

The individual is tempted to act according to a distorted posterior. This posterior is the one obtained under the preferred likelihood consistent with the event S_ℓ. For any event $S \subseteq \mathcal{S}$, define[10]

$$\ell_S^* \in \arg\max_{\ell \in \Delta(S)} \max_{x \in \Delta(X)} u(x;\ell). \tag{4.5}$$

The distorted likelihood ℓ_S^* only assigns positive probability to states in event S. Once a state has probability 0, its probability cannot be distorted. Moreover, the distortion only depends on the event, not on the true likelihood. There is no guarantee that there is a unique likelihood satisfying Equation (4.5). However, the model allows the identification of one likelihood among those satisfying Equation (4.5) from choices over contingent menus.

The time 2 expected utility of act f computed with the distorted posterior at event S is

$$\sum_s p_{\ell_S^*}(s)\, u(f_s;\ell_S^*).$$

The distorted likelihood influences expected utility through two channels. It affects tastes over outcomes $u(x;\ell_S^*)$ and posterior beliefs, which are the Bayesian update of the prior p under the distorted likelihood ℓ_S^*.

At time 2, after having observed the menu realisation M from contingent menu F, the individual chooses an act f from M to maximise a weighted combination of the expected utilities under the true and distorted likelihood. Moreover, she suffers a cost proportional to the utility difference between the chosen act and the optimal act under the distorted

[10]Such likelihoods always exist since u is continuous and both $\Delta(X)$ and $\Delta(S)$ for each event S are compact.

likelihood. For each event S, the utility representation over menus at each likelihood ℓ is as follows:

$$\begin{aligned}\mathcal{U}(M;\ell) = &\max_{f\in M}\left[\sum_s p_\ell(s)\,u(f_s;\ell) + \alpha_\ell \sum_s p_{\ell^*_{S_\ell}}(s)\,u\left(f_s;\ell^*_{S_\ell}\right)\right] \\ &- \alpha_\ell \max_{f'\in M}\sum_s p_{\ell^*_{S_\ell}}(s)\,u\left(f'_s;\ell^*_{S_\ell}\right),\end{aligned} \tag{4.6}$$

where the positive number $\alpha_\ell \geq 0$ represents the weight assigned to the distorted expected utility. When choosing act f from menu M after realisation of the likelihood ℓ, the utility cost of temptation over menus is the difference between the second and the third terms in Equation (4.6). Specifically, it is the utility difference between the chosen act and the optimal act under the distorted likelihood. The representation implies that, for each event S, when the true likelihood coincides with ℓ^*_S, the preferred likelihood, there is no temptation. When the individual receives the information she prefers, there is no reason to distort it.

I can now describe preferences over contingent menus. The individual chooses the contingent menu F anticipating the indirect utility $\mathcal{U}(M;\ell_{M,F})$ from each possible menu realisation M, so that each contingent menu F is evaluated by the expected utility

$$\mathscr{U}(F) = \sum_M \sum_s p(s)\,F_s(M)\,\mathcal{U}(M;\ell_{M,F}). \tag{4.7}$$

To summarise, the model is as follows. When choosing the contingent menu F, the individual anticipates that her BDP will lead her to update prior beliefs p deviating from Bayes rule after observing the menu realisation M. This deviation is represented by the distortion of the true likelihood ℓ to ℓ^*, which leads to updating the prior beliefs with Bayes rule via ℓ^*. The interpretation is that, ex ante, she would like to choose from any menu to maximise her expected utility under the Bayesian update. However, she is tempted to maximise her expected utility under the Bayesian update of the prior given the distorted likelihood. She is sophisticated and foresees the temptation of choosing according to the distorted posterior, influencing both preferences and beliefs. Thus, there is a trade-off between acting according to the Bayesian and distorted posterior beliefs.

4.3.1 Axioms

In this section, I list the axioms on the preference relation $\succsim$ over contingent menus yielding the representation in Section 4.3. I begin with the standard axioms that allow a continuous utility representation of preferences over contingent menus to be obtained.

Axiom 4.1. *(**Order**). The ranking $\succsim$ is complete and transitive.*

Axiom 4.2. *(**Continuity**). For each contingent menu F the sets*

$$\{F' \mid F' \succsim F\} \text{ and } \{F' \mid F' \precsim F\}$$

are closed.

I now introduce an axiom structuring the attitude to information. Since in this model the individual has preferences for the information she receives, contingent menus with the same informativeness play a special role. First, I define the support of a contingent menu F

$$\mathcal{M}_F := \{M \in \mathcal{M} \mid F_s(M) > 0 \text{ for some } s \in \mathcal{S}\}.$$

Recall that a likelihood is a probability distribution over states defined in Equation (4.3)

$$\ell_{M,F}(s) := \frac{F_s(M)}{\sum_{s'} F_{s'}(M)}. \tag{4.3}$$

Definition 4.1. *(**Identical Inference (II)**) Two contingent menus F and F' satisfy **identical inference** if, for each menu $M \in \mathcal{M}_F \cap \mathcal{M}_{F'}$, their likelihood is the same $\ell_{M,F} = \ell_{M,F'}$.*

Two contingent menus F, F' satisfying II have the property that, when a menu M is realised from a probabilistic mixture of them, inference about the state is the same regardless of whether it comes from F or F'. To state independence, I first define the relevant mixture operations. As usual, a mixture of two acts delivers in each state a probability distribution that is the mixture of the one induced by the two acts. For any two acts f, f', state s and outcome x

$$(\lambda f + (1-\lambda) f')_s(x) = \lambda f_s(x) + (1-\lambda) f'_s(x).$$

As is standard in the menu choice literature, a mixture of two menus is a menu of mixed acts, one of which is in the first menu and the other in the second menu. For any two menus M and M' and $0 \leq \lambda \leq 1$,

$$\lambda M + (1-\lambda) M' = \{\lambda f + (1-\lambda) f' \mid f \in M, f' \in M'\}.$$

I now define mixtures of contingent menus. The contingent menu $\lambda F + (1-\lambda) F'$ delivers a distribution of menus conditional on each state s, which is a mixture of F_s and F'_s. For any two contingent menus F, F', state s and menu M

$$(\lambda F + (1-\lambda) F')_s(M) = \lambda F_s(M) + (1-\lambda) F'_s(M).$$

I now impose a weak version of independence that holds only among II contingent menus.

Axiom 4.3. *(**II Independence**). For all $0 < \lambda \leq 1$ and contingent menus F, F', F'' such that F and F'' satisfy II and F' and F'' satisfy II, $F \succsim F'$ if and only if $\lambda F + (1-\lambda) F'' \succsim \lambda F' + (1-\lambda) F''$.*

This axiom constrains preferences to depend on the realised likelihood. The intuition for the axioms is as follows. Under expected utility, independence holds to induce preferences that are linear in probabilities. However, if preferences depend on the information received, the standard independence axiom is not appropriate. This is because mixing two contingent menus affects the likelihoods that they induce for each of their menu realisation. Preferences over information should not be linear in such mixtures of likelihoods. II Independence imposes indifference only for mixtures of contingent menus inducing the same likelihood from menus in their common support.

Restricted to II contingent menus, the intuition for independence is the standard one in the menu choice literature. Adapted to the present setting, it is as follows. Consider a lottery over contingent menus delivering F with probability λ and F' with probability $1-\lambda$. The intuition for independence suggests that $F \sim F'$ iff such a mixture is indifferent to both F and F'. Once justification for indifference between probabilistic mixtures and $\lambda F_s + (1-\lambda) F'_s$ is provided, the intuition for independence is complete. Under the latter, first, the state is realised, and the individual updates her beliefs and then chooses from the available menu of mixed acts. Under probabilistic mixtures, randomisation among contingent menus is performed before the individual's choice. Hence, indifference between the two amounts to indifference to the timing of resolution of these sources of uncertainty. I briefly comment on such indifference in the conclusion.

Identify with y the contingent menu delivering in every state the singleton menu containing the act yielding the outcome y with probability 1. To avoid trivial cases, I assume the following.

Axiom 4.4. *(**Nondegeneracy**). There exist outcomes y, y' in X for which $y \succ y'$.*

Next, I adapt to the present setting the Set-Betweenness axiom of the menu choice literature. The intuition behind the axiom is that the individual prefers not to expand the available menu with ex-ante suboptimal acts, as these create temptation. A new notation is needed. For any contingent menu F, menu M in its support and menu M' outside its support, I denote with $F_{M \to M'}$ a contingent menu equivalent to F except that any realisation of M is substituted with M'. The menu M' should not be in the support of F; otherwise, $\ell_{M', F_{M \to M'}}$ would not be identical to $\ell_{M,F}$.

Axiom 4.5. *(**Set-Betweenness**). For all contingent menus F and menus M, M'*

$$F \succsim F_{M \to M'} \Rightarrow F \succsim F_{M \to M \cup M'} \succsim F_{M \to M'}.$$

The rationale for this assumption is the same as in the menu choice literature (Gul & Pesendorfer, 2001), except it holds conditional on observing a menu realisation. The preference $F \succsim F_{M \to M'}$ indicates that the individual would rather choose from M than from M', all else equal. Temptation cannot increase utility; hence, the individual prefers not to expand M with M', which contains ex-ante dominated options. Since the two

contingent menus are otherwise equivalent at the ex ante stage, the ranking in the axiom follows.

The next axiom is the main departure from the literature. Its role is to ascribe the source of temptation to belief distortion due to BDP. The idea is that the individual should not distort her beliefs when observing a likelihood that satisfies her BDP preferences. To state the axiom, I must define such a best likelihood. For this purpose, a few definitions are needed.

The following is the set of all contingent menus that induce likelihood ℓ whenever menu M realises

$$\mathcal{C}_{\ell}^{M} := \{F \mid \ell_{M,F} = \ell\}.$$

Then, I define the set of preferred outcomes at likelihood ℓ[11]

$$X_{\ell} := \left\{x \in \Delta(X) \;\middle|\; F \succsim F_{\{x\}\to\{x'\}} \text{ for all } x' \in \Delta(X) \text{ and some } F \in \mathcal{C}_{\ell}^{\{x\}}\right\}.$$

Owing to II Independence, the ranking between any two menus M and M' does not depend on the specific contingent menu F, as long as $\ell_{M,F} = \ell$. Therefore, II Independence implies that "for some" in the definition of X_{ℓ} is equivalent to "for all". A generic element of X_{ℓ} is x_{ℓ}. For illustration, I will show that in terms of the representation in Section 4.3 each x_{ℓ} satisfies the following:

$$x_{\ell} \in \arg\max_{x \in \Delta(x)} u(x;\ell).$$

Fix a collection of outcomes $(x_s)_{s\in\mathcal{S}}$. For each ℓ, construct the contingent menu F^{ℓ} such that $F_s^{\ell}(\{x_{\ell}\}) = \ell(s)$ and $F_s^{\ell}(\{x_s\}) = 1 - F_s^{\ell}(\{x_{\ell}\})$ for all s.[12] For each S, define the likelihoods

$$\ell_S^* \in \left\{\ell \in \Delta(S) \;\middle|\; F^{\ell} \succsim F^{\ell'} \text{ for all } \ell' \in \Delta(S)\right\}. \tag{4.8}$$

I show in Theorem 4.1 that any ℓ_S^* satisfies Equation (4.5) for each S:

$$\ell_S^* \in \arg\max_{\ell\in\Delta(S)} \max_{x\in\Delta(X)} u(x;\ell). \tag{4.5}$$

These likelihoods can be interpreted as follows. Say the individual can choose an outcome in $\Delta(X)$, whose realisation does not depend on the state. For each event S, the likelihood ℓ_S^* is the likelihood that the individual would prefer to observe at that event. The likelihood determined via this procedure reflects the individual's preferences over information when it is not instrumental for choice.

One more piece of notation is necessary to state the axiom. Define for each menu M and likelihood ℓ the set of acts

[11] Since $\succsim$ is continuous and $\Delta(X)$ is compact, each X_{ℓ} is nonempty.

[12] An example of the construction of such a contingent menu is shown in Appendix C.2.

$$\mathcal{F}_{M,\ell} := \left\{ f \in M \;\middle|\; F \succsim F_{\{f\}\to\{f'\}} \text{ for all } f' \in M \text{ and some } F \in \mathcal{C}_\ell^{\{f\}} \right\}.$$

Fix a menu M and a likelihood ℓ induced by it. Then, $\mathcal{F}_{M,\ell}$ is interpreted as the set of ex ante best acts in M at that likelihood. If the individual could commit, she would always choose acts from this set.[13] To illustrate, I will show that, in terms of the representation, acts in $\mathcal{F}_{M,\ell}$ are those maximising Equation (4.4) and therefore satisfying for each ℓ and M the following

$$f \in \arg\max_{f' \in M} \sum_s p_\ell(s)\, u(f'_s; \ell).$$

I now state the axiom.

Axiom 4.6. *(**Strategic Rationality for Best Likelihood (SRBL**)). For each:*

- *couple of menus M, M';*
- *contingent menu F such that $\ell_{M,F} = \ell$;*

if $\mathcal{F}_{M\cup M',\ell} \cap \mathcal{F}_{M\cup M',\ell^*_{S_\ell}} \neq \emptyset$ for at least one $\ell^*_{S_\ell}$, then

$$F \succsim F_{M\to M'} \Rightarrow F \sim F_{M\to M\cup M'}.$$

The intuition for the axiom is as follows. First, notice that the axiom implies that at the preferred likelihoods ℓ^*_S there is never temptation, as the antecedent is always satisfied. There is no reason to distort beliefs if they are the desired ones. However, the axiom imposes more. There is no temptation whenever the optimal choice under the true and preferred likelihoods coincides. The intuition is that, in this case, there is no trade-off between acting according to the true or distorted likelihood, and therefore no demand for commitment.

It is instructive to consider what would happen if II Independence is strengthened to hold among all contingent menus. In fact, the interaction between II Independence and SRBL allows interpretation of the distorted belief updating in the functional form as coming from BDP and not from other cognitive phenomena. Assume that the individual observes likelihood ℓ, which is different from one of her preferred likelihoods $\ell^*_{S_\ell}$ at that event. Say the same act in the menu M at her disposal is optimal under both likelihoods. Then, she faces no temptation and picks this act. When II Independence is strengthened to hold for all contingent menus, regardless of their informational content, the individual has no BDP. In terms of the representation, this means that u only depends on outcomes, not on likelihoods. Preferences over likelihoods are flat, and SRBL implies that there is

[13] Each menu M is a subset of the set of acts $\Delta(X)^{\mathcal{S}}$. Since $\mathcal{S}$ is finite, $\Delta(X)^{\mathcal{S}}$ is the cartesian product of compact spaces. By Theorem 2.61 in Aliprantis & Border (2006, p. 52), the cartesian product of compact spaces is compact. A menu M is thus a closed subset of a compact space, and is therefore compact. By compactness of M and continuity of preferences $\succsim$, each $\mathcal{F}_{M,\ell}$ is nonempty.

never temptation. The classical version of independence and SRBL together imply that the individual is always strategically rational and the model collapses to expected utility with Bayesian updating. SRBL traces the source of temptation to having an optimal choice that is different under the true and preferred information, and its antecedent holds for all likelihoods when there are no BDP.

The next axiom is an adaptation of state independence to the current setting. The notation fsf' indicates an act equivalent to f in state s and to f' in all states $s' \neq s$. For each state s and menus M, M', define the menu $MsM' := \{fsf' \mid f \in M, f' \in M'\}$. The menus of outcomes are denoted with $L \subseteq \Delta(X)$.

Axiom 4.7. *(**State Independence**). For all contingent menus F, menus L, L', M and states s, s'*

$$F \succsim F_{LsM \to L'sM} \;\Rightarrow\; F \succsim F_{Ls'M \to L's'M}.$$

The contingent menus F and $F_{LsM \to L'sM}$ are equivalent except in one realisation. The first offers a choice of outcomes from L in state s, whereas the second from L' in the same state. The axiom requires that the ranking of the two contingent menus is preserved when the state is changed. The intuition is that the individual's preferences over menus of outcomes are independent of the state in which the lottery realises.

Finally, I assume full support.

Axiom 4.8. *(**Full Support**). For each state s, there exist contingent menus F and F' such that for all menus M it holds that $F_{s'}(M) = F'_{s'}(M)$ for each $s' \neq s$ and $F \nsim F'$.*

If two contingent menus are always indifferent whenever they are equivalent in each state except one, then that state can be omitted without loss of generality.

4.4 Results

Representation. The main result of this paper links axioms to the utility representation in Equations (4.6) and (4.7). The proofs are in Appendix C.1. I report the utility representation and its properties.

Definition 4.2. *A ranking $\succsim$ over contingent menus is a BDP if it is represented by*

$$\mathscr{U}(F) = \sum_M \sum_s p(s)\, F_s(M)\, \mathcal{U}(M; \ell_{M,F}), \tag{4.9}$$

$$\begin{aligned} \mathcal{U}(M;\ell) = &\max_{f \in M} \left[\sum_s p_\ell(s)\, u(f_s;\ell) + \alpha_\ell \sum_s p_{\ell^*_{S_\ell}}(s)\, u\left(f_s; \ell^*_{S_\ell}\right) \right] \\ &- \alpha_\ell \max_{f' \in M} \sum_s p_{\ell^*_{S_\ell}}(s)\, u\left(f'_s; \ell^*_{S_\ell}\right), \end{aligned} \tag{4.10}$$

where:

1. $u : \Delta(X) \times \Delta(\mathcal{S}) \to \mathbb{R}$ *is linear in mixtures in* $\Delta(X)$*, jointly continuous, bounded and nonconstant for each* ℓ*;*

2. p *is a full-support probability distribution over* $\mathcal{S}$*;*

3. $\alpha_\ell \geq 0$ *for each likelihood* ℓ*;*

4. p_ℓ *is the Bayesian posterior of* p *under likelihood* ℓ*;*

5. ℓ_S^* *satisfies Equation* (4.11) *for each* S

$$\ell_S^* \in \arg\max_{\ell \in \Delta(S)} \max_{x \in \Delta(X)} u(x; \ell). \tag{4.11}$$

I can now state the main result.

Theorem 4.1. *The ranking* $\succsim$ *over contingent menus satisfies the axioms Order, Continuity, II Independence, Nondegeneracy, Set-Betweenness, SRBL, State Independence and Full Support if and only if it is a BDP.*

If an individual's preferences over contingent menus satisfy the axioms in the statement of Theorem 4.1, she behaves as if she anticipates distorting her beliefs to satisfy her BDP and act according to such distorted beliefs.

Uniqueness. I describe the uniqueness properties of the representation. Denote with $\ell^* = (\ell_S^*)_{S \in 2^S}$ a collection of preferred likelihoods satisfying condition (4.8), one for each event. Denote with p_ℓ the vector of posterior beliefs induced by observing likelihood ℓ, one for each s. Finally, $\alpha = (\alpha_\ell)_{\ell \in \Delta(S)}$.

Corollary 4.1. *Let* (u, p, α, ℓ^*) *represents* $\succsim$*, then* $(u', p', \alpha', \ell'^*)$ *also represents* $\succsim$ *if and only if:*

1. *there exists* $(a, b) \in \mathbb{R}_{++} \times \mathbb{R}$ *such that*

$$u' = au + b \quad \text{and} \quad p' = p;$$

2. *for each likelihood* ℓ*, if* $\alpha'_\ell \neq \alpha_\ell$*, then* $\ell = \ell_{S_\ell}^*$*;*

3. $\ell_S'^* = \ell_S^*$ *for each event* S*.*

It is instructive to compare the uniqueness properties of the representation with previous results from the literature on BDP. Eliaz & Spiegler (2006) and Liang (2017) show that preferences over posterior beliefs unique up to monotonic transformations cannot be identified from choices of information and acts alone. The lack of uniqueness is because the mean of the posteriors is equal to the prior beliefs. In the language of the current

model, they establish that the function $\mathcal{U}'$ represents the same ranking as $\mathcal{U}$ if and only if there exists $(a, b, c) \in \mathbb{R}_{++} \times \mathbb{R} \times \mathbb{R}^S$ such that for each likelihood ℓ

$$\mathcal{U}'(\cdot; \ell) = a\mathcal{U}(\cdot; \ell) + b - \sum_s c(s)\, p_\ell(s). \tag{4.12}$$

When taking expectations over $\mathcal{U}$, the term $\sum_s c(s)\, p_\ell(s)$ averages to a constant for all likelihoods.[14] Eliaz & Spiegler (2006) provide examples showing how the lack of uniqueness prohibits the identification of preferred beliefs. An individual's choices of information can reveal only preferences for probability distributions which mean is the prior. Non-Bayesian updating is thus responsible for the identification result in Corollary 4.1. The uniqueness properties of $\mathcal{U}$ are inherited by the function u. Owing to the functional form of $\mathcal{U}$, the term $\sum_s c(s)\, p_\ell(s)$ must necessarily be null for Equation (4.12) to hold. Since $u(\cdot; \ell_S^*)$ appears but the average of $\sum_s c(s)\, p_{\ell_S^*}(s)$, with the distorted likelihood, is not a constant, c must be 0 for the transformation to represent the same ranking.

Identification of the model components allows elaborating on the behavioural meaning of α_ℓ. First, define conditional preferences at likelihood ℓ as follows:

$$M \succsim_\ell M' \text{ if } F \succsim F_{M \to M'} \text{ for some } F \text{ such that } \ell_{M,F} = \ell,$$

where "for some F" is equivalent to "for all F" under the axioms. The ranking $\succsim_\ell$ is represented by $\mathcal{U}(\cdot; \ell)$. Consider act f and outcomes $x, x' \in \Delta(X)$ such that for some ℓ

$$\{f\} \succ_\ell \{f, x\} \succ_\ell \{x\}\ ,\ \{f\} \succ_\ell \{f, x'\} \succ_\ell \{x'\} \text{ and } \{x\} \nsim_\ell \{x'\}.$$

In other words, outcomes x and x' are tempting when choosing from menus $\{f, x\}$, $\{f, x'\}$ and x and x' are not indifferent at likelihood ℓ. Then

$$\mathcal{U}(\{f\}; \ell) - \mathcal{U}(\{f, x\}; \ell) = \alpha_\ell \left(u(x; \ell) - \sum_s p_\ell(s)\, u(f_s; \ell) \right),$$
$$\mathcal{U}(\{f\}; \ell) - \mathcal{U}(\{f, x'\}; \ell) = \alpha_\ell \left(u(x'; \ell) - \sum_s p_\ell(s)\, u(f_s; \ell) \right).$$

[14] The algebra is as follows:

$$\begin{aligned}
\sum_M \sum_s p(s) F_s(M) \sum_{s'} c(s')\, p_{\ell_{M,F}}(s') &= \sum_M \sum_s p(s) F_s(M) \sum_{s'} c(s') \frac{F_{s'}(M)\, p(s')}{\sum_{s''} F_{s''}(M)\, p(s'')} \\
&= \sum_M \sum_s c(s) F_s(M)\, p(s) \\
&= c.
\end{aligned}$$

Subtracting the two equations, the following expression for α_ℓ yields

$$\alpha_\ell = \frac{\mathcal{U}(\{f,x\};\ell) - \mathcal{U}(\{f,x'\};\ell)}{u(x;\ell) - u(x';\ell)}.$$

Theorem 9 in Gul & Pesendorfer (2001) allows interpreting α_ℓ as a measure of self-control. The ranking $\succsim_\ell$ exhibits less self-control than $\succsim'_\ell$ if for all menus of acts M and M', the ranking $M \succ_\ell M \cup M' \succ_\ell M'$ implies that the same must hold for $\succsim'_\ell$. In the current setting, self-control is relative to acting according to distorted beliefs. Therefore, the interpretation of α_ℓ can be adapted as a measure of the strength of motivated reasoning. As in Epstein (2006), the number α_ℓ is an absolute measure. The difference $\mathcal{U}(\{f\};\ell) - \mathcal{U}(\{f,x\};\ell)$ is the utility cost of self-control when the lottery x is available but the act f is chosen. Then, α_ℓ, is the rate at which the cost of resisting temptation increases as x improves, as measured by $u(x;\ell)$. It is the marginal cost of self-control at likelihood ℓ.

4.5 Discussion

In this section, I discuss the model and its relationship with the previous literature. First, it is instructive to compare this model to Epstein (2006) and Gul & Pesendorfer (2001), which are the closest in the decision theory literature. In the present model, the distortion of the likelihood induces a change in both beliefs, via non-Bayesian updating, and tastes, trough BDP. A change in both tastes and beliefs constitutes a departure from the previous literature. In Epstein (2006), temptation arises because of non-Bayesian updating due to cognitive biases, not BDP. The individual does not change tastes as represented by u but suffers from updating biases. She is thus tempted to act according to her biased posterior beliefs. In Gul & Pesendorfer (2001), instead, temptation arises because of a change in tastes. After observing her menu, the individual is tempted to choose according to new tastes v. In the present model, both sources of temptation are present and linked to each other. The individual distorts her posterior beliefs *because* her tastes u depend on them. These distortions have structure, as distorted tastes are those the individual would have if the true likelihood was her preferred one ℓ^*.

Second, I discuss the relationships with other models of BDP and non-Bayesian updating. Compared with previous models, there are two main distinctions. First, I do not take a stance on the cognitive process underlying belief distortion. The main advantage of this methodological stance is that the object of choice is observable and that predictions do not rely on a psychological interpretation. Instead, previous studies have resorted to various cognitive assumptions. As an example, in Brunnermeier & Parker (2005) and Köszegi (2006), the individual chooses her beliefs. Assuming that beliefs are chosen makes theories hard to test and obfuscates their revealed preference foundations. Spiegler (2008) shows that in the two models above the individual violates Independence of Irrelevant alternatives. The implication is that the individual's ranking of options de-

pends on the set of options itself, and predictions vary significantly. A second relevant model is that of Bénabou (2015) and Bénabou & Tirole (2016), where the individual chooses the probability with which she forgets a signal and updates her beliefs by recalling such probability. Their model features multiple-selves with different preferences playing a game whose equilibrium is a forgetting strategy. The drawback of such a modelling approach is that the relevant unit of choice is not a single individual, which complicates the interpretation of the theory and leaves degrees of freedom to conduct welfare analyses. Second, in the present model non-Bayesian updating is not disjointed from BDP. The models above instead make two disjointed assumptions: that individuals' well-being depends on their beliefs and how they update their beliefs. Instead, here knowledge of the function u implies knowledge of how beliefs are distorted.

The informational parsimony in assumptions does not come at a cost to consistency with the empirical evidence. The robust stylised fact that individuals update suitably when facing good news but fail to properly account for bad news is consistent with the model.[15] When observing a likelihood different from the preferred likelihood, individuals exhibit non-Bayesian updating, but they do not when they face what they want to hear.

I consider the posterior obtained by updating conditional on the likelihood ℓ^*, distorted compared with the Bayesian update of the subjective prior p. This is in contrast to previous literature, in which individuals distorted their beliefs compared with an objective probability distribution that is considered "true" or has an empirical counterpart observable by the modeller (Brunnermeier & Parker, 2005; Yariv, 2002). The model thus reconciles motivated belief updating to subjective Bayesianism in the tradition of Savage (1972). Such distortion operates by interpreting the objective likelihood ℓ as ℓ^*, which is arguably a mistake. I provide a second interpretation of the model in the concluding remarks relying on a more "rational" view of such a decision criterion.

The particular choice of preferred likelihood requires justification. First, why consider the preferred likelihood given that the choice is from the menus of objective lotteries? An alternative is to identify the preferred likelihood when the choice is from among all the acts. The outcome of acts depends on the realisation of the uncertain state. Therefore, such a preferred likelihood would reflect not only preferences over beliefs but also an evaluation of the instrumental value of beliefs. To illustrate why this procedure is conceptually confusing, consider an individual with no BDP. Her "preferred beliefs" would be those which put all the weight on states inducing her preferred outcome under some act. The reason for limiting choice to objective lotteries is thus to identify only the BDP component of the preferred likelihood, not the instrumental one. To see this intuition formally, consider that if I were to strengthen II Independence to hold for all contingent menus, then for any event S, all likelihoods are indifferent, as discussed above. If the individual has no BDP, her choices from objective lotteries do not depend on information

[15] See Eil & Rao (2011), Garrett & Sharot (2017), Möbius et al. (2022), Drobner & Goerg (2024), among others.

and preferences on likelihoods are flat. Assume that I considered preferred likelihoods given that the choice is from some menu of acts M. The reasoning above would not hold; the preferred likelihoods of an individual satisfying independence are those that induce her preferred outcomes with probability 1 for some act. Therefore, SRBL together with Independence would imply that an individual with no BDP is tempted to act according to such degenerate beliefs, even if there is no reason to have them. Independence and SRBL would not deliver standard expected utility with Bayesian updating.

Second, why consider the preferred likelihood when the choice is from all objective lotteries and not one conditional on each possible menu of objective lotteries? This procedure would make the preferred likelihood depend on the event and on the outcomes in X that could be induced by the available acts and thus on the menu realisation. The model would be more complex without a clear gain in scope. Moreover, the representation would be weaker, as its components depend directly on the contingent menu. Regardless, the expression of SRBL does not depend on the definition of best likelihood and can accommodate other interpretations.

4.6 Application: Polarisation

In this section, I develop a simple application of the model to show how BDP can lead to belief polarisation. I assume a sender wants to persuade an individual with BDP to take a specific action. The example is a simple variant of the judge and prosecutor example in Kamenica & Gentzkow (2011) and has various interpretations.

There is a binary state space $\mathcal{S} = \{0, 1\}$. An individual chooses an action $a \in \{0, 1\}$. She has an identity $i \in [0, 1]$, representing a belief she would like to hold over the uncertain state. The sender and the individual have a common prior over state 1 of $p = 3/10$. The individual's utility of choosing action a at state s and likelihood ℓ is

$$w(a; s, \ell) = -(a - s)^2 - (p_\ell - i)^2,$$

where p_ℓ is the posterior belief of state 1 at likelihood ℓ.[16] She would like to match the state, but also hold beliefs that match with her identity. When $i = 0$, the individual has the same preferences as in Kamenica & Gentzkow (2011). Her optimal action ex ante, under the prior belief, is $a = 0$.

Assume the sender wants to steer the individual toward choosing $a = 1$. The sender can choose among Blackwell experiments, mappings between states and distributions over action recommendations $E : \mathcal{S} \to \Delta(\{0, 1\})$, and commits to reporting the signal, as in Bayesian Persuasion. The action recommendation a realised from the experiment E induces the likelihood over states

[16] In the language of the model in the paper, each action is an AA act mapping state realisations to differences between actions and the state.

$$\ell_{a,E}(s) = \frac{E_s(a)}{\sum_{s'} E_{s'}(a)}.$$

Once the individual observes a signal, her preferences over actions are BDP as in Definition 4.2. Because $p_{\ell^*} = i$, as long as a signal does not rule out any state, preferences over actions are as follows:

$$\begin{aligned}
&- \left[p_\ell (a-1)^2 + (1-p_\ell)(a-0)^2 + (p_\ell - i)^2\right] \\
&- \alpha \left[i(a-1)^2 + (1-i)(a-0)^2\right] \\
&+ \alpha \left[(1-i)\right].
\end{aligned}$$

Otherwise, if a signal reveals the state, the second and third term of the equation above cancel out and preferences over actions reduce to:

$$- \left[p_\ell (a-1)^2 + (1-p_\ell)(a-0)^2 + (p_\ell - i)^2\right].$$

The individual has an identity of $i > 1/2$, i.e., should would like to have a belief favourable to the sender. The sender wants to maximise the probability that the individual chooses action $a = 1$. The optimal experiment is the following:[17]

	$E_s(a)$	0	1
$7/10$	0	q	$1-q$
$3/10$	1	0	1

Table 4.1: Optimal experiment for $i \in (1/2, 1]$.

where

$$q = \min\left\{\frac{4 - 10\alpha(2i-1)}{7(1-\alpha(2i-1))}, 1\right\}.$$

When $\alpha = 0$ or $i = 1/2$, the optimal experiment is the same as in Kamenica & Gentzkow (2011). The sender can induce the individual to choose action $a = 1$ more often when the individual has $i > 1/2$ compared to what she can do in the standard case when $i = 0$. Moreover, the sender can do better when α, the strength of motivated reasoning, is higher.

If the sender can target individuals with different identities, then they will have different beliefs. Even if two individuals with different identities observe the realisation of the same experiment, they will distort beliefs differently. This simple example illustrates how BDP can lead to belief polarisation even when individuals are subject to the same information.[18]

[17] Computations for this section are in Appendix C.3.

[18] For example, Kahan et al. (2012) results suggest that division over climate change stem from a desire of individuals to form beliefs in line with those held by others with whom they share close ties.

4.7 Conclusion

In this paper, I develop a theory of BDP and belief updating that can be tested by observing choices over contingent menus. I conclude by providing a second interpretation of the model and discussing a few implications of the analysis.

The interpretation of the model in the main text is that the individual distorts signals in the direction of her BDP preferences and updates beliefs with Bayes rule via distorted signals. However, the model admits a second interpretation, closer in spirit to that of Epstein (2006). When observing the realisation of the contingent menu, the individual might revise her prior beliefs rather than distort the signal. The revised prior beliefs are updated according to the true signal, leading to posterior beliefs satisfying BDP. The conceptual distinction between these two interpretations relates to the supposed "irrationality" of motivated reasoning. Distorting the objective likelihood of a signal is arguably a mistake. However, revising a prior belief is not necessarily irrational. Since it is a subjective belief, there is no objective counterpart to qualify it as "wrong". Such an interpretation is possible in this model because it features subjective beliefs contrary to the objective lotteries framework in the past literature. Under this second interpretation, discussing the trade-off between accuracy and utility from beliefs is meaningless, as there is no "accurate" belief. Moreover, regardless of the subjective or objective nature of the belief constituting a benchmark, both the current and previous models feature individuals having both the "true" belief and the "distorted" belief in mind. If the individual can formulate the trade-off between accuracy and utility, she knows what is accurate, but how can she believe something else then? This trade-off is a critical component of the BDP literature (Bénabou & Tirole, 2016). If motivated reasoning is interpreted as a rational model of decision-making, the process leading to belief revision inconsistent with Bayes rule cannot be explained in terms of an accuracy-utility trade-off. There is a second element that favours the change in prior interpretation. The model requires independence, which I justified with indifference to probabilistic mixtures. Why should the individual distort the likelihood of a signal and not the probability mixture of two contingent menus? The change in the prior interpretation is consistent with the correct processing of information coming from objective lotteries of contingent menus in the model. I thus challenge the common wisdom that the accuracy-utility trade-off is a conceptually appealing tool for theories of motivated belief updating. A trade-off between material and belief-based utility, such as the one formulated in the first example in Section 4.2, seems more intuitive.

Adopting a model of BDP leading to non-Bayesian updating has implications for agreement theorems in the style of Aumann (1976). Two individuals with the same prior beliefs but different BDP have distinct posteriors beliefs, even if these are common knowledge. Individuals with the same BDP will instead have the same posterior beliefs. This does not necessarily follow from previous models, in which BDP and non-Bayesian

updating are disjointed assumptions. This violation of Aumann makes BDP suitable candidates for explaining phenomena of polarisation and assortativity on the basis of preferences, as hinted in the application in Section 4.6.

References

Aliprantis, C. D., & Border, K. C. (2006). *Infinite dimensional analysis : A hitchhiker's guide* (3rd Ed. ed.). Berlin: Springer. 56, 62

Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving. *Journal of Political Economy*, *125*(3), 625–653. 47

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199–205. 51, 55

Aumann, R. J. (1976). Agreeing to disagree. *The annals of statistics*, *4*(6), 1236–1239. 70, 71

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review*, *85*(5), 1337–1343. 47

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. 47

Battigalli, P., & Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, *60*(3), 833–882. 50, 51

Bénabou, R. (2015). The economics of motivated beliefs. *Revue d'économie politique*(5), 665–685. 48, 67

Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141–64. 48, 51, 67, 70

Brunnermeier, M. K., & Parker, J. A. (2005). Optimal expectations. *American Economic Review*, *95*(4), 1092–1118. 48, 52, 66, 67

Caplin, A., & Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, *116*(1), 55–79. 48

Chan, T. Y., Liao, L., Martin, X., & Wang, Z. (2024). Avoiding Peer Information and Its Effects on Charity Crowdfunding: A Field Experiment. *Management Science*, *70*(4), 2272–2293. 47

Cobb-Clark, D. A., Dahmann, S. C., Kamhöfer, D. A., & Schildberg-Hörisch, H. (2022). The predictive power of self-control for life outcomes. *Journal of Economic Behavior & Organization*, *197*, 725–744. 50

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*, 67–80. 52

Daniel, K., & Hirshleifer, D. (2015). Overconfident investors, predictable returns, and excessive trading. *Journal of Economic Perspectives*, *29*(4), 61–88. 48, 53

Derksen, L., Kerwin, J. T., Reynoso, N. O., & Sterck, O. (2024). Healthcare appointments as commitment devices. *The Economic Journal*, ueae077. 55

Dillenberger, D., & Raymond, C. (2020). Additive-belief-based preferences. *PIER Working Paper*. 51

Drobner, C., & Goerg, S. J. (2024). Motivated Belief Updating and Rationalization of Information. *Management Science*, *70*(7), 4583–4592. 47, 67

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: Asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. 47, 67

Eliaz, K., & Spiegler, R. (2006). Can anticipatory feelings explain anomalous choices of information sources? *Games and Economic Behavior*, *56*(1), 87–104. 49, 64, 65

Epstein, L. G. (2006). An axiomatic model of non-Bayesian updating. *The Review of Economic Studies*, *73*(2), 413–436. 51, 66, 70

Epstein, L. G., & Kopylov, I. (2007). Cold feet. *Theoretical Economics*, *2*, 231–259. 51

Garrett, N., & Sharot, T. (2017). Optimistic update bias holds firm: Three tests of robustness following Shah et al. *Consciousness and cognition*, *50*, 12–22. 67

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, *78*(1), 35–71. 47

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of economic literature*, *55*(1), 96–135. 48, 51, 52, 53

Grossman, Z., & van der Weele, J. (2013). Self-image and strategic ignorance in moral dilemmas. *Working Paper*. 52

Gul, F., & Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, *69*(6), 1403–1435. 60, 66

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature climate change*, *2*(10), 732–735. 48, 69

Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615. 68, 69

Karlsson, N., Loewenstein, G., & Seppi, D. (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty*, *38*(2), 95–115. 47

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, *4*(4), 673–707. 48, 66

Köszegi, B. (2010). Utility from anticipation and personal equilibrium. *Economic Theory*, *44*, 415–444. 48, 52

Liang, Y. (2017). Information-dependent expected utility. *Available at SSRN 2842714*. 49, 51, 64

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, *68*(11), 7793–7817. 47, 67

Niehaus, P. (2014). A theory of good intentions. *San Diego, CA: University of California and Cambridge, MA: NBER*, *111*. 52

Oster, E., Shoulson, I., & Dorsey, E. R. (2013). Optimal expectations and limited medical testing: Evidence from Huntington disease. *American Economic Review*, *103*(2), 804–830. 47

Ozdenoren, E. (2002). Completing the state space with subjective states. *Journal of Economic Theory*, *105*(2), 531–539. 51

Rommeswinkel, H., Chang, H.-C., & Hsu, W.-T. (2023). Preference for Knowledge. *Journal of Economic Theory*, *214*, 105737. 49, 51

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications. 51, 67

Sicherman, N., Loewenstein, G., Seppi, D. J., & Utkus, S. P. (2016). Financial attention. *The Review of Financial Studies*, *29*(4), 863–897. 47

Siniscalchi, M. (2011). Dynamic choice under ambiguity: Dynamic choice under ambiguity. *Theoretical Economics*, *6*(3), 379–421. 52

Spiegler, R. (2008). On two points of view regarding revealed preference and behavioral economics. *The foundations of positive and normative economics: A handbook*, 95–115. 66

Thornton, R. L. (2008). The demand for, and impact of, learning HIV status. *American Economic Review*, *98*(5), 1829–1863. 47

Van der Weele, J. J. (2014). Inconvenient truths: Determinants of strategic ignorance in moral dilemmas. *Available at SSRN 2247288*. 52

von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton University Press. 49

Yariv, L. (2002). I'll see it when I believe it? A simple model of cognitive consistency. *Working Paper*. 67

Chapter A

Appendix for Chapter 2

A.1 Equal Sacrifice in Games

I map normal-form games to claim problems.[1] This exercise allows defining equal sacrifice universalisation for any sacrifice rule. I restrict attention to two-player games. Here $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the non-negative and positive real numbers, respectively.

A **claim problem** is an ordered list $(\{1,2\},(x_i)_{i\in I})$ where $\{1,2\}$ is the set of players and $x_i \in \mathbb{R}_{++}$ is the claim of individual i. An **award** is $y_i \in \mathbb{R}_+$ satisfying $0 \leq y_i \leq x_i$ for all i. In my formulation, the claim of each player in a game is the maximal expected utility for consequences he can obtain, which I denote with $\overline{V_i}$.[2] Therefore, $x_i = \overline{V_i}$ for all i. An **allocation rule** maps claims to awards $\pi : \mathbb{R}^2_{++} \to \mathbb{R}^2_+$. An **equal sacrifice function** is a continuous, strictly increasing and hence invertible function $R : \mathbb{R}_{++} \to \mathbb{R}$. The equal sacrifice allocation rule relative to the function R and sacrifice $s \in \mathbb{R}_+$ is

$$\pi_R\left(x_i, x_{-i}\right) := \left(R^{-1}\left(R\left(\overline{V_i}\right) - s\right)\right)_{i\in I}.$$

As an example, the equal loss rule $\pi_R\left(x_i, x_{-i}\right) = \left(x_i - s\right)_{i\in I}$ in the main text has $R\left(x_i\right) = x_i$ for all x_i. In a game, utilities depend on actions, so denote

$$U_i^c\left(\alpha_i, \alpha_{-i}\right) = \sum_{a_i, a_{-i}} \alpha_i(a_i)\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}})$$

for each (α_i, α_{-i}). An action profile inducing the maximal expected utility for i is $\left(\alpha_i^*, \alpha_{-i}^*\right)$ so that $U_i^c\left(\alpha_i^*, \alpha_{-i}^*\right) = \overline{V_i}$. Then, action α_i^{Rs} induces sacrifice s relative to the function R if

$$R^{-1}\left(R\left(U_i^c\left(\alpha_i^*, \alpha_{-i}^*\right)\right) - s\right) = U_i^c\left(\alpha_i^{Rs}, \alpha_{-i}^*\right).$$

A player exhibits equal sacrifice universalisation with respect to R if his universalisation function is $T_{\alpha_i^*, \alpha_{-i}^*}\left[\alpha_i^{Rs}\right] = \alpha_{-i}^{Rs}$.

[1] I redirect the interested reader to Thomson (2019) for a general treatment. Notice that the model in this section is not related to game-theoretic analyses of claim problems surveyed by Thomson (2013). The only purpose is to determine players' universalisation functions, not to distribute a given endowment.

[2] In this appendix, I employ the utility representation for simplicity, but everything could be defined only using ordinal preferences $\succsim_i$.

A.2 Proofs

Proof of Theorem 2.1. I omit necessity and focus on sufficiency. First, for each α_{-i}, consider the set of actions

$$\Delta^{\alpha_{-i}}(A_i) = \left\{\alpha_i \in \Delta(A_i) \mid T_{\alpha_i^*, \alpha_{-i}^*}(\alpha_i) = \alpha_{-i}\right\}.$$

This is the set of all actions that are universalised to α_{-i}. Since the game is reduced, all actions in $\Delta^{\alpha_{-i}}(A_i)$ that induce the same act are the same action. Moreover, preferences $\succsim_i$ satisfies independence when restricted to $\Delta^{\alpha_{-i}}(A_i)$ by Universalisation Independence. By Weak order, Non-triviality and Theorem 4 in Battigalli et al. (2017), preferences $\succsim_i$ are represented by

$$U_i^{\alpha_{-i}}(\alpha_i) = \sum_{a_i, a_{-i}} \alpha_i(a_i)\mu_i(a_{-i})u_i'(\rho_{a_i,a_{-i}}, \alpha_{-i}), \tag{A.1}$$

for each $\alpha_i \in \Delta^{\alpha_{-i}}(A_i)$, for some function u_i' unique up to affine transformations and unique beliefs μ_i.

Now consider the set of actions inducing constant acts

$$\Delta^c(A_i) = \left\{\alpha_i \in \Delta(A_i) \mid \rho_{\alpha_i,a_i}(x) = \rho_{\alpha_i,a_i'}(x) \text{ for each pair } a_{-i}, a_{-i}' \text{ and } x\right\}.$$

The game restricted to action inducing constant acts is also reduced, and by Lotteries independence, preferences $\succsim_i$ restricted to constant acts are a continuous weak order satisfying independence. With a slight abuse of notation, I denote the probability consequence x realises under the constant act induced by action $\alpha_i \in \Delta^c(A_i)$ with $\rho_{\alpha_i}(x)$. Again by Theorem 4 in Battigalli et al. (2017) preferences $\succsim_i$ over actions inducing constant acts can be represented by

$$U_i^c(\alpha_i) = \sum_x \rho_{\alpha_i}(x)\, u_i(x), \tag{A.2}$$

for each $\alpha_i \in \Delta^c(A_i)$, for some u_i unique up to affine transformations. Equations (A.1) and (A.2) imply that $u_i'(x, \alpha_{-i})$ should represent the same ordering as $u(x)$ for each α_{-i}, and therefore that preferences over actions in $\Delta^{\alpha_{-i}}$ are represented by

$$U_i^s(\alpha_i) = \sum_{a_i, a_{-i}} \alpha_i(a_i)\mu_i(a_{-i})u_i(\rho_{a_i,a_{-i}}), \tag{A.3}$$

for each α_{-i}.

Now consider the set of actions inducing the same act

$$\Delta^f(A_i) = \{\alpha_i \in \Delta(A_i) \mid \rho_{\alpha_i} = f\}.$$

By Universalisation evaluation, preferences $\succsim_i$ over $\Delta^f(A_i)$ must be consistent with preferences with constant acts. By Equation (A.2) and by Universalisation evaluation, these are represented by

$$U_i^c(\alpha_i) = \sum_{a_i, a_{-i}} \alpha_i(a_i) T_{\alpha_i^*, \alpha_{-i}^*}[\alpha_i](a_{-i}) u_i(\rho_{a_i, a_{-i}}), \tag{A.4}$$

for each $\alpha_i \in \Delta^f(A_i)$.

In the next step, I will "patch" the functions U_i^s in Equation (A.3) and U_i^c in Equation (A.4) into a unique function which is a convex combination of the two. In particular, I exploit the fact that the two functions are mixture linear on overlapping subdomains. Each mixed action α_i induces a pair of acts $\left(\rho_{\alpha_i}, \rho_{\alpha_i, T_{\alpha_i^*, \alpha_{-i}^*}[\alpha_i]}\right)$, where the second is a constant act. Pairs of acts lie in a connected convex domain with a product structure. By Theorem 2.2 in Chateauneuf & Wakker (1993), preferences $\succsim_i$ on the whole domain of mixed actions can thus be represented by

$$U_i(\alpha_i) = (1-\kappa) U_i^s(\alpha_i) + \kappa U_i^c(\alpha_i)$$

for each α_i, and the result follows.

□

Proof of Proposition 2.1. For the profile (α, α) to be a *SKE*, it must be the case that, for each mixed action α' and player i

$$\sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i}) u_i(\rho_{a_i, a_{-i}}) \geq \sum_{a_i, a_{-i}} \alpha'(a_i)\alpha'(a_{-i}) u_i(\rho_{a_i, a_{-i}}). \tag{A.5}$$

Under *HK* preferences, player i evaluates action α by[3]

$$U_i(\alpha) = \sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i}) u_i(\rho_{a_i, a_{-i}}).$$

For (α, α) to be a Nash Equilibrium in a game between two *HK*, it must be the case that, for each α' and i

$$U_i(\alpha) = \sum_{a_i, a_{-i}} \alpha(a_i)\alpha(a_{-i}) u_i(\rho_{a_i, a_{-i}}) \geq \sum_{a_i, a_{-i}} \alpha'(a_i)\alpha'(a_{-i}) u_i(\rho_{a_i, a_{-i}}) = U_i(\alpha') \tag{A.6}$$

Equation (A.5) and (A.6) are equivalent, one is satisfied only when the other is too, which concludes the proof. □

Proof of Proposition 2.2. For the profile (α_i, α_{-i}) to be a *MKE*, it must be the case that, for all players i and real numbers $r \geq 0$

[3]Preferences in games are usually represented by utility functions over action profiles. Since *HK* payoff only depends on his action, I can stick with my notation without loss.

$$\sum_{a_i,a_{-i}} \alpha_i(a_i)\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}) \geq \sum_{a_i,a_{-i}} r\cdot\alpha_i(a_i)r\cdot\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}). \tag{A.7}$$

Under *MHK* preferences relative to the profile (α_i, α_{-i}), player i evaluates action α_i by[4]

$$U_i\left(r\cdot\alpha_i\right) = \sum_{a_i,a_{-i}} r\cdot\alpha_i(a_i)r\cdot\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}).$$

For (α_i, α_{-i}) to be a Nash Equilibrium in a game between two *MHK*, it must be the case that, for all players i and real numbers $r \geq 0$

$$U_i(\alpha_i) = \sum_{a_i,a_{-i}} \alpha_i(a_i)\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}) \geq \sum_{a_i,a_{-i}} r\cdot\alpha_i(a_i)r\cdot\alpha_{-i}(a_{-i})u_i(\rho_{a_i,a_{-i}}) = U_i(r\cdot\alpha_i) \tag{A.8}$$

Equation (A.7) and (A.8) are equivalent, one is satisfied only when the other is too, which concludes the proof. □

I here prove a version of Proposition 2.3 that holds for all equal sacrifice rules.

Proposition A.1. *Assume the game is symmetric. Then, if an action is optimal under ESU with respect to any R, it is also optimal under HM.*

Proof of Proposition A.1. I employ the notation of Appendix A.1. Pick a profile implementing the maximal expected utility for material consequences for player i, denoted $\left(\alpha_i^*, \alpha_{-i}^*\right)$. An action α_i^{Rk} inducing sacrifice k for rule R satisfies the following:

$$R^{-1}\left(R\left(U_i^\rho\left(\alpha_i^*, \alpha_{-i}^*\right)\right) - k\right) = U_i^\rho\left(\alpha_i^{Rk}, \alpha_{-i}^*\right).$$

Since the game is symmetric, the profile $\left(\alpha_i^*, \alpha_{-i}^*\right)$ also induces a maximal consequence for player $-i$ as $U_i^\rho\left(\alpha_i^*, \alpha_{-i}^*\right) = U_{-i}^\rho\left(\alpha_{-i}^*, \alpha_i^*\right)$. Then, the condition for equal sacrifice of $-i$ is equivalent to the one of i, that implies $\alpha_i^{Rk} = \alpha_{-i}^{Rk}$ for every k. The counterfactual consequence function of player i, if he has *ESU* preferences, is thus $\phi_{\alpha_i^{Rk},a_{-i}} = \rho_{\alpha_i^{Rk},\alpha_i^{Rk}}$, which is the same as the one of *HK*. Therefore, if an action is optimal under *ESU*, it is also optimal under *HK*. □

References

Battigalli, P., Cerreia-Vioglio, S., Maccheroni, F., & Marinacci, M. (2017). Mixed extensions of decision problems under uncertainty. *Economic Theory*, *63*(4), 827–866. 76

[4]The same point of the previous footnote holds. Since *MHK* payoff only depends on his action, I can stick with my notation without loss.

Chateauneuf, A., & Wakker, P. (1993). From local to global additive representation. *Journal of Mathematical Economics*, *22*(6), 523–545. 77

Thomson, W. (2013). Game-theoretic analysis of bankruptcy and taxation problems: Recent advances. *International Game Theory Review*, *15*(03), 1340018. 75

Thomson, W. (2019). *How to divide when there isn't enough.* Cambridge University Press. 75

Chapter B

Appendix for Chapter 3

B.1 Proofs

Proof of Proposition 3.1. First, I show that if a mechanism implements a meritocratic social choice function, it is meritocratic with respect to merit criteria $\geq_i$ which agree with $\geq_{\mathcal{R}}$ for each i. Assume that M implements a meritocratic f, that is, for each preference profile R and for every Nash equilibrium strategy profile $a^*(R)$ it holds that

$$g(a^*(R)) = f(R).$$

Since f is meritocratic, for each individual i and profile R_{-i}, for all preferences R_i, R_i', it holds that

$$R_i \geq_{\mathcal{R}} R_i' \implies f(R_i, R_{-i}) \geq_i f(R_i', R_{-i}).$$

Equivalently,

$$R_i \geq_{\mathcal{R}} R_i' \implies g(a_i^*(R_i, R_{-i}), a_{-i}^*(R)) \geq_i g(a_i^*(R_i', R_{-i}), a_{-i}^*(R)),$$

and therefore the mechanism is meritocratic with respect to some merit criteria such that

$$(a_i^*(R_i, R_{-i}), a_{-i}^*(R_i, R_{-i})) \geq_{A_i} (a_i^*(R_i', R_{-i}), a_{-i}^*(R_i', R_{-i})).$$

I show that each merit criterion $\geq_{A_i}$ must agree with $\geq_{\mathcal{R}}$. Suppose, for the sake of contradiction, that for some individual i and some fixed a_{-i} the criteria $\geq_{A_i}$ and $\geq_{\mathcal{R}}$ do not agree. Then there exist two actions a_i, a_i' and corresponding preferences R_i, R_i' such that

$$R_i \geq_{\mathcal{R}} R_i' \quad \text{but} \quad (a_i^*(R_i', R_{-i}), a_{-i}^*(R_i', R_{-i})) \geq_{A_i} (a_i^*(R_i, R_{-i}), a_{-i}^*(R_i, R_{-i})).$$

Then, the mechanism would not be implementing the meritocratic f, as for the mechanism to be meritocratic it must be the case that

$$g(a_i^*(R_i', R_{-i}), a_{-i}^*(R_i', R_{-i})) \geq_i g(a_i^*(R_i, R_{-i}), a_{-i}^*(R_i, R_{-i})),$$

but

$$f(R_i, R_{-i}) \geq_i f(R_i', R_{-i}).$$

Conversely, assume that M is meritocratic with respect to some criteria $\geq_{A_i}$. Then, for each individual i and profile R_{-i}, for all preferences R_i, R_i', for each Nash equilibrium profile $a^*(R)$, it holds that

$$(a_i^*(R_i, R_{-i}), a_{-i}^*(R_i, R_{-i})) \geq_{A_i} (a_i^*(R_i', R_{-i}), a_{-i}^*(R_i', R_{-i}))$$

$$\Longrightarrow$$

$$g(a_i^*(R_i, R_{-i}), a_{-i}^*(R_i, R_{-i})) \geq_i g(a_i^*(R_i', R_{-i}), a_{-i}^*(R_i', R_{-i})).$$

Then, such mechanism implements a meritocratic social choice function f with respect to some criterion $\geq_{\mathcal{R}}$. By an argument similar to the one of the previous paragraph, it follows that each $\geq_{A_i}$ and $\geq_{\mathcal{R}}$ agree. □

Proof of Proposition 3.2. Fix an individual i and consider two preference relations R_i and R_i', each having a unique maximal element $x^*(R_i)$ and $x^*(R_i')$, respectively. Fix any profile a_{-i} and consider actions a_i, a_i' such that

$$g(a_i, a_{-i}) = x^*(R_i) \quad \text{and} \quad g(a_i', a_{-i}) = x^*(R_i').$$

By the agreement condition between the merit criteria $\geq_{A_i}$ and $\geq_{\mathcal{R}}$, it must hold that

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \iff R_i \geq_{\mathcal{R}} R_i'.$$

Since $\geq_{A_i}$ satisfies Pareto Merit, substituting $x^*(R_i)$ and $x^*(R_i')$ delivers

$$(a_i, a_{-i}) \geq_{A_i} (a_i', a_{-i}) \iff \left[x^*(R_i)\, R_j\, x^*(R_i') \text{ for all } j \text{ and } x^*(R_i)\, P_j\, x^*(R_i') \text{ for some } j\right].$$

By the agreement condition, it follows that

$$R_i \geq_{\mathcal{R}} R_i' \iff \left[x^*(R_i)\, R_j\, x^*(R_i') \text{ for all } j \text{ and } x^*(R_i)\, P_j\, x^*(R_i') \text{ for some } j\right].$$

□

Proof of Proposition 3.3. I start with item 1. I first show the necessity of assumptions. Suppose the mechanism is a Proportional Meritocracy.

Meritocracy. Define the merit criterion so that $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$ means $\ell_i \geq \ell_i'$. Then by the monotonicity of α_i, it holds that $\alpha_i(\ell_i, \ell_{-i}) \geq \alpha_i(\ell_i', \ell_{-i})$ whenever $\ell_i \geq \ell_i'$. Hence

$$g_i(\ell_i, \ell_{-i}) \;=\; \alpha_i(\ell_i, \ell_{-i})\, c^{-1}\left(\sum_j \ell_j\right) \;\geq\; \alpha_i(\ell_i', \ell_{-i})\, c^{-1}\left(\sum_j \ell_j\right) \;=\; g_i(\ell_i', \ell_{-i}),$$

which means a more meritorious action yields a strictly preferred outcome for i. Thus the mechanism is meritocratic.

Welfarist Reward. In a Proportional Meritocracy, each individual i's consumption is

$$\alpha_i(\ell)\, c^{-1}\left(\sum_j \ell_j\right).$$

Since each individual preferences are increasing in consumption, Welfarist Reward is satisfied.

Conditional Labour Monotonicity. Since α_i is monotonic in ℓ_i, if $\ell_i \geq \ell_i'$, holding ℓ_{-i} fixed, then $\alpha_i(\ell_i, \ell_{-i}) \geq \alpha_i(\ell_i', \ell_{-i})$. Thus $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$.

Scale-Invariance. Since α_i is homogeneous of degree zero, scaling ℓ to $\lambda\ell$ does not change the values of $\alpha_i(\cdot)$. Hence if $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$, the same ordering persists after multiplying all labour inputs by any $\lambda > 0$.

I now show that the assumptions imply the mechanism is a Proportional Meritocracy.

Total Feasibility. Maximal total consumption is

$$c^{-1}\left(\sum_j \ell_j\right).$$

Since preferences are increasing in consumption, a no-waste argument implies

$$\sum_{i=1}^{n} g_i(\ell) = c^{-1}\left(\sum_j \ell_j\right).$$

Define the shares

$$\alpha_i(\ell) \;:=\; \frac{g_i(\ell)}{c^{-1}\left(\sum_j \ell_j\right)}.$$

Then $\sum_i \alpha_i(\ell) = 1$. I prove that $\alpha_i(\ell)$ is monotonic in ℓ_i, holding ℓ_{-i} fixed, and homogeneous of degree zero.

Monotonicity. By Labour Monotonicity, $\ell_i \geq \ell_i'$ implies $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$. Because M is meritocratic and $\geq_i = R_i$, by Welfarist Reward, it follows that $g_i(\ell_i, \ell_{-i})$ is strictly better, i.e., a larger consumption, than $g_i(\ell_i', \ell_{-i})$. Hence $\alpha_i(\ell_i, \ell_{-i}) \geq \alpha_i(\ell_i', \ell_{-i})$. So α_i is non-decreasing in ℓ_i.

Scale-Invariance. By Scale-Invariance, $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$ implies $(\lambda\ell_i, \lambda\ell_{-i}) \geq_{A_i} (\lambda\ell_i', \lambda\ell_{-i})$. Since $\ell_i \geq \ell_i' \iff (\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$, this ordering persists under scaling, and hence $\alpha_i(\lambda\ell) = \alpha_i(\ell)$. Thus α_i is homogeneous of degree zero.

Together, these conditions imply the mechanism

$$g_i(\ell) \;=\; \alpha_i(\ell)\;\; c^{-1}\left(\sum_j \ell_j\right)$$

is a Proportional Meritocracy. This completes the proof of item 1.

I now prove item 2. I first show that if

$$g_i(\ell) \;=\; \frac{\ell_i}{\sum_j \ell_j}\, c^{-1}\left(\sum_j \ell_j\right),$$

then Labour Monotonicity holds. As before, one checks that $\sum_i \alpha_i(\ell) = 1$, $\alpha_i(\ell)$ is strictly increasing in ℓ_i, and $\alpha_i(\lambda\ell) = \alpha_i(\ell)$. Defining $(\ell_i, \ell_{-i}) \geq_{A_i} (\ell_i', \ell_{-i})$ to mean $\ell_i \geq \ell_i'$ satisfies Labour Monotonicity, because

$$\ell_i \;\geq\; \ell_i' \;\iff\; \frac{\ell_i}{\sum_j \ell_j} \;\geq\; \frac{\ell_i'}{\sum_j \ell_j} \;\iff\; \alpha_i(\ell) \;\geq\; \alpha_i(\ell').$$

I now show that the assumptions imply item 2. Assume the mechanism is meritocratic, satisfies Welfarist Reward, Labour Monotonicity and Scale-Invariance. By part 1., it is already a Proportional Meritocracy. I must show that

$$\alpha_i(\ell) = \frac{\ell_i}{\sum_j \ell_j}.$$

Since Labour Monotonicity means

$$\ell_i \;\geq\; \ell_i' \quad\iff\quad (\ell_i, \ell_{-i}) \;\geq_{A_i}\; (\ell_i', \ell_{-i}),$$

meritocracy plus Welfarist Reward implies $\ell_i \geq \ell_i' \iff \alpha_i(\ell) \geq \alpha_i(\ell')$. Also, Scale-Invariance forces $\alpha_i(\lambda\ell) = \alpha_i(\ell)$. Hence $\alpha_i(\ell)$ depends only on the ratio $\ell_i / \sum_j \ell_j$. That is,

$$\alpha_i(\ell) \;=\; F\left(\frac{\ell_i}{\sum_j \ell_j}\right)$$

for some strictly increasing function F. Since $\sum_i \alpha_i(\ell) = 1$, a standard argument (Aczél, 1966, ch. 2 Th. 2) yields $F(x) = x$. Thus

$$\alpha_i(\ell) \;=\; \frac{\ell_i}{\sum_j \ell_j}.$$

Lastly, I show agreement with the merit criterion $\geq_{\mathcal{R}}$. The argument proceeds by contradiction: if MRS_{R_i} is not uniformly smaller than $\mathrm{MRS}_{R_i'}$, the agreement fails. Suppose there exist preferences R_i and R_i' such that $\mathrm{MRS}_{R_i}(\ell_i, y_i) \leq \mathrm{MRS}_{R_i'}(\ell_i, y_i)$ in some regions, but strictly greater in some other region of (ℓ_i, y_i). I show this violates agreement.

In the region where R_i has $\mathrm{MRS}_{R_i} < \mathrm{MRS}_{R_i'}$, it is easier to make R_i better off with fewer consumption units for the same labour. One can select a labour level ℓ_i at which

$$g_i(\ell_i, \ell_{-i}) \;=\; \frac{\ell_i}{\sum_j \ell_j}\, c^{-1}\left(\sum_j \ell_j\right)$$

is strictly preferred by R_i to the outcome from a smaller labour ℓ_i'. In the region where R_i has MRS_{R_i} greater than $\mathrm{MRS}_{R_i'}$, one can choose a smaller labour ℓ_i' that yields

an outcome $g_i(\ell_i', \ell_{-i})$ preferred by R_i' to the outcome of the bigger labour ℓ_i. Hence one obtains

$$g(\ell_i, \ell_{-i}) \; R_i \; g(\ell_i', \ell_{-i}) \quad \text{and} \quad g(\ell_i', \ell_{-i}) \; R_i' \; g(\ell_i, \ell_{-i}).$$

However, agreement demands

$$\ell_i \; \geq \; \ell_i' \quad \Longleftrightarrow \quad R_i \; \geq_{\mathcal{R}} \; R_i'.$$

But also reversing the roles ($\ell_i' \geq \ell_i$) would force $R_i' \geq_{\mathcal{R}} R_i$, and thus I reached a contradiction.

References

Aczél, J. (1966). *Lectures on functional equations and their applications* (Vol. 19). Academic press. 83

Chapter C

Appendix for Chapter 4

C.1 Proofs

Proof of Theorem 4.1. Necessity is omitted. I only prove sufficiency.

First, I study the properties of conditional rankings on menus induced by preferences over contingent menus. For each likelihood ℓ, the ranking $\succsim$ induces the conditional ranking $\succsim_\ell$ on the set of menus $\mathcal{M}$ defined as

$$M \succsim_\ell M' \text{ if } F \succsim F_{M\to M'} \text{ for some } F \text{ such that } \ell_{M,F} = \ell,$$

where F and $F_{M\to M'}$ coincide except for one realisation delivering M in the first when the second delivers $M' \notin \mathcal{M}_F$. Any two contingent menus F and $F_{M\to M'}$ satisfy II, as they coincide except when delivering M and M' inducing the same likelihood. By II Independence, any mixture of F, $F_{M\to M'}$ with a third contingent menu $F_{M\to M''}$ preserves their ranking. Therefore, for all likelihoods ℓ, the ranking $\succsim_\ell$ satisfies Independence on the set of menus $\mathcal{M}$. Due to Independence, the ranking between any two menus M and M' under $\succsim_\ell$ does not depend on the specific contingent menu F, as long as $\ell_{M,F} = \ell$. Therefore, Independence implies that "for some" in the above definition of $\succsim_\ell$ is equivalent to "for all".

I now study properties of conditional preferences on particular singleton menus. The set of all contingent menus is $\mathcal{C}$. For any collection of likelihoods indexed by menus $\widehat{\ell} = \left(\widehat{\ell}_M\right)_{M\in\mathcal{M}}$, consider the set of contingent menus

$$\mathcal{C}_{\widehat{\ell}} := \left\{ F \in \mathcal{C} \;\middle|\; \ell_{M,F} = \widehat{\ell}_M \text{ and } M = \{f\} \text{ for one } f \in \Delta(X)^{\mathcal{S}} \text{ for all } M \in \mathcal{M}_F \right\}.$$

Contingent menus in any $\mathcal{C}_{\widehat{\ell}}$ only deliver distributions of singleton menus containing an act for each state s and the same likelihood $\widehat{\ell}_{\{f\}}$ for each menu realisation $\{f\}$ in their support. The set of such contingent menus is a set of AA acts $F : \mathcal{S} \to \Delta^\circ(M)$, each inducing the same likelihood at time 1 for each of the menu realisations in their support. The ranking $\succsim$ on each $\mathcal{C}_{\widehat{\ell}}$ satisfies Order, Continuity and Independence, and hence has a standard AA representation. By standard results (Fishburn, 1970, Theorem 13.1 pag. 176) preferences over contingent menus in $\mathcal{C}_{\widehat{\ell}}$ have the following representation

$$\mathscr{U}(F) = \sum_{\{f\}} \sum_{s} F_s(\{f\})\, U\left(f; \widehat{\ell}_{\{f\}}\right), \tag{C.1}$$

for all $F \in \mathcal{C}_{\widehat{\ell}}$, where $U\left(f; \widehat{\ell}_{\{f\}}\right)$ represents the conditional ranking $\succsim_{\widehat{\ell}_{\{f\}}}$ over singleton menus.

The next step is to employ Theorem 1 in Liang (2017) to show that Order, Continuity, II Independence and Nondegeneracy imply that preferences over menus depend on the likelihood these induce. Liang (2017) has a different setting, assumes the set of outcomes is infinite and has different axioms, so a few modifications of his proof are needed.

I now prove two preliminary lemmas that allow me to employ Liang's result. Recall that fsf' is an act equivalent to f in state s and to f' in all states $s' \neq s$.

Lemma C.1. *Assume $\succsim$ satisfies Order, Continuity, and II Independence. Then, for any collection of likelihoods $\widehat{\ell}$, if $F \in \mathcal{C}_{\widehat{\ell}}$ and $\widehat{\ell}_{\{f'\}}(s) = 0$ for some f' and s, then $F \sim F_{\{f'\}\to\{fsf'\}}$ for all f.*

Proof. Preferences over contingent menus in each $\mathcal{C}_{\widehat{\ell}}$ are represented by Equation (C.1)

$$\mathscr{U}(F) = \sum_{\{f\}} \sum_{s} F_s(\{f\})\, U\left(f_s; \widehat{\ell}_{\{f\}}\right), \tag{C.1}$$

and therefore

$$\begin{aligned}\mathscr{U}\left(F_{\{f'\}\to\{fsf'\}}\right) = &\sum_{\{f''\}\neq\{fsf'\}} \sum_{s'} \left(F_{\{f'\}\to\{fsf'\}}\right)_{s'}(\{f''\})\, U\left(f''_s; \widehat{\ell}_{\{f\}}\right) \\ &+ \sum_{s'} \left(F_{\{f'\}\to\{fsf'\}}\right)_{s'}(\{fsf'\})\, U\left((fsf')_{s'}; \widehat{\ell}_{\{f\}}\right).\end{aligned}$$

Due to the definition of fsf', it follows that the last term is equal to

$$\begin{aligned}&\sum_{s'\neq s} \left(F_{\{f'\}\to\{fsf'\}}\right)_{s'}(\{fsf'\})\, U\left(f'_{s'}; \widehat{\ell}_{\{f\}}\right) \\ &+ \left(F_{\{f'\}\to\{fsf'\}}\right)_{s}(\{fsf'\})\, U\left(f_s; \widehat{\ell}_{\{f\}}\right).\end{aligned}$$

If $\widehat{\ell}_{\{f'\}}(s) = 0$ then $\widehat{\ell}_{\{fsf'\}}(s) = 0$ and $\left(F_{\{f'\}\to\{fsf'\}}\right)_{s}(\{fsf'\}) = 0$. The last term of the equation is thus equal to zero. Therefore, changing f' to f in state s leads to indifference. □

Lemma C.2. *The ranking $\succsim$ satisfies Order, Continuity, II Independence, Nondegeneracy if and only there exist a function $\mathscr{U}$ representing $\succsim$ that is mixture linear in contingent menus satisfying II.*

Proof. The set of finite probability distributions on $\mathcal{M}$ is a mixture space and each contingent menu $F : \mathcal{S} \to \Delta^{\circ}(\mathcal{M})$ is an AA act inducing only lotteries with finite support. Any two mixtures of two contingent menus satisfying II also satisfy II. Therefore, if $F \succsim F' \succsim F''$ and F, F'' satisfy II, there exists a unique λ such that $\lambda F + (1-\lambda) F'' \sim F'$ (Fishburn, 1970, Theorem 8.3 pag. 112).

Consider two contingent menus such that $F \succ F'$ and $F_s(\{x_s\}) = 1$ and $F'_s(\{x'_s\}) = 1$ for all s. These contingent menus induce a distinct singleton menu in each state, and therefore each of their realisation reveals the state. These two contingent menus exist by Nondegeneracy. For each contingent menu G such that $F \succ G \succ F'$, define $\mathscr{U}(G)$ to be equal to the unique λ such that $\lambda F + (1-\lambda) F' \sim G$.

For each contingent menu G satisfying II with F' and such that $G \succ F$, define $\mathscr{U}(G)$ to be $1/\lambda$ where λ is the unique number such that $\lambda G + (1-\lambda) F' \sim F$. For each contingent menu G satisfying II with F and such that $G \prec F'$, define $\mathscr{U}(G)$ to be $\lambda/(\lambda-1)$ where λ is the unique number such that $\lambda F + (1-\lambda) G \sim F'$.

Consider a contingent menu such that $G \prec F'$ but G and F do not satisfy II. This means that, for some s, $\{x_s\} \in \mathcal{M}_G$, the support of menus induced by the contingent menu G and $\ell_{\{x_s\},G}(s) \neq 1$. By Lemma C.1, $F \sim F_{\{x_s\} \to \{fs'x_s\}}$ for all f and $s' \neq s$, since $\ell_{\{x_s\},F}(s') = 0$ for all $s' \neq s$. Because the support of any contingent menu is finite, there is always an act f such that the menu $\{fs'x_s\}$ is outside the support of G. Therefore, it is enough to define $\mathscr{U}(G)$ using the procedure above and mixing it with $F_{\{x_s\} \to \{fs'x_s\}}$. A similar construction works if $G \succ F$ but G and F' do not satisfy II.

By Proposition 1 in Liang (2017), the utility function $\mathscr{U}$ is well-defined, represents the ranking $\succsim$, and is linear in mixtures of contingent menus satisfying II. □

I now prove that Order, Continuity, II Independence, Nondegeneracy are equivalent to the following expected utility representation.

Proposition C.1. *The ranking $\succsim$ satisfies Order, Continuity, II Independence, Nondegeneracy if and only if it can be represented by*

$$\mathscr{U}(F) = \sum_M \sum_s p(s) F_s(M) \mathcal{U}(M; \ell_{M,F}) \tag{C.2}$$

for all contingent menus F, where $\mathcal{U} : \mathcal{M} \times \Delta(\mathcal{S}) \to \mathbb{R}$ is continuous and bounded on both $\mathcal{M}$ and $\Delta(\mathcal{S})$ and $p \in \Delta(\mathcal{S})$.

Proof. Since preferences over contingent menus $\succsim$ satisfy Order, II Independence, Continuity, and Nondegeneracy, by Lemmas C.1, C.2 and Theorem 1 in Liang (2017), the necessity and sufficiency of the representation in Equation (C.2) holds.

I now prove the continuity of $\mathcal{U}$ by contrapositive, I show that if it is not continuous, then Continuity does not hold. Suppose that $\mathcal{U}$ is not continuous at some point (M_0, ℓ_0) in $\mathcal{M} \times \Delta(\mathcal{S})$. Then, there exists $\varepsilon > 0$ such that for every $\delta > 0$, there is a point (M, ℓ) satisfying:

$$d(M, M_0) < \delta, \quad \|\ell - \ell_0\| < \delta, \quad \text{and} \quad |\mathcal{U}(M;\ell) - \mathcal{U}(M_0;\ell_0)| \geq \varepsilon,$$

where d is the Hausdorff metric. Consider a sequence $(M_n, \ell_n) \to (M_0, \ell_0)$. Construct a sequence of contingent menus F_n such that for each n, the menu M_n has $\sum_s p(s) F_{s,n}(M_n) > 0$, and the likelihood ℓ_{M_n,F_n}. Since the mapping from F_n to

$$\left(\sum_s p(s) F_{s,n}(M_n), \ell_{M_n,F_n}\right)$$

is continuous, and $\mathscr{U}(F_n)$ depends on $\mathcal{U}(M;\ell_{M_n,F_n})$, the discontinuity in $\mathcal{U}$ at (M_0,ℓ_0) leads to a discontinuity in $\mathscr{U}(F_n)$ at the corresponding F_0. This contradicts Continuity, which requires $\mathscr{U}$ to be continuous in F (Fishburn, 1970, Theorem 3.5 p. 36). Therefore, $\mathcal{U}$ must be continuous on $\mathcal{M}\times\Delta(\mathcal{S})$. Since $\mathcal{U}$ is continuous on a compact space, it is bounded. □

I now study the shape of $\mathcal{U}$. Because $\succsim$ satisfies Set-Betweenness, each $\succsim_\ell$ inherits the property

$$M \succsim_\ell M' \implies M \succsim_\ell M \cup M' \succsim_\ell M', \tag{C.3}$$

for all menus M, M'. By Order and Continuity, all $\succsim_\ell$ satisfy analogous properties. Since each ranking $\succsim_\ell$ satisfies Order, Continuity, Independence, and Set-Betweenness as expressed in Equation (C.3), by Theorem 1 in Kopylov (2009) it can be represented by

$$\mathcal{U}(M;\ell) = \max_{f\in M}\left\{U(f;\ell) + V(f;\ell) - \max_{f'\in M} V(f';\ell)\right\} \tag{C.4}$$

for all menus M and likelihoods ℓ, where the functions $U(\cdot;\ell)$ and $V(\cdot;\ell)$ are continuous, bounded and linear in convex combinations of acts. Moreover, for all likelihoods ℓ, the functions $U(\cdot;\ell)$ and $V(\cdot;\ell)$ satisfy the mixture space axioms.

The state dependent version of AA theorem and their continuity imply that these have the following functional forms

$$U(f;\ell) = \sum_s u(f_s;\ell,s)\ ,\ V(f;\ell) = \sum_s v(f_s;\ell,s) \tag{C.5}$$

for all acts $f:\mathcal{S}\to\Delta(X)$. Each $u(\cdot;\ell,s)$ and $v(\cdot;\ell,s)$ is continuous, bounded, and linear in mixtures of lotteries over X.

The next step is to obtain the stronger state-independent subjective expected utility representation. Recall that $L\subseteq\Delta(X)$ denotes a menu of outcome lotteries and, for each state s and menus M, M', the definition of the menu $MsM' := \{fsf' \mid f\in M, f'\in M'\}$.

Lemma C.3. *Assume $\succsim$ satisfies Order, State Independence, and Nondegeneracy. Then, for all menus L, L', M, likelihoods ℓ and states s, s' the ranking $\succsim_\ell$ satisfies*

$$LsM \succsim_\ell L'sM \ \Rightarrow\ Ls'M \succsim_\ell L's'M. \tag{C.6}$$

Moreover, for all menus M, likelihoods ℓ and states s, if $LsM \sim_\ell L'sM$ for all menus L and L', then $\ell(s) = 0$.

Proof. By definition of $\succsim_\ell$, the antecedent of condition (C.6) holds if $F \succsim F_{LsM \to L'sM}$ for some contingent menu F such that $\ell_{LsM,F} = \ell$. By State Independence, for each state s' and F, it holds that $F \succsim F_{Ls'M \to L's'M}$, which implies $Ls'M \succsim_\ell L's'M$.

If $\ell(s) \neq 0$ under the hypothesis of the second part of the Lemma, then Order and State Independence would imply that $L \sim_\ell L'$ for all L, L', contradicting Nondegeneracy. □

I now consider again preferences over contingent menus in $\mathcal{C}_{\widehat{\ell}}$. By Proposition C.1, these are represented by the following expected utility function

$$\mathscr{U}(F) = \sum_{\{f\}} \sum_{s} p(s) F_s(\{f\}) U\left(f; \widehat{\ell}_{\{f\}}\right). \tag{C.7}$$

By the first part of Lemma C.3 and Nondegeneracy, each function u in the first part of Equation (C.5) has the subjective expected utility form (Fishburn, 1970, Theorem 13.2 pag. 177), and therefore Equation (C.7) can be rewritten as

$$\mathscr{U}(F) = \sum_{\{f\}} \sum_{s} \left(\sum_{s'} p(s') F_{s'}(\{f\}) \right) p_{\widehat{\ell}_{\{f\}}}(s) u\left(f_s; \widehat{\ell}_{\{f\}}\right).$$

The probability a menu $\{f\}$ realises is $\sum_{s'} p(s') F_{s'}(\{f\})$, while the probability state s and menu $\{f\}$ realise is $p(s) F_s(\{f\})$. By the chain rule, $p_{\widehat{\ell}_{\{f\}}}$ must be the Bayesian posterior of p and therefore for each menu $\{f\}$ and state s

$$p_{\widehat{\ell}_{\{f\}}}(s) := \frac{\widehat{\ell}_{\{f\}}(s) p(s)}{\sum_{s'} \widehat{\ell}_{\{f\}}(s') p(s')} = \frac{F_s(\{f\}) p(s)}{\sum_{s'} F_{s'}(\{f\}) p(s')}.$$

Without loss of generality, for each likelihood ℓ and state s

$$u(\cdot\,; \ell, s) = p_\ell(s) u(\cdot\,; \ell), \tag{C.8}$$

where p_ℓ is the Bayesian update of the prior p under the likelihood ℓ. By Full Support the prior p has full support. By Equation (C.8), the first part of Equation (C.5) can be rewritten as

$$U(f; \ell) = \sum_{s} p_\ell(s) u(f_s; \ell). \tag{C.9}$$

Lastly, I study the shape of V. For this purpose, I need to construct a specific class of contingent menus. The following is the set of all contingent menus inducing likelihood ℓ whenever menu M realises

$$\mathcal{C}_\ell^M := \{F \in \mathcal{C} \mid \ell_{M,F} = \ell\}.$$

Next, I define the set of preferred outcomes at likelihood ℓ

$$X_\ell := \left\{ x \in \Delta(X) \;\middle|\; F \succsim F_{\{x\}\to\{x'\}} \text{ for all } x' \in \Delta(X) \text{ and some } F \in \mathcal{C}_\ell^{\{x\}} \right\}.$$

A generic element of X_ℓ is x_ℓ. Define contingent menu $\overline{F}$ so that $\overline{F}_s(\{x_s\}) = 1$ for all s, with $x_s \in \Delta(X)$. This is a contingent menu whose all menu realisations are singletons containing an outcome and revealing the state. Normalise $\mathscr{U}\left(\overline{F}\right) = 0$. For each event S, for each likelihoods $\ell \in \Delta(S)$, construct contingent menus $\overline{F}^\ell$ as follows. First, $\overline{F}^\ell$ coincides with $\overline{F}$ outside S, that is $\overline{F}_s^\ell(M) = \overline{F}_s(M)$ for each state $s \notin S$ and menu M. For each $s \in S$, two properties must hold:

1. $\overline{F}_s^\ell(\{x_\ell\}) = 1 - \overline{F}_s^\ell(\{x_s\})$;
2. $\ell_{\{x_\ell\},\overline{F}^\ell} = \ell$.

The contingent menu $\overline{F}^\ell$ always induces the likelihood ℓ when the singleton menu $\{x_\ell\}$ realises. By Theorem 1 in Liang (2017), for any contingent menu $\overline{F}^\ell$, there always exist a unique decomposition such that

$$\frac{1}{|\mathcal{S}|}\overline{F}^\ell + \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)\overline{F} = \lambda F^\ell + (1-\lambda)\overline{F},$$

where the contingent menu F^ℓ satisfies properties 1. and 2., plus $F_s^\ell(\{x_\ell\}) = \ell(s)$, and $\lambda = \frac{1}{|\mathcal{S}|}\sum_s \overline{F}_s^\ell(\{x_\ell\})$.

By Lemma C.2 and because $\mathscr{U}\left(\overline{F}\right) = 0$, it follows that

$$\frac{1}{|\mathcal{S}|}\mathscr{U}\left(\overline{F}^\ell\right) + \left(\frac{|\mathcal{S}|-1}{|\mathcal{S}|}\right)\mathscr{U}\left(\overline{F}\right) = \lambda\mathscr{U}\left(F_s^\ell\right) + (1-\lambda)\mathscr{U}\left(\overline{F}\right)$$
$$\mathscr{U}\left(\overline{F}^\ell\right) = \sum_s F_s^\ell(\{x_\ell\})\,\mathscr{U}\left(F_s^\ell\right).$$

In the construction of Proposition C.1, the function $\mathcal{U}(\{x_\ell\};\ell)$ is defined to be $\mathscr{U}\left(F_s^\ell\right)$. Moreover, since $\{x_\ell\}$ is a singleton menu whose only element is an outcome, by Equations (C.4) and (C.9), the function $\mathcal{U}(\{x_\ell\};\ell)$ represents the same ordering as $u(x_\ell;\ell)$. Therefore, observing choices among menus constructed as F_s^ℓ allows identifying the preferred likelihoods according to the ranking represented by u.

For each event S, define the likelihoods

$$\ell_S^* \in \left\{ \ell \in \Delta(S) \;\middle|\; F^\ell \succsim F^{\ell'} \text{ for all } \ell' \in \Delta(S) \right\}. \tag{C.10}$$

These are the likelihoods in the statement of SRBL. Because $\mathscr{U}\left(F^\ell\right)$ represents the same ordering as $u(x_\ell;\ell)$, and because $x_\ell \in X_\ell$ for each ℓ, likelihoods satisfying condition (C.10) also satisfy Equation (4.5), for each event S

$$\ell_S^* \in \arg\max_{\ell\in\Delta(S)} \max_{x\in\Delta(X)} u(x;\ell). \tag{4.5}$$

The next step is to use SRBL to show that for each event S, for each likelihood $\ell \in \Delta(S)$

$$V(\cdot;\ell) = \alpha_\ell U(\cdot;\ell_S^*) + \beta_\ell \tag{C.11}$$

for some ℓ_S^* satisfying condition (C.10), with $\alpha_\ell \geq 0$ for each ℓ and for some number β_ℓ. That is, the temptation ranking at event S coincides with the ranking U when one of the preferred likelihoods at that event realises. I prove the claim by contrapositive, i.e., if $V(\cdot;\ell)$ represents any other ranking, then SRBL is violated.

Assume that, for an event S and a likelihood $\ell \in \Delta(S)$, the temptation ranking $V(\cdot;\ell)$ does not represent the same ranking as any $U(\cdot;\ell_S^*)$ over acts, for all ℓ_S^* satisfying condition (C.10). The implication is that there exist two acts f, f' such that $U(f;\ell_S^*) > U(f';\ell_S^*)$ and $V(f';\ell) > V(f;\ell)$. Consider the menus $M = \{f\}$ and $M' = \{f'\}$, and a likelihood ℓ such that the antecedent of SRBL holds, that is

$$\mathcal{F}_{M\cup M',\ell} \cap \mathcal{F}_{M\cup M',\ell_S^*} \neq \emptyset \text{ for at least one } \ell_S^*, \tag{C.12}$$

where one such ℓ always exists, since the set of likelihoods satisfying condition (C.10) is non-empty and taking ℓ from that set suffices. By Equations (C.2) and (C.12), there is a common maximal element of $U(\cdot;\ell)$ and $U(\cdot;\ell_S^*)$, for one ℓ_S^*, in $M \cup M'$, which by hypothesis must be f. However, the maximal element of $V(\cdot;\ell)$ is f', which, by Equation (C.2), implies that $F \succ F_{M\to M\cup M'}$, in violation of SRBL. Therefore, if Equation (C.11) is violated, then SRBL is violated.

By Equations (C.4), (C.9) and (C.11) it follows that for each menu M, event S and likelihood $\ell \in \Delta(S)$

$$\begin{aligned}\mathcal{U}(M;\ell) = &\max_{f\in M}\left\{\sum_s p_\ell(s)\, u(f_s;\ell) + \alpha_\ell \sum_s p_{\ell_S^*}(s)\, u(f_s;\ell_S^*)\right\} \\ &- \max_{f'\in M} \alpha_\ell \sum_s p_{\ell_S^*}(s)\, u(f'_s;\ell_S^*),\end{aligned} \tag{C.13}$$

which, together with Equation (C.2) delivers the representation.

□

Proof of Corollary 4.1. I omit the necessity part of the statement. Suppose that both (u, p, α, ℓ^*) and $(u', p', \alpha', \ell'^*)$ represent $\succsim$, where $\ell^* = (\ell_S^*)_{S\in 2^S}$ is a collection of likelihoods satisfying condition (C.10), one for each event. I first show the uniqueness properties of u and p. By AA subjective expected utility theorem, u' represents the same ranking as u if and only if, for all likelihoods ℓ, there exists $a_\ell, b_\ell \in \mathbb{R}_{++} \times \mathbb{R}$ such that

$$u'(\cdot;\ell) = a_\ell u(\cdot;\ell) + b_\ell \text{ and } p'_\ell = p_\ell. \tag{C.14}$$

For each likelihood $\ell \in \Delta(S)$, either

$$\alpha_\ell \neq 0 \text{ and } \ell_S^* \neq \ell \tag{C.15}$$

or not. The vector (u, p, α, ℓ^*) violates Equation (C.15) if and only if $(u', p', \alpha', \ell'^*)$ also does. If this is the case, either $\alpha_\ell = 0$, from which also $\alpha'_\ell = 0$, or $\ell = \ell_S^*$, and then, by Equation (C.13), any couple of $\alpha_\ell, \alpha'_\ell$ preserves the ordinal ranking. If Equation (C.15) holds, Theorem 4 in Gul & Pesendorfer (2001) implies that

$$V'(\cdot;\ell) = A_\ell V(\cdot;\ell) + B_{V,\ell} \tag{C.16}$$
$$U'(\cdot;\ell) = A_\ell U(\cdot;\ell) + B_{U,\ell} \tag{C.17}$$

where, by Theorem 4.1, for each event S and likelihood $\ell \in \Delta(S)$

$$V(f;\ell) = \alpha_\ell \sum_s p_{\ell_S^*}(s)\, u(f_s;\ell_S^*)$$

for all f and one ℓ_S^* satisfying condition (C.10). For each act f, event S and likelihood $\ell \in \Delta(S)$

$$\begin{aligned} U'(f;\ell) &= \sum_s p_\ell(s)\, u'(f_s;\ell) \\ &= \sum_s p_\ell(s)\,(a_\ell u(f_s;\ell) + b_\ell) \\ &= a_\ell \sum_s p_\ell(s)\, u(f_s;\ell) + \sum_s p_\ell(s)\, b_\ell \\ &= a_\ell U(f;\ell) + B_{U,\ell}, \end{aligned}$$

from which $A_\ell = a_\ell$ for all likelihoods ℓ. By the uniqueness result in Theorem 1 of Liang (2017), the function $\mathcal{U}'$ represents the same ranking as $\mathcal{U}$ if and only if there exist $(a, b, c) \in \mathbb{R}_{++} \times \mathbb{R} \times \mathbb{R}^S$ such that for all likelihoods ℓ

$$\mathcal{U}'(\cdot;\ell) = a\mathcal{U}(\cdot;\ell) + b - \sum_s c(s)\, p_\ell(s).$$

Define the set of all contingent menus only containing singleton menus in their support

$$\overline{\mathcal{C}} := \left\{ F \in \mathcal{C} \;\middle|\; M = \{f\} \text{ for some } f \in \Delta(X)^S \text{ for all } M \in \mathcal{M}_F \right\}.$$

By Theorem 4.1, preferences over $\overline{\mathcal{C}}$ are represented by

$$\mathscr{U}(F) = \sum_{\{f\}} \sum_s p(s)\, F_s(\{f\})\, U\left(f; \ell_{\{f\},F}\right)$$

for all $F \in \overline{\mathcal{C}}$. Therefore, U is defined as $\mathcal{U}$ for singleton menus, and it inherits its uniqueness properties (Kopylov, 2009). In the present setting

$$\begin{aligned}\mathcal{U}'(\{f\};\ell) &= a_\ell U(f;\ell) + B_{U,\ell} \\ &= a_\ell \mathcal{U}(\{f\};\ell) + B_{U,\ell},\end{aligned}$$

from which $a_\ell = a$ and $B_{U,\ell} = b - \sum_s c(s)\, p_\ell(s)$ for all ℓ. Because of the functional form of $\mathcal{U}$, the function $\mathcal{U}'$ only represents the same ranking if $c = 0$. In fact, for each event S and likelihood $\ell \in \Delta(S)$, substitution into the representation delivers

$$\begin{aligned}\mathcal{U}'(M;\ell) = \max_{f\in M} \Bigg\{ & U'(f;\ell) + V'(f;\ell) - \max_{f\in M} V'(f;\ell) \Bigg\} \\ = \max_{f'\in M} \Bigg\{ & aU(f;\ell) + b - \sum_s c(s)\, p_\ell(s) \\ & + \alpha_\ell \left(aU(f;\ell^*_S) + b - \sum_s c(s)\, p_{\ell^*_S}(s) \right) \\ & - \max_{f'\in M} \alpha_\ell \left(aU(f';\ell^*_S) + b - \sum_s c(s)\, p_{\ell^*_S}(s) \right) \Bigg\}.\end{aligned}$$

When taking expectations of $\mathcal{U}$, averaging the term $\sum_s c(s)\, p_\ell(s)$ makes it constant and equal to the prior for any likelihood, allowing to preserve the ranking.[1] However, the same does not hold for $\sum_s c(s)\, p_{\ell^*_S}(s)$. Therefore, to preserve ordinal equivalence, c must be null, otherwise the expression $\sum_s c(s)\, p_{\ell^*_S}(s)$ does not average to the prior and U' does not represent the same ranking as U. Moreover, by Equation (C.14) it follows that $b_\ell = b$ for each ℓ.

Next, I derive the uniqueness of α_ℓ for each likelihood ℓ. For each event S and likelihood $\ell \in \Delta(S)$, substitution of u in the expression of V delivers

$$\begin{aligned}V'(f,\ell) &= \alpha'_\ell \sum_s p_{\ell^*_S}(s)\,(au(f_s;\ell^*_S) + b) \\ &= a\alpha_\ell \sum_s p_{\ell^*_S}(s)\, u(f_s;\ell^*_S) + B_{V,\ell} \\ &= aV(f,\ell) + B_{V,\ell}\end{aligned}$$

[1] Consider the representation in Theorem 4.1. The algebra is as follows:

$$\begin{aligned}\sum_M \sum_s p(s)\, F_s(M) \sum_{s'} c(s')\, p_{\ell_{M,F}}(s') &= \sum_M \sum_s p(s)\, F_s(M) \sum_{s'} c(s')\, \frac{F_{s'}(M)\, p(s')}{\sum_{s''} F_{s''}(M)\, p(s'')} \\ &= \sum_M \sum_s c(s)\, F_s(M)\, p(s) \\ &= c\,.\end{aligned}$$

for all f, where the last equality follows from Equation (C.16) and the fact that $A_\ell = a$ for all ℓ. Therefore, $\alpha_\ell = \alpha'_\ell$ for each ℓ to preserve the ordinal ranking.

Lastly, I show that, for each event S, if the likelihood ℓ'^*_S part of ℓ'^* represents the same preferences over contingent menus as ℓ^*_S, then $\ell'^*_S(s) = \ell^*_S(s)$ for all states $s \in S$. I prove it by contrapositive, if this is not the case, then SRBL is violated.

First, I show that any two ℓ^*_S and ℓ'^*_S must induce the same posterior beliefs. Fix an event S and assume that both ℓ'^*_S and ℓ^*_S satisfy Equation (C.10) and that $p_{\ell'^*_S}(s) \neq p_{\ell^*_S}(s)$ for some s. Then, $U(\cdot, \ell'^*_S)$ does not represent the same ordering over acts of $U(\cdot, \ell^*_S)$. Assume this is not the case and these represent the same ordering over acts. Then, the representation of the ranking over constant acts $u(\cdot; \ell'^*_S)$ must be an affine transformation of $u(\cdot; \ell^*_S)$. However, since by hypothesis $p_{\ell'^*_S}(s) \neq p_{\ell^*_S}(s)$ for some s, and by Equation (C.9), $U(\cdot; \ell'^*_S)$ does not represent the same ordering over acts of $U(\cdot; \ell^*_S)$, which is absurd.

By Theorem 4.1, preferences over menus at ℓ^*_S can be represented by the following

$$\mathcal{U}(M; \ell^*_S) = \max_{f \in M} \left\{ U(f; \ell^*_S) + \alpha_{\ell^*_S} U(f; \ell'^*_S) - \max_{f' \in M} \alpha_{\ell^*_S} U(f'; \ell'^*_S) \right\},$$

for each menu M, since ℓ'^*_S satisfies condition (C.10). For any menu M, the antecedent of SRBL holds, as ℓ^*_S satisfies Equation (C.10) and therefore trivially

$$\mathcal{F}_{M \cup M', \ell^*_S} \cap \mathcal{F}_{M \cup M', \ell''^*_S} \neq \emptyset \text{ for at least one } \ell''^*_S \text{ satisfying Equation (C.10).}$$

However, if $U(\cdot; \ell'^*_S)$ and $U(\cdot; \ell^*_S)$ do not represent the same ordering over acts, its consequent will not hold in general. Consider two acts f, f' such that $U(f; \ell'^*_S) > U(f'; \ell'^*_S)$ and $U(f'; \ell'^*_S) > U(f; \ell'^*_S)$ and construct menus $M = \{f\}$ and $M' = \{f'\}$. By Equation (C.2), $F \succ F_{M \to M \cup M'}$ for any F such that $\ell_{M,F} = \ell^*_S$, in violation of SRBL. Therefore, if $p_{\ell'^*_S}(s) \neq p_{\ell^*_S}(s)$ for some s, then SRBL is violated.

Fixing a prior p, there is a one to one relationship between likelihood and posterior. For each likelihood ℓ and state s

$$\ell(s) = \frac{\frac{p_\ell(s)}{p(s)}}{\sum_{s'} \frac{p_\ell(s')}{p(s')}}.$$

By Equation (C.14), the prior p representing preferences is unique. Therefore, since the posteriors satisfy $p_{\ell'^*_S}(s) = p_{\ell^*_S}(s)$ for each s, it follows that $\ell'^*_S(s) = \ell^*_S(s)$ for each s. □

C.2 Construction of Best Likelihoods

I here provide an example of the construction of contingent menus that allows to identify the preferred likelihoods in Theorem 4.1. I start from these two contingent menus, F and $\overline{F}$, where the utility of the second is normalised to 0:

$$\begin{matrix} F & \overline{F} \end{matrix}$$
$$\begin{bmatrix} \{x_\ell\} & (0.1, 0.4) \\ \{y\} & (0.9, 0) \\ \{z\} & (0, 0.6) \end{bmatrix} \qquad \begin{bmatrix} \{x_{s_1}\} & (1, 0) \\ \{x_{s_2}\} & (0, 1) \end{bmatrix}.$$

The aim is to represent a combination of these two contingent menus as a combination of three contingent menus, one for each menu in the support of F. The combination will be such that a part of it averages to give 0 utility and the rest gives the utility of each menu in the support of F at the likelihood it induces in that menu, preserving their realisation probability. Each of the three contingent menus is then defined as the utility of choosing from their corresponding menu in F at their likelihood.

Each of these contingent menus must contain one menu realisation from F at the same likelihood. So, for each menu in the support of F, I change the probability of its realisation in each state until it coincides with its normalised likelihood. Then , I fill the rest of the contingent menu with elements from $\overline{F}$. As an example, for $\{x_\ell\}$

$$\begin{bmatrix} \{x_\ell\} & (0.2, 0.8) \\ \{x_{s_1}\} & (0.8, 0) \\ \{x_{s_2}\} & (0, 0.2) \end{bmatrix}.$$

All three of them are as follows:

$$\begin{bmatrix} \{x_\ell\} & (0.2, 0.8) \\ \{x_{s_1}\} & (0.8, 0) \\ \{x_{s_2}\} & (0, 0.2) \end{bmatrix} \qquad \begin{bmatrix} \{y\} & (1, 0) \\ \{x_{s_2}\} & (0, 1) \end{bmatrix} \qquad \begin{bmatrix} \{z\} & (0, 1) \\ \{x_{s_1}\} & (1, 0) \end{bmatrix}.$$

I must construct a linear combination of these that coincides with a linear combination of the original contingent menus F and $\overline{F}$. In F, conditional on a state, a menu M realises with probability $F_s(M)$. In the three new contingent menus, conditional on a state, the probability that a menu M in the support of F realises is its normalised likelihood, namely $\frac{F_s(M)}{\sum_{s'} F_{s'}(M)}$. Therefore, to make the conditional probability of realisation coincides, the weight on each new contingent menu must be $\sum_{s'} F_{s'}(M)$, to cancel the denominator. However, summing for each menu yields

$$\sum_{M \in \mathcal{M}_F} \sum_{s'} F_{s'}(M) = \mid \mathcal{S} \mid,$$

which is greater than 1. Therefore, the weight on each new contingent menu should be $\frac{\sum_{s'} F'_s(M)}{|\mathcal{S}|}$, which results in the following linear combination:

$$0.25 \begin{bmatrix} \{x_\ell\} & (0.2, 0.8) \\ \{x_{s_1}\} & (0.8, 0) \\ \{x_{s_2}\} & (0, 0.2) \end{bmatrix} + 0.45 \begin{bmatrix} \{y\} & (1, 0) \\ \{x_{s_2}\} & (0, 1) \end{bmatrix} + 0.3 \begin{bmatrix} \{z\} & (0, 1) \\ \{x_{s_1}\} & (1, 0) \end{bmatrix}.$$

Since the conditional probability of each menu has been divided by the number of states, the probability of realisation of F in combination with $\overline{F}$ should be $\frac{1}{|\mathcal{S}|}$, to make conditional probabilities coincide

$$\frac{1}{2}\begin{bmatrix}\{x_\ell\} & (0.1,0.4)\\ \{y\} & (0.9,0)\\ \{z\} & (0,0.6)\end{bmatrix}+\frac{1}{2}\begin{bmatrix}\{x_{s_1}\} & (1,0)\\ \{x_{s_2}\} & (0,1)\end{bmatrix}$$

$$=0.25\begin{bmatrix}\{x_\ell\} & (0.2,0.8)\\ \{x_{s_1}\} & (0.8,0)\\ \{x_{s_2}\} & (0,0.2)\end{bmatrix}+0.45\begin{bmatrix}\{y\} & (1,0)\\ \{x_{s_2}\} & (0,1)\end{bmatrix}+0.3\begin{bmatrix}\{z\} & (0,1)\\ \{x_{s_1}\} & (1,0)\end{bmatrix}.$$

The conditional probability any menu in $\overline{F}$ realising also coincides in both linear combination, as it is

$$\begin{aligned}\sum_{M\in\mathcal{M}_F}\frac{\sum_{s'}F_{s'}(M)}{|\mathcal{S}|}\left(1-\frac{F_s(M)}{\sum_{s''}F_{s''}(M)}\right)&=\frac{1}{|\mathcal{S}|}\sum_{M\in\mathcal{M}_F}\left(\sum_{s''}F_{s''}(M)-F_s(M)\right)\\&=\frac{1}{|\mathcal{S}|}\left(\sum_{M\in\mathcal{M}_F}\sum_{s''}F_{s''}(M)-\sum_{M\in\mathcal{M}_F}F_s(M)\right)\\&=\frac{1}{|\mathcal{S}|}(|\mathcal{S}|-1)\\&=\frac{|\mathcal{S}|-1}{|\mathcal{S}|}.\end{aligned}$$

C.3 Computation for Section 4.6

I here provide the computations for the application in Section 4.6. The sender chooses a vector of q_i, r_i for all i, i.e., an experiment as follows:

	$E_s(a)$	0	1
$^7/_{10}$	0	q_i	$1-q_i$
$^3/_{10}$	1	$1-r_i$	r_i

Table C.1: Experiment for individual i.

He must be sure that each i follows the action recommendation. As long as q_i and r_i are in the interior of the unit interval, the programme is :

$$\max_{q_i, r_i} r_i \left(\frac{3}{10}\right) + (1 - q_i)\left(\frac{7}{10}\right)$$

$s.t.$

$IC_{0\to 1}$:

$$-\frac{7}{10}\left(\frac{q_i}{\frac{3}{10}(1-r_i)+\frac{7}{10}q_i}\right)\cdot 0 - \frac{3}{10}\left(\frac{1-r_i}{\frac{3}{10}(1-r_i)+\frac{7}{10}q_i}\right)\cdot 1 - \alpha\left[(1-i)\cdot 0 + i\cdot 1\right]$$
$$> -\frac{7}{10}\left(\frac{q_i}{\frac{3}{10}(1-r_i)+\frac{7}{10}q_i}\right)\cdot 1 - \frac{3}{10}\left(\frac{1-r_i}{\frac{3}{10}(1-r_i)+\frac{7}{10}q_i}\right)\cdot 0 - \alpha\left[(1-i)\cdot 1 + i\cdot 0\right]$$

$IC_{1\to 0}$:

$$-\frac{7}{10}\left(\frac{1-q_i}{\frac{3}{10}r_i+\frac{7}{10}(1-q_i)}\right)\cdot 1 - \frac{3}{10}\left(\frac{r_i}{\frac{3}{10}r_i+\frac{7}{10}(1-q_i)}\right)\cdot 0 - \alpha\left[(1-i)\cdot 1 + i\cdot 0\right]$$
$$> -\frac{7}{10}\left(\frac{1-q_i}{\frac{3}{10}r_i+\frac{7}{10}(1-q_i)}\right)\cdot 0 - \frac{3}{10}\left(\frac{r_i}{\frac{3}{10}r_i+\frac{7}{10}(1-q_i)}\right)\cdot 1 - \alpha\left[(1-i)\cdot 0 + i\cdot 1\right]$$

The terms $(p_\ell - i)^2$ and $\alpha\left[(i-i)\right]$ appear in both sides of all inequalities and are therefore omitted. If $r_i = 1$, then when the individual observes the recommendation to choose $a = 0$ the state is revealed and $IC_{0\to 1}$ reduces to $0 < 1$, that is always satisfied. The constraint $IC_{1\to 0}$ becomes:

$$IC_{1\to 0}: \frac{7}{10}\left(\frac{1-q_i}{\frac{3}{10}+\frac{7}{10}(1-q_i)}\right) + \alpha(1-i)$$
$$< \frac{3}{10}\left(\frac{1}{\frac{3}{10}+\frac{7}{10}(1-q_i)}\right) + \alpha i,$$

from which

$$q_i = \frac{4 - 10\alpha(2i-1)}{7(1-\alpha(2i-1))}.$$

When $\alpha = 0$ or $i = 1/2$, q_i coincides with the solution of the Bayesian Persuasion problem with standard preferences. The derivatives of q_i with respect to α and i are both negative.

C.4 Notation

Symbol [elements]	Name	Mathematical object
X	outcomes	compact metric set
$\Delta(X)$ $[x, y]$	(lotteries over) outcomes	compact metric set
$\mathcal{S}$ $[s, s']$	states	finite set
$S \subseteq \mathcal{S}$	events	finite set
f, f'	acts	functions $\mathcal{S} \to \Delta(X)$
$\mathcal{M}$ $[M, M']$	menus	compact metric set
$\Delta^{\circ}(\mathcal{M})$	finite lotteries over menus	compact metric set
$\mathcal{C}$ $[F, F']$	contingent menus	functions $\mathcal{S} \to \Delta^{\circ}(\mathcal{M})$
$\succsim$	preference	subset of $\mathcal{C} \times \mathcal{C}$
ℓ	likelihoods	probability distribution over $\mathcal{S}$
p	prior beliefs	probability distribution over $\mathcal{S}$
p_{ℓ}	posterior beliefs	probability distribution over $\mathcal{S}$
u	utility functions	functions $\Delta(X) \times \Delta(\mathcal{S}) \to \mathbb{R}$
$\mathcal{U}$	utility functions	functions $\mathcal{M} \times \Delta(\mathcal{S}) \to \mathbb{R}$
$\mathscr{U}$	utility functions	functions $\mathcal{C} \to \mathbb{R}$

Table C.2: Symbols, their names, and corresponding mathematical objects.

References

Fishburn, P. C. (1970). *Utility theory for decision making.* New York: Wiley. 85, 86, 88, 89

Gul, F., & Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, *69*(6), 1403–1435. 92

Kopylov, I. (2009). Temptations in General Settings. *The B.E. Journal of Theoretical Economics*, *9*(1). 88, 92

Liang, Y. (2017). Information-dependent expected utility. *Available at SSRN 2842714*. 86, 87, 90, 92