A Foundation for Universalisation in Games

Enrico Mattia Salonia[*]

October 7, 2024

Abstract

Revealed preference theory equates choices with preferences over the consequences these choices induce. Nevertheless, if a decision criterion prescribes an act for reasons unrelated to its consequences, the inference drawn regarding preferences can be misleading. I study the behaviour of non-consequentialist individuals who have preferences for universalisation. They choose the action that, in a counterfactual scenario where it is also chosen by everyone else, leads to their preferred consequences. I develop a model for individuals who value their choices in light of the counterfactual consequences they induce. Choices are interpreted as revealing a preference for counterfactual consequences. I impose axioms 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.

1 Introduction

What would I get if everyone behaved 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 acted as they do. 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 frameworks (Roemer, 2010).

[*] Toulouse School of Economics. mattia.salonia@tse-fr.eu. I am indebted to Ingela Alger and François Salanié for countless discussions, comments, and invaluable guidance throughout various stages of this project. I also thank, Matteo Broso, Annalisa Costella, Philippe De Donder, Franz Dietrich, Mostapha Diss, Alberto Grillo, Peter Hammond, Pau Juan Bartroli, Giacomo Rubbini, Karine Van Der Straeten and participants at various workshops and conferences for helpful feedback. I acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics).

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 choice-theoretic model and introduce axioms that characterise preferences for universalisation. This model enables the unification of previous ones and provides a justification for existing empirical identification practices. 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, not consequences, as the primitive. An example in Section 1.1 illustrates that behaviour consistent with preferences for universalisation in a game cannot be rationalised by a ranking over material consequences. This motivates the use of the choice-theoretic model of Luce & Raiffa (1957), where the object of choice is an element of an action set. An individual chooses an action under uncertainty. The chosen action and the realisation of an uncertain state lead to a material consequence. The novelty is that these also lead to a counterfactual consequence, i.e., what would happen under a different realisation of the uncertain state. The individual cares about both the material and the counterfactual consequences of his actions. If, in a game, opponents' actions are interpreted as the uncertain state, an individual with prefer-

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

ences for universalisation cares about the counterfactual consequence that obtains if everyone else acts as he does.

The first result of this paper, Proposition 1, establishes the requirement for a weak order over actions to be equivalent to a weak order over pairs of material and counterfactual consequences. The necessary and sufficient condition for this equivalence is a property I call Extended Consequentialism, which requires that two actions yielding the same distribution of material and counterfactual consequences must be ranked equally. If an individual is a consequentialist with respect to an expanded domain of consequences that includes more than just material features, his ranking over actions corresponds to a ranking over these extended consequences.

The main result, Theorem 1, provides an expected utility representation of preferences over actions that are additive in material and counterfactual concerns. Extended Consequentialism implies that preferences are not sensitive to the correlation between material and counterfactual consequences. Using an argument from the literature on conjoint measurement (Fishburn, 1970, Ch. 11), I show that the absence of preferences for correlation guarantees that material and counterfactual concerns are aggregated additively. Because the theorem is silent on the shape of preferences over sure consequences, it reveals that universalisation and pro-social preferences are distinct attitudes; it is possible for an individual to exhibit both, consistent with empirical evidence (van Leeuwen & Alger, 2024). The theorem implies that welfare analysis for individuals with preferences for counterfactual consequences 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 non-consequentialist preferences, as they desire to induce a specific counterfactual consequence. 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 complementing the axioms of Theorem 1, 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). In particular, the axioms link rankings over actions to the universalisation counterfactual envisioned by the individual. I discuss the testability of these axioms and how they provide guidance for empirical research. I comment on the difference between my foundation for Kantian Equilibrium and that of Roemer. I argue that his model's properties can be preserved by abandoning the distinction between the "optimisation protocol", a concept he introduces, and preferences, resulting in a more parsimonious framework.

The model allows me to 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 differ-

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

ence in aim satisfaction, that is, an equal sacrifice. I show that this counterfactual evaluation 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.

The paper is organized as follows: In Section 2, I introduce the primitives of the model and the main axioms. The main theorem is presented in Section 3. In Section 4, I derive characterisations of various models of universalisation. A novel definition of universalisation, called equal sacrifice universalisation, is introduced in Section 5. Section 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), featuring a mapping between pairs of actions and states to consequences. I consider a two-dimensional alternative in which also a counterfactual consequence is induced by a second mapping, that describes what would happen under a different realisation of the state. The analysis builds on an observation by Battigalli et al. (2017). The authors show under which conditions Luce & Raiffa's approach is equivalent to the more tractable one of Anscombe & Aumann (1963). The requirement guaranteeing equivalence of the two is an assumption of consequentialism. The individual in Luce & Raiffa must be indifferent between two actions that induce the same distribution of consequences. I extend consequentialism to comprise both material and counterfactual consequences. The approach allows rationalising universalisation without expanding the set of consequences.

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. I argue in the motivating example in Section 1.1 that the more structured approach of this paper is necessary to characterise preferences for universalisation.

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 introduction of counterfactual consequences 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.

The theory developed here is distinct from various forms of regret aversion (Loomes & Sugden, 1982). In regret aversion, individuals consider their payoff in a counterfactual scenario where they have chosen differently. In contrast, the theory presented here involves individuals considering a counterfactual scenario where the uncertain state has a different realisation.

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 Equilibrium. 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 the literature investigating 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) and Ponthiere (2024) study Epictetusian and Stoic preferences, while Shi (2024) introduces a preference for a social minimum consumption level.

1.1 Illustrative example

I illustrate the contribution of this paper through an example. I show the issues arising with universalisation if consequences are exclusively material. In these cases, choices cannot be rationalised by a complete and transitive preference relation and predictions depend on the labelling of actions. 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 1: Preference reversal.

Assume the row player has beliefs μ, highlighted in blue in Table 1 and 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, maximises 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 red in Table 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.

That the mere inclusion of material consequences, monetary rewards in this example, does not allow to incorporate all the relevant features of a decision problem is not new. Context-dependent preferences by Gilboa & Schmeidler (2003) can handle this issue. The authors derive an expected utility representation for preferences ranking pairs of actions and state realisations, corresponding, in both this paper and theirs, to opponents' actions. Context-dependent preferences rationalise universalisation in this example, as one would obtain a real number in the table above such that ℓ is preferred to both m and r regardless of column player's action. Although this works, such a general model is silent on the determinants of preferences and makes it difficult to test a particular hypothesis, such as the presence of preferences for universalisation. Moreover, they consider beliefs as primitive objects, making it difficult to distinguish universalisation from magical thinking. One solution is to use a less general model that specifies the relevant features of the decision problem. For example, in psychological games preferences depend on both material consequences and players' beliefs (Geanakoplos et al., 1989; Battigalli & Dufwenberg, 2009). With reciprocity, preferences depend on previous opponents' actions (Charness & Rabin, 2002). In these cases, beliefs or previous actions are consequences. The primitives of these models can be elicited or observed. I take a similar route, studying a decision problem in which preferences are observable. Such a modelling choice gives empirical guidance, as I discuss in Section 3. I defend again this methodological stance in Section 6.

This example shows that universalisation depends 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 5, I present a novel definition of universalisation, relying on the general theory, that is equivalent under any redescription of actions.

2 Model

In this section, I introduce the primitives of the model and axioms to derive a general functional representation of preferences of which particular cases are studied in the rest of the paper. For any set X, I denote with $\Delta(X)$ the set of finite probability distributions over X.

Primitives. I study decision problems under uncertainty defined as follows.

Definition 1. *A **decision problem** is an ordered list $\mathcal{D} = \left(A, S, C, \rho, \phi, \succsim^A\right)$, featuring:*

- *a set of actions A, the set of mixed actions is $\Delta(A)$;*
- *a finite set of states of the world S;*
- *a set of consequences C;*
- *a **material** consequence function $\rho : A \times S \to C$;*
- *a **counterfactual** consequence function $\phi : A \times S \to C$;*
- *a ranking over mixed actions $\succsim^A$.*

The material consequence function maps actions and state realisations to consequences. It is related to the nature of the decision problem. The counterfactual consequence function is instead a feature of the individual. It describes the link between his behaviour and a counterfactual consequence he envisions. A counterfactual consequence function is such that, for all actions a and states s, it holds that $\phi_{a,s} = \rho_{a,s'}$ for one specific s'. The counterfactual consequence is the material consequence under an alternative realisation of the state of the world. Counterfactual consequence functions are unobservable. However, they can be inferred from behaviour under certain conditions. I will discuss this point when I introduce models of universalisation that allow such an inference.

Each pair of pure action a and state realisation s induces a pair of material and counterfactual consequences $(c, c') = (\rho_{a,s}, \phi_{a,s})$ where $c, c' \in C$. Any mixed action $\alpha \in \Delta(A)$ induces an Anscombe & Aumann (AA) act $\rho_\alpha : S \to \Delta(C)$ leading to consequence c under state s with probability $\rho_{\alpha,s}(c) = \alpha(\{a \in A \mid \rho_{a,s} = c\})$. The same holds for ϕ_α. I refer to ρ_α as the material act and to ϕ_α as the counterfactual act. Each mixed action α induces therefore the AA act $(\rho \circ \phi)_\alpha : S \to \Delta(C \times C)$, mapping states to distribution over pairs of consequences. I assume there exist mixed actions that, under various states, can induce every possible pair of distributions of consequences. A sufficient condition for this to hold is that for each pair of consequences (c, c') there exists an action a such that $(\rho_{a,s}, \phi_{a,s}) = (c, c')$ for all states s. This is a standard richness assumption.

The consequence functions allow me to study under which conditions choice of an action can be interpreted as a revealed preference for consequences. It is impossible to investigate this link in von Neumann & Morgenstern, Anscombe & Aumann or Savage decision problems. In von Neumann & Morgenstern, individuals choose lotteries over consequences. There is no conceptual distinction between action and consequence. As for Anscombe & Aumann and Savage, the objects of choice are acts, functions from states to consequences. In the language of this paper, a Savage act is the section at A of the material consequence function $\rho_a : S \to C$. These are richer than von Neumann & Morgenstern, but still collapse the relation between action and consequence.

Axioms. Throughout the paper, I assume preferences satisfy the von Neumann & Morgenstern axioms.

Axiom 1. *(vNM) For all actions $\alpha, \alpha', \alpha''$:*

1. *(**Weak Order**) the ranking $\succsim^A$ is complete and transitive;*
2. *(**Continuity**) the sets*

$$\{\lambda \in [0,1] \mid \lambda\alpha + (1-\lambda)\alpha' \succsim^A \alpha''\} \text{ and } \{\lambda \in [0,1] \mid \alpha'' \succsim^A \lambda\alpha + (1-\lambda)\alpha'\}$$

   are closed;

3. *(**Independence**) if $\alpha \sim^A \alpha'$ then $\frac{1}{2}\alpha + \frac{1}{2}\alpha'' \sim^A \frac{1}{2}\alpha' + \frac{1}{2}\alpha''$.*

To avoid trivial cases, I assume the following.

Axiom 2. *(**Non-degeneracy**) There exist actions α, α' such that $\alpha \succ^A \alpha'$.*

The crucial axiom in this paper imposes that the individual is indifferent between two actions that induce the same material and counterfactual acts, namely the same distribution over consequences in each state.

Axiom 3. ***(Extended consequentialism)*** *For all actions α, α', if $\rho_\alpha = \rho_{\alpha'}$ and $\phi_\alpha = \phi_{\alpha'}$, then $\alpha \sim^A \alpha'$.*

Extended consequentialism allows the individual to prefer an action to another even if these two induce the same material act. In consequentialist models this possibility is ruled out. One could also be indifferent between actions that induce the same counterfactual act, but different material act. As I show in Section 4, this is the case for individuals behaving according to Kantian Equilibrium and for the extreme case of Homo Moralis, Homo Kantiensis. I refer to such a preference as purely non-consequentialist. Being either purely consequentialist or non-consequentialist is consistent with Extended consequentialism. In general, individuals who care about both material and counterfactual consequences satisfy Extended consequentialism without being purely consequentialist or non-consequentialist. Extended consequentialism is also a reduction condition. The individual cannot prefer a compound lottery over actions that leads to the same acts as a simple lottery. Lastly, Extended consequentialism rules out preferences for the correlation structure between the two acts. A weakening of Extended consequentialism allowing for preferences sensitive to the correlation between material and counterfactual acts requires indifference between actions α and α' if $(\rho \circ \phi)_\alpha = (\rho \circ \phi)_{\alpha'}$.[3]

The next axiom requires a new piece of notation. Define an incomplete ranking over acts $f \in \Delta(C)^S$ as follows: $f \succsim f' \iff \alpha \succsim^A \alpha'$ for all α, α' such that $\rho_\alpha = \phi_\alpha = f$ and $\rho_{\alpha'} = \phi_{\alpha'} = f'$.

Axiom 4. ***(Separability)*** *For all actions α, α', the following two conditions hold:*

1. *if $\phi_\alpha = \phi_{\alpha'}$, $\rho_\alpha = f$ and $\rho_{\alpha'} = f'$, then $f \succsim f' \iff \alpha \succsim^A \alpha'$;*
2. *if $\rho_\alpha = \rho_{\alpha'}$, $\phi_\alpha = f$ and $\phi_{\alpha'} = f'$, then $f \succsim f' \iff \alpha \succsim^A \alpha'$.*

Separability imposes that the ranking over acts in the two dimensions, material and counterfactual, is the same.

3 Functional Representation

From Actions to Consequences. My first result shows that choices of actions can be interpreted as revealing a preference over pairs of material and counterfactual acts if and only if Extended consequentialism holds.

Proposition 1. *Assume the ranking $\succsim^A$ is a weak order. Then, the ranking over AA acts $\succsim^C$ defined as*

$$(\rho \circ \phi)_\alpha \succsim^C (\rho \circ \phi)_{\alpha'} \iff \alpha \succsim^A \alpha'$$

is a weak order if and only if $\succsim^A$ satisfies Extended consequentialism.

[3] The axiom can thus be interpreted as an adaptation to an AA setting of a requirement of in conjoint measurement (Fishburn, 1970, p. 149).

All proofs are in Appendix B. Proposition 1 shows that Extended consequentialism is crucial to derive a ranking over acts from the primitive ranking over actions. Since the counterfactual consequence function is not observed, one viable option is to observe a strict preference for an action compared with another that induces the same material act. Chen & Schonger (2022) opt for this identification choice, relying on a simple theory where individuals have lexicographic preferences for "moral" actions. A second way consistent with Proposition 1 is to specify the counterfactual consequence function ϕ and structurally estimate a preference for specific counterfactual consequences. Miettinen et al. (2020) and van Leeuwen & Alger (2024) and take this route to investigate the presence of preferences for universalisation in lab experiments.

Linear Aggregation. I show next that the vNM axioms, Non-degeneracy, and Extended consequentialism characterise an individual who behaves as if he has two rankings over material and counterfactual acts. The same conditions guarantee that the two rankings are aggregated linearly. Moreover, if Separability holds, then the two rankings are the same. Theorem 1 links axioms to a functional representation of preferences over actions.

THEOREM 1. *The ranking $\succsim^A$ satisfies vNM, Non-degeneracy and Extended consequentialism if and only if there exist nonconstant functions $u^\rho, u^\phi : C \to \mathbb{R}$ and a probability distribution $\mu \in \Delta(S)$ such that, for all actions α, α',*

$$\alpha \succsim_A \alpha' \iff U(\alpha) \geq U(\alpha')$$

where

$$U(\alpha) = \sum_s \mu(s) \sum_c \rho_{\alpha,s}(c)\, u^\rho(c) + \sum_s \mu(s) \sum_c \phi_{\alpha,s}(c)\, u^\phi(c)\,. \tag{1}$$

Moreover, $\succsim^A$ satisfies vNM, Non-degeneracy, Extended consequentialism and Separability if and only if there exists a nonconstant function $u : C \to \mathbb{R}$ and $\lambda \in [0,1]$ such that for all actions α, α',

$$\alpha \succsim_A \alpha' \iff U(\alpha) \geq U(\alpha')$$

where

$$U(\alpha) = (1-\lambda) \sum_s \mu(s) \sum_c \rho_{\alpha,s}(c)\, u(c) + \lambda \sum_s \mu(s) \sum_c \phi_{\alpha,s}(c)\, u(c)\,. \tag{2}$$

The functions u^ρ, u^ϕ and u are unique up to similar positive affine transformations and μ is unique.

Theorem 1 states that choices of mixed actions satisfying the axioms are consistent with the following utility function: when choosing α, the individual evaluates the probability that each state s realises according to his subjective belief μ; each state realisation induces two distributions over consequences, $\rho_{\alpha,s}$ and $\phi_{\alpha,s}$; each consequence is then evaluated according to the functions u^ρ, u^ϕ or u. The functional representation of $\succsim^A$ aggregates material and counterfactual concerns linearly.

I do not derive the form of u^ρ, u^ϕ and u; 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^ρ. The axioms here imply nothing about such shape. The representation allows the individual, as an example, to both exhibit preferences for universalisation, as captured by the shape of ϕ and, say, inequity aversion, as captured by u^ϕ. Then, in a game, the individual would choose the action that, if implemented by everyone else as well, satisfies his inequity averse preference. Theorem 1 clarifies that pro-social and non-consequentialist preferences are not exclusive. On the contrary, these two can coexist. In the next section, I provide foundations for game-theoretic notions of universalisation building on Theorem 1.

4 Preferences for universalisation

In this section, I study particular cases of the functional representation of preferences in Theorem 1. I introduce axioms to derive several versions of preferences for universalisation in games. For this aim, I must link individual decision problems to games. Unfortunately, how to construct such a link is not clear.[4] For the purpose of this paper, I proceed as follows. I define games assuming players have preferences over mixed actions profiles, as usual. I study decision problems in which the set of uncertain states for each player is the set of opponents' actions. I impose axioms on preferences over mixed actions in individual decision problems to obtain functional forms of preferences for universalisation and compare them with their game-theoretic counterpart in the literature. I use the same symbol $\succsim_i^A$ for both player's i preferences over mixed actions profiles in the game and his preferences over mixed actions in his decision problem. Restricting attention to two-player game suffices.

Definition 2. *A* ***normal-form game*** *is an ordered list* $\mathcal{G} = \left(\{1,2\}, C, \left(A_i, \succsim_i^A\right)_{i\in\{1,2\}}, \rho, \phi\right)$ *featuring:*[5]

- *set of players* $\{1,2\}$*;*
- *set of consequences* C*;*

[4] Mariotti (1995) raises the problem. Battigalli (1996) and Hammond (1998) provide further discussion. Perea (2024) develops a new promising approach building on context-dependent preferences by Gilboa & Schmeidler (2003).

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

- *finite set of actions A_i for each player i;*
- *material consequence function $\rho : A_i \times A_{-i} \to C$;*
- *counterfactual consequence function $\phi : A_i \times A_{-i} \to C$;*
- *player i's ranking over mixed actions $\succsim_i^A$.*[6]

As in the previous section, pure actions profile (a_i, a_{-i}) induces the material consequence $\rho_{a_i,a_{-i}}$. Mixed actions profile (α_i, α_{-i}) induces instead a probability distribution over material consequences $\rho_{\alpha_i,\alpha_{-i}} \in \Delta(C)$. The probability that any consequence c realises under mixed actions profile (α_i, α_{-i}) is $\rho_{\alpha_i,\alpha_{-i}}(c)$. The same holds for counterfactual distributions of consequences $\phi_{\alpha_i,\alpha_{-i}} \in \Delta(C)$. When playing the game $\mathcal{G}$, player i is facing the decision problem $\mathcal{D}_i = \left(A_i, A_{-i}, C, \rho, \phi, \succsim_i^A\right)$ as in definition 1. In all the following sections, I assume ρ and ϕ are the same for both players.

Preferences over mixed actions profiles are represented by a utility function that is consistent with equation 2. In other words,

$$U_i(\alpha_i, \alpha_{-i}) = (1-\lambda)\sum_{a_{-i}} \alpha_{-i}(a_{-i}) \sum_{c} \rho_{\alpha_i,a_{-i}}(c)\, u(c) + \lambda \sum_{a_{-i}} \alpha_{-i}(a_{-i}) \sum_{c} \phi_{\alpha_i,a_{-i}}(c)\, u(c) \tag{3}$$

for all i and (α_i, α_{-i}). These are the preferences $U_i(\alpha_i)$ individual i has in his decision problem, as in equation 2, when his beliefs coincide with α_{-i}. To study the relation between equilibria in games and decision problems, I introduce the notion of an optimal action, an action that is preferred to all available actions in the decision problem.

Definition 3. *A mixed action $\alpha_i \in \Delta(A_i)$ is **optimal** in decision problem $\mathcal{D}_i$ if it is maximal for the ranking $\succsim_i^A$, i.e., $\alpha_i \in \left\{\alpha_i' \in \Delta(A_i) \;\middle|\; \alpha_i' \succsim_i^A \alpha_i'' \text{ for all } \alpha_i'' \in \Delta(A_i)\right\}$.*

In the following subsections, I study conditions under which preferences over mixed actions are equivalent to various notions of universalisation. 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.

4.1 Homo Kantiensis and Simple Kantian equilibrium

In this section, I restrict attention to symmetric games, where $A_1 = A_2 = A$ and $U_1(\alpha_1, \alpha_2) = U_2(\alpha_2, \alpha_1)$ for all mixed actions profiles (α_1, α_2). Simple Kantian Equilibrium is defined as follows.

Definition 4. *An actions profile (α, α) constitutes a **Simple Kantian Equilibrium** (SKE) of the symmetric game $\mathcal{G}$ if, for all players i and actions α'*

[6] Action sets should be rich enough for Theorem 1 to hold. In any game, only a subset of these actions are feasible.

$$\sum_c \rho_{\alpha,\alpha}(c)\, u_i(c) \geq \sum_c \rho_{\alpha',\alpha'}(c)\, u_i(c)\,.$$

A symmetric mixed actions profile constitutes a Simple Kantian Equilibrium if it induces the best distribution over material consequences over all symmetric mixed action profiles. I show that a mixed action in a *SKE* profile is an optimal mixed action for *Homo Kantiensis* preferences.[7]

Definition 5. *A ranking over mixed actions $\succsim_i^A$ is a* ***Homo Kantiensis*** *(HK) preference if it is represented by*

$$U_i(\alpha) = \sum_c \rho_{\alpha,\alpha}(c)\, u_i(c)\,,$$

for all α.

When evaluating any mixed action α, a *Homo Kantiensis* only considers the distribution over material consequences induced when his opponent chooses α as well.

I first impose conditions on preferences in decision problems to obtain *HK* preferences. For the purpose, I derive an incomplete ranking over distributions of counterfactual consequences. Consider an action α inducing the same distribution over counterfactual consequences in any state, and therefore a constant counterfactual act $\phi_{\alpha,s} = \phi_{\alpha,s'}$ for all s, s'. For all distributions over consequences $\gamma, \gamma' \in \Delta(C)$, I define $\gamma \succsim_i^\phi \gamma' \iff \alpha \succsim_i^A \alpha'$ for all α, α' such that $\rho_\alpha = \rho_{\alpha'}$, $\phi_{\alpha,s} = \gamma$ and $\phi_{\alpha',s} = \gamma'$ for all s, s'. This definition allows me to introduce the next axiom.

Axiom 5. ***(Universalisation Counterfactual)*** *For all mixed actions α_i, α_i',*

$$\alpha_i \succsim_i^A \alpha_i' \iff \rho_{\alpha_i,\alpha_i} \succsim_i^\phi \rho_{\alpha_i',\alpha_i'}.$$

Universalisation Counterfactual imposes that any mixed action α_i is preferred to α_i' whenever the distribution over material consequences induced when both players choose α_i is preferred to the distribution induced when both players choose α_i', as measured by the ranking $\succsim_i^\phi$. The axiom allows characterising *HK* preferences.

Proposition 2. *The ranking over mixed actions $\succsim_i^A$ satisfies vNM, Non-degeneracy, Extended consequentialism and Universalisation Counterfactual if and only if it is a HK preference.*

Proposition 2 shows that when Universalisation Counterfactual complements the axioms in the first part of Theorem 1, then the individual acts as a *HK*. A corollary is that an optimal action in a decision problem with *HK* preferences is also part of a symmetric mixed action profile constituting a *SKE*.

[7] These are also equivalent to preferences in Laffont (1975).

Corollary 1. *Action α is optimal according to a HK preference if and only if it is part of a SKE profile in a symmetric game.*

These results allow 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.[8] 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.

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.[9] Outside contexts of long repeated interactions and adaptive dynamics, an action in a Nash Equilibrium profile is played by an expected utility maximiser holding correct conjectures about opponents' behaviour.[10] 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 preference but different optimisation protocols. If they choose differently in the same problem, this would be defined as them having different preferences.

I show with Corollary 1 that there is no need to rely on informal arguments regarding how individuals optimise. Behaviour consistent with *SKE* reveals a preference for more desirable counterfactual consequences. 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 opti-

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

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

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

misation 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 for material consequences u^{ρ} is the same as that for counterfactuals u^{ϕ}. Nevertheless, this equivalence does not imply the individual would be indifferent between receiving a monetary amount and acting to induce it as a counterfactual consequence. 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.[11]

Corollary 1 also offers a novel interpretation of 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 employing mixed actions is therefore the adherence to a non-consequentialist attitude.

One last point is that *SKE* does not require strategic stability or correctness of beliefs. There are no conditions on individuals' beliefs regarding their opponents for an action to constitute a *SKE*.

4.2 Homo Moralis

In this section, I exploit the representation in Theorem 1 to derive Homo Moralis preferences as a special case of equation 2. In the context of two-player symmetric games, Homo Moralis preferences are defined as follows.

Definition 6. *A ranking over mixed actions $\succsim_i^A$ is a **Homo Moralis** (HM) preference if it is represented by*

$$U_i(\alpha_i) = (1-\kappa)\sum_{a_{-i}} \mu_i(a_{-i}) \sum_c \rho_{\alpha_i,a_{-i}}(c)\, u_i(c) + \kappa \sum_c \rho_{\alpha_i,\alpha_i}(c)\, u_i(c)\,, \tag{4}$$

for all α_i.

A Homo Moralis maximises a convex combination between expected material payoff and expected material payoff in a counterfactual scenario where both individuals play his action. Contrary to *SKE*, *HM* is a preference, it does not require joint optimality.

A *HM* with $\kappa = 1$ is a *HK* and his preferences are represented by equation 5. In the general formulation with an intermediate κ, Universalisation Counterfactual must be relaxed. The next

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

axiom establishes that Universalisation Counterfactual must be satisfied only for actions that induce the same material act.

Axiom 6. *(**Partial Universalisation Counterfactual**) For all mixed actions α_i, α_i' such that $\rho_\alpha = \rho_{\alpha'}$,*

$$\alpha_i \succsim_i^A \alpha_i' \iff \rho_{\alpha_i,\alpha_i} \succsim_i^\phi \rho_{\alpha_i',\alpha_i'}.$$

Universalisation Counterfactual implies Partial Universalisation Counterfactual. The axiom allows obtaining the following result.

Proposition 3. *The ranking over mixed actions $\succsim_i^A$ satisfies vNM, Non-degeneracy, Extended consequentialism, Separability and Partial Universalisation Counterfactual if and only if it is a HM preference.*

A *HM* is not only interested in the universalisation counterfactual, 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. It is possible that a *HM* believes his opponent will act differently from him, allowing to both derive material and counterfactual payoff and having correct beliefs in Nash equilibrium. This marks the difference between universalisation and magical thinking (Daley & Sadowski, 2017).

The fact that the ranking over material and counterfactual consequences in *HM* is the same, by Separability, greatly simplifies testing Partial Universalisation Counterfactual, the key axiom. The utility function u can be estimated by observing choices of actions which are equivalent in the counterfactual dimension. Of course, if the analyst does not know ϕ, this is not an easy task, but under the assumption that the counterfactual is related to others' behaviour observing choices over bets suffices. Because the ranking in the material and counterfactual dimensions is the same, Partial Universalisation Counterfactual can be tested by observing choices of actions that are equivalent in the material dimension. Therefore, Proposition 3 offers guidance on how to investigate for the presence of *HM*, and therefore *HK*, under ancillary assumptions, with a simple test.

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 5.

4.3 Multiplicative Kantian equilibrium

Here I employ Theorem 1 to study the relationship between optimality in decision problems and Multiplicative Kantian Equilibrium.[12] 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, but the extension of Theorem 1 to such a setting would come at a technical cost that bears no conceptual benefits. 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 symmetric games and assume players only have two pure actions available. 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.

Definition 7. *An actions profile (α_i, α_{-i}) constitutes a **Multiplicative Kantian Equilibrium** (MKE) of the game $\mathcal{G}$ if, for all players i and real numbers $r \geq 0$*

$$\sum_c \rho_{\alpha_i,\alpha_{-i}}(c)\, u_i(c) \geq \sum_c \rho_{r\cdot\alpha_i, r\cdot\alpha_{-i}}(c)\, u_i(c)\,. \tag{5}$$

A mixed actions profile constitutes a Multiplicative Kantian Equilibrium if it induces the best distribution over material consequence over all mixed action profiles that can be obtained by multiplying both actions by r.[13] 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.

The difference between *MKE* and *SKE* is the counterfactual consequence function, illustrated in Figure 1. An individual envisioning the *SKE* counterfactual only conceives both players choosing the same action. In the context of two-player symmetric games with two actions, these profiles correspond to the diagonal of the square representing mixed actions. Instead, *MKE* actions are multiplicative deviations from a specific profile. Counterfactual evaluations 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 counterfactuals are identical.

[12] An equivalent analysis delivers similar results for Additive Kantian Equilibrium (Roemer, 2019).

[13] According to this definition, $(0, 0)$ is always a *MKE* if it is available.

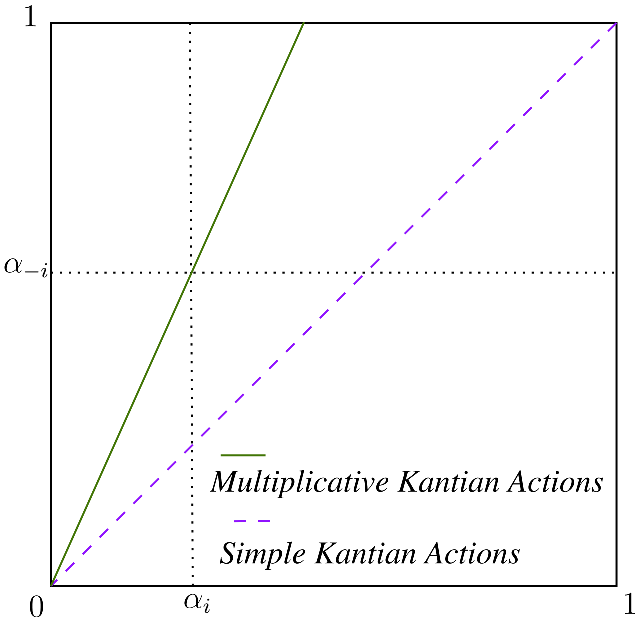

Figure 1: Counterfactual action profiles of Simple and Multiplicative Kantian Equilibria.

Paralleling the analysis of *SKE*, I show that mixed actions in a *MKE* profile are optimal for Multiplicative Homo Kantiensis preferences, a novel concept, defined as follows.

DEFINITION 8. *A ranking over mixed actions $\succsim_i^A$ is a **Multiplicative Homo Kantiensis (MHK)** preference relative to the profile (α_i, α_{-i}) if it is represented by*

$$U_i\left(r \cdot \alpha_i\right) = \sum_c \rho_{r \cdot \alpha_i, r \cdot \alpha_{-i}}\left(c\right) u_i\left(c\right),$$

for all real numbers r.

A Multiplicative Homo Kantiensis considers the multiplicative counterfactual of *MKE*, instead of the symmetric one of *HK*. The following axiom allows deriving MKE preferences.

AXIOM 7. *(**Multiplicative Universalisation Counterfactual**) For all numbers r, r',*

$$r \cdot \alpha_i \succsim_i^A r' \cdot \alpha_i' \iff \rho_{r \cdot \alpha_i, r \cdot \alpha_{-i}} \succsim_i \rho_{r' \cdot \alpha_i, r' \cdot \alpha_{-i}},$$

for one profile (α_i, α_{-i}).

Proposition 4 links Multiplicative Universalisation Counterfactual to *MHK*.

PROPOSITION 4. *The ranking over mixed actions $\succsim_i^A$ satisfies vNM, Non-degeneracy, Extended consequentialism, and Multiplicative Universalisation Counterfactual if and only if it a MHK preference.*

As for *SKE*, *MKE* is obtained from optimality in a decision problem in which individuals have *MHK* preferences.

COROLLARY 2. *If both players have MHK preferences relative to the profile (α_i, α_{-i}) and α_i, α_{-i} are optimal according to $\succsim_i$ and $\succsim_{-i}$, then (α_i, α_{-i}) is a MKE.*

Corollary 2 reveals that for a profile to constitute a *MKE*, it is required that players have the same counterfactual consequence function, i.e., that they consider the same reference profile. In *SKE*, this is guaranteed because only the diagonal of the square is relevant and there is no possibility of mismatch. This begs the question of how players coordinate on a specific line. I argue that *MKE* can be reached if individuals construct their counterfactual consequence functions from beliefs.

Say that player i has belief μ_i, fix an action α_i and consider the counterfactual distribution over consequences $\phi_{r \cdot \alpha_i, \mu_i} = \rho_{r \cdot \alpha_i, r \cdot \mu_i}$ for all real numbers r. Player i takes (α_i, μ_i) as a reference point and evaluates deviations given their distance from it, as measured by r. If beliefs are not correct, he mis-coordinates with his opponent. Consider the example in Figure 2, where players have lines of different slopes as counterfactuals. Player 1 believes his opponent will pick α_2, and α_1 is such that the profile is preferred to any other on the line induced by the counterfactual ϕ_1. Player 2 believes 1 will choose α_1', and (α_1', α_2') is his favourite profile compared to the counterfactuals on the line passing thorough it from the origin. Then, α_1 and α_2' are optimal actions, but do not necessarily constitute a *MKE*.

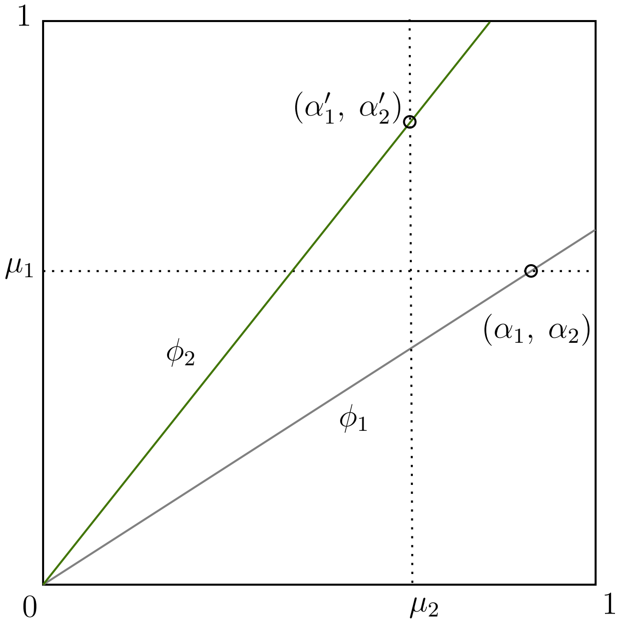

Figure 2: Multiplicative counterfactuals under different reference profiles.

The analysis reveals that one of the main differences between *MKE* and *SKE* is that the former, contrary to the latter, requires correctness of beliefs. Such a requirement is reminiscent of the choice-theoretic requirements for Nash Equilibrium.

Contrary to *SKE*, *MKE* can be defined outside symmetric settings 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 asymmetric games.

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*, and hence to *HK* for purely non-consequentialist individuals.

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 material consequence of the game as a currency. In game theory, this is usually players' *vNM* 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 material consequences.

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

Table 2: Prisoners' dilemma.

Row player 1 obtains his highest material reward in the consequence $(3, 0)$, induced by the profile (a_1', a_2). I define α_1^k as the mixed action which, compared to a_1', given a_2, induces a reduction in expected material payoff by k, i.e., $3 - \left[2\alpha_1^k + 3\left(1 - \alpha_1^k\right)\right] = k$. The profile leading to the greatest material reward for column player 2 is instead (a_1, a_2'). As for player 2, his mixed actions α_2^k induces the difference $3 - \left[2\alpha_2^k + 3\left(1 - \alpha_2^k\right)\right] = k$. Assume player 1, when picking any action α_1^k, considers the counterfactual where 2 chooses α_2^k. 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 material payoff. In this game, $\alpha_1^k = \alpha_2^k = k$ for all k. Whenever they deviate from the action profile that yields their preferred material outcome, both players consider a counterfactual scenario in which their opponents also deviate by choosing the same action. Therefore the counterfactual is identical to the one of *HK* and *HM*. The actions that lead to the highest reward under this counterfactual evaluation are a_1 for 1 and a_2 for 2, i.e., $\alpha_1^k = \alpha_2^k = 1$, 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 implementing his preferred material consequence. Second, he evaluates each action considering the induced difference in material 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, I consider general equal sacrifice rules. The results and arguments in this section hold for any equal sacrifice rule. Under the assumption that preferences over action profiles can be represented by equation 3, one can denote with $\left(\alpha_i^*, \alpha_{-i}^*\right)$ a profile that induces the maximal material consequence for player i, i.e., that is maximal according to the ranking represented by $\sum_{a_{-i}} \alpha_{-i}\left(a_{-i}\right) \sum_c \rho_{\alpha_i, a_{-i}}(c)\, u(c)$. Actions α_i^k leading to absolute sacrifice k satisfy:

$$\sum_{a_{-i}} \alpha_{-i}^*\left(a_{-i}\right) \sum_c \rho_{\alpha_i^*, a_{-i}}(c)\, u(c) - \sum_{a_{-i}} \alpha_{-i}^*\left(a_{-i}\right) \sum_c \rho_{\alpha_i^k, a_{-i}}(c)\, u(c) = k. \tag{6}$$

ESU is then defined as follows.

Definition 9. *A ranking over mixed actions $\succsim_i^A$ is a* ***Equal Sacrifice Universalisation*** *preference if it is represented by*

$$U_i\left(\alpha_i^k\right) = \sum_c \rho_{\alpha_i^k, \alpha_{-i}^k}(c)\, u_i(c), \tag{7}$$

for all real numbers k.

When choosing α^k, the individual evaluates the scenario where his opponent deviates from his optimal profile to induce the same sacrifice. Neither $\left(\alpha_i^*, \alpha_{-i}^*\right)$ nor α_i^k are guaranteed to be unique. Of course, it is possible to consider convex combinations between material and counterfactual payoffs, as in *HM*. I do not axiomatize *ESU*, which would require introducing a structure on the set of consequences to evaluate sacrifices. I conjecture that under such a structure, an axiomatic analysis similar to that in previous sections could characterize *ESU*.

The key difference between *ESU* and previous concepts is that it does not assume the structure of the counterfactual evaluation. Rather, it depends on the game at hand. I illustrate this point in the battle of the sexes. Numbers are again Bernoulli utilities for material consequences.

1 \ 2	a_2	a'_2
a_1	$2, 1$	$0, 0$
a'_1	$0, 0$	$1, 2$

Table 3: Asymmetry.

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

Table 4: Common Actions.

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

Table 5: Symmetry.

Consider the game in Table 3 on the left and assume throughout that preferences in the material and counterfactual dimension are the same. 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 material payoff of both players is 2, in (a_1, a_2) and (a'_1, a'_2). The action α_1^k of player 1 inducing a sacrifice of k solves $2 - 2\alpha_1^k = k$ and hence $\alpha_1^k = \frac{2-k}{2}$. The equivalent for player 2 is $2 - (2 - 2\alpha_2^k) = k$ which implies $\alpha_2^k = \frac{k}{2} = 1 - \alpha_1^k$. The optimum for *ESU* is reached at $k = \frac{2}{3}$ with $\alpha_1^k = \alpha_2^k = \frac{1}{2}$ which, if picked by both players, leads to a common expected material 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 re-labelling of actions from Table 3 to 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 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 $\left(\frac{1}{2}, \frac{1}{2}\right)$, i.e., the optimal actions of *ESU*. Not only the optimal profile coincides, but also the set of profiles considered in the counterfactual evaluation is identical. The re-labelling of actions from the first to the third table amounts to changing any mixed action α_2^k to $1 - \alpha_1^k$, which leads to $a_2 = a'_1$ and $a'_2 = a_1$ and switching columns. This is exactly the *ESU* counterfactual.

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. It holds for any equal sacrifice rule, not only absolute sacrifice, as shown in the proof in Appendix B.

Proposition 5. *Assume the game $\mathcal{G}$ is symmetric. Then, if an action is optimal under ESU, it is also optimal under HK.*

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. Future research might explore the correspondence between equal sacrifice rules and counterfactuals. This would constitute a step forward in the comprehension of the "ethos" implementing specific conceptions of justice, a line of research suggested by Maniquet (2019).

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.[14]

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 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.[15] When this is not enough, constrained versions of equal sacrifice, developed by Stovall (2020), can be employed.

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

[15] 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.

6 Conclusion

I have developed a choice-theoretic model to account for non-consequentialist preferences for universalisation. I derived a representation theorem for preferences that evaluate counterfactual scenarios from a collection of standard axioms. I also highlighted the conceptual difference between non-consequentialist and pro-social attitudes. Then, I complemented the axioms to derive preferences for universalisation. 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 decision problem. 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 counterfactual 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. What, then, is the difference between universalisation and consequentialist pro-social attitudes? John 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. It may be welfarist in the evaluation of the counterfactual consequence, but, as I have shown, this is not necessary.

Appendix

A Equal sacrifice in games

I map normal-form games to claim problems.[16] 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 $\left(I, (x_i)_{i \in I}\right)$ where $I = \{1, 2\}$ 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 material consequences he can obtain, denoted $\overline{U_i^\rho}$. Therefore, $x_i = \overline{U_i^\rho}$ 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 $k \in \mathbb{R}_+$ is

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

As an example, the equal loss rule $\pi_R\left(x_i, x_{-i}\right) = \left(x_i - k\right)_{i \in I}$ 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

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

for all $\left(\alpha_i, \alpha_{-i}\right)$. If it exists, an actions profile inducing the maximal expected utility for i is $\left(\alpha_i^*, \alpha_{-i}^*\right)$ so that $U_i^\rho\left(\alpha_i^*, \alpha_{-i}^*\right) = \overline{U_i^\rho}$. Then, action α_i^{Rk} induces sacrifice k relative to the function R if

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

A player exhibits equal sacrifice universalisation with respect to R if his counterfactual function is $\phi_{\alpha_i^{Rk}, a_{-i}} = \rho_{\alpha_i^{Rk}, \alpha_{-i}^{Rk}}$ for all α_i, a_{-i} and sacrifice k.

The profiles $\left(\alpha_i^*, \alpha_{-i}^*\right)$ inducing the maximal expected utility are not unique in general. The same holds for deviations α_i^{Rk}. Therefore, without further assumptions, an equal sacrifice rule is not uniquely related to a counterfactual consequence function.

B Proofs

Proof of Proposition 1. First, I prove that if $\succsim_A$ is a weak order and satisfies Extended consequentialism, then $\succsim^C$ is a weak order over AA acts. Consider the pairs of mixed actions (α, β) and (α', β') that induce the same acts $f, f' \in \Delta\left(C \times C\right)^S$:

[16] 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 the counterfactual envisioned by players, not to distribute a given endowment.

$$(\rho \circ \phi)_\alpha = (\rho \circ \phi)_\beta = f;$$
$$(\rho \circ \phi)_{\alpha'} = (\rho \circ \phi)_{\beta'} = f'.$$

I assumed the action set is rich enough for these actions to exist. By Extended consequentialism, $\alpha \sim^A \beta$ and $\alpha' \sim^A \beta'$. Transitivity of $\succsim^A$ implies $\alpha \succsim^A \alpha' \iff \beta \succsim^A \beta'$. For all acts, define $f \succsim^C f'$ as in the statement of the proposition: $f \succsim^C f'$ if actions α, α' exist such that $(\rho \circ \phi)_\alpha = f, (\rho \circ \phi)_{\alpha'} = f'$ and $\alpha \succsim^A \alpha'$. The ranking $\succsim^C$ is a weak order because $\succsim_A$ is a weak order.

Second, I show that if $\succsim^C$ is a weak order then Extended consequentialism is satisfied. I proceed by contrapositive, i.e., I prove that if $\succsim_A$ does not satisfy Extended consequentialism, then $\succsim^C$ is not a weak order. If Extended consequentialism does not hold, then there exist actions α and α' that induce the same act f such that $\alpha \succ^A \alpha'$. Then, by definition $f \succ^C f$, which violates reflexivity and hence completeness of $\succsim^C$. □

Before proceeding to the proof of Proposition 1, I show the independence of Extended consequentialism and Separability. First, Extended consequentialism does not imply Separability, as the former only disciplines preferences between actions that induce the same material and counterfactual acts, contrary to the latter. Second, Separability does not imply Extended consequentialism. I provide an example where the first is satisfied but the second is not. Consider two actions α, α' inducing the same acts (f, f'). Assume $\alpha \succ^A \alpha'$, which is possible because Extended consequentialism is not required. Separability only implies $f \succsim f$ and $f' \succsim f'$.

Proof of Theorem 1. That representations 1 and 2 imply the hypotheses is easy to check. I focus on the other direction of the statement.

Part 1. The ranking $\succsim^A$ satisfies vNM, Non-degeneracy, and and therefore, by Theorem 4 in Battigalli et al. (2017), it is represented by a nonconstant function $v : C \times C \to \mathbb{R}$ and a probability distribution $\mu \in \Delta(S)$ such that

$$U(\alpha) = \sum_s \mu(s) \sum_a \alpha(a)\, v\left((\rho \circ \phi)_{a,s}\right) \text{ for all } \alpha, \tag{8}$$

where v is unique up to positive affine transformations and μ is unique.

I now show that Extended consequentialism allows restricting v functional form. Consider a point $(c_0, c_0') \in C \times C$ and define $u^\rho, u^\phi : C \to \mathbb{R}$ so that

$$u^\rho(c_0) + u^\phi(c_0') = v(c_0, c_0') \tag{9}$$

$$u^\rho(c) = v(c, c_0') - u^\phi(c_0') \text{ for all } c \in C \tag{10}$$

$$u^\phi(c') = v(c_0, c') - u^\rho(c_0) \text{ for all } c' \in C. \tag{11}$$

Adding equations 10 and 11

$$u^{\rho}(c) + u^{\phi}(c') = v(c, c'_0) + v(c_0, c') - u^{\rho}(c_0) - u^{\phi}(c'_0).$$

By equation 9

$$u^{\rho}(c) + u^{\phi}(c') = v(c, c'_0) + v(c_0, c') - v(c_0, c'_0). \tag{12}$$

Now, consider two acts inducing in each state s the following distributions over pairs of consequences

$$\frac{1}{2}(c_s, c'_s) + \frac{1}{2}(c_0, c'_0) \text{ and } \frac{1}{2}(c_s, c'_0) + \frac{1}{2}(c_0, c'_s).$$

These acts induce distributions having the same marginals in each state

$$\left(\frac{1}{2}c_s + \frac{1}{2}c_0, \frac{1}{2}c'_s + \frac{1}{2}c'_0\right) \text{ and } \left(\frac{1}{2}c_s + \frac{1}{2}c_0, \frac{1}{2}c'_0 + \frac{1}{2}c'_s\right).$$

By Extended consequentialism, the actions inducing these two acts must be indifferent and therefore

$$v(c_s, c'_s) + v(c_0, c'_0) = v(c_s, c'_0) + v(c_0, c'_s) \text{ for each } s. \tag{13}$$

Since the previous equality holds for each c_s, substituting 13 in 12 delivers

$$u^{\rho}(c) + u^{\phi}(c') = v(c, c').$$

From equation 8, for each action α

$$U(\alpha) = \sum_s \mu(s) \sum_a \alpha(a) u^{\rho}(\rho_{a,s}) + \sum_s \mu(s) \sum_a \alpha(a) u^{\phi}(\phi_{a,s}). \tag{14}$$

The functions u^{ρ} and u^{ϕ} inherit the cardinal uniqueness of v. There exist numbers $q > 0$ and r such that, for each c, c'

$$\begin{aligned} v'(c, c') &= qv(c, c') + r \\ &= qu^{\rho}(c) + qu^{\phi}(c') + r \end{aligned}$$

from which $u'^{\rho}(c) = qu^{\rho}(c) + r^{\rho}$ where $r^{\rho} = r + qu^{\phi}(c'_0) - u'_{\phi}(c'_0)$.

Part 2. Denote $U^{\rho}(\alpha) = \sum_s \mu(s) \sum_a \alpha(a) u^{\rho}(\rho_{a,s})$ and $U^{\phi}(\alpha)$ equivalently. Say $\alpha \succsim^A \alpha'$ with $\phi_\alpha = \phi_{\alpha'}$, $\rho_\alpha = f$ and $\rho_{\alpha'} = f'$. By equation 14, it must be the case that $U^{\rho}(f) \geq U^{\rho}(f')$. Now consider β such that $\rho_\beta = \rho_{\beta'}$, $\phi_\beta = f$ and $\rho_{\beta'} = f'$. By Separability, it must be the case that $U^{\phi}(f) \geq U^{\phi}(f')$. The above must hold for all acts f. Hence, U^{ρ} and U^{ϕ} represent the same ranking on the same domain and must thus be related by positive affine transformations so that $U^{\phi} = q^{\phi}U^{\rho} + h^{\phi}$. Fix $(1-\lambda)V = U^{\rho}$, then the representation follows with $\lambda = q^{\phi}$ where $u : C \to \mathbb{R}$. □

Proof of Proposition 2. By vNM, Non-degeneracy, Extended consequentialism and Theorem 1, the ranking $\succsim_i^\phi$ is represented by $\sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \phi_{\alpha_i,a_{-i}}\left(c\right) u_i^\phi\left(c\right)$ for a belief μ_i and a utility function u_i^ϕ. By Universalisation Counterfactual, $\alpha_i \succsim_i^A \alpha' \iff \rho_{\alpha_i,\alpha_i} \succsim_i^\phi \rho_{\alpha_i',\alpha_i'}$, and therefore

$$\begin{aligned}
\alpha_i \succsim_i^A \alpha_i' &\iff \sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \phi_{\alpha_i,a_{-i}}\left(c\right) u_i^\phi\left(c\right) \geq \sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \phi_{\alpha_i',a_{-i}}\left(c\right) u_i^\phi\left(c\right)\\
&\iff \sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \rho_{\alpha_i,\alpha_i}\left(c\right) u_i^\phi\left(c\right) \geq \sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \rho_{\alpha_i',\alpha_i'}\left(c\right) u_i^\phi\left(c\right)\\
&\iff \sum_c \rho_{\alpha_i,\alpha_i}\left(c\right) u_i^\phi\left(c\right) \geq \sum_c \rho_{\alpha_i',\alpha_i'}\left(c\right) u_i^\phi\left(c\right)
\end{aligned}$$

for all actions α_i, α_i', from which the result obtains. □

Proof of Corollary 1. A mixed action α is maximal according to a *HK* preference $\succsim_i^A$ if, for all mixed actions α'

$$\sum_c \rho_{\alpha,\alpha}\left(c\right) u_i\left(c\right) \geq \sum_c \rho_{\alpha',\alpha'}\left(c\right) u_i\left(c\right). \tag{15}$$

If a game is symmetric, if α is optimal for player i it is also optimal for player $-i$. Therefore, equation 15 holds for all players i and mixed actions α', which are the conditions for (α, α) to be a *SKE*. □

Proof of Proposition 3. By vNM, Non-degeneracy, Extended consequentialism, Separability and Theorem 1, the ranking $\succsim_i^\phi$ is represented by $\sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \phi_{\alpha_i,a_{-i}}\left(c\right) u_i\left(c\right)$ for a belief μ_i and a utility function u_i. By Partial Universalisation Counterfactual, for all α_i, α_i' such that $\rho_\alpha = \rho_{\alpha'}$ it holds that $\alpha_i \succsim_i^A \alpha_i' \iff \rho_{\alpha_i,\alpha_i} \succsim_i^\phi \rho_{\alpha_i',\alpha_i'}$, and therefore

$$\begin{aligned}
\alpha_i \succsim_i^A \alpha_i' &\iff \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \phi_{\alpha_i,a_{-i}}\left(c\right) u_i\left(c\right) \geq \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \phi_{\alpha_i',a_{-i}}\left(c\right) u_i\left(c\right)\\
&\iff \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{\alpha_i,\alpha_i}\left(c\right) u_i\left(c\right) \geq \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{\alpha_i',\alpha_i'}\left(c\right) u_i\left(c\right)\\
&\iff \sum_c \rho_{\alpha_i,\alpha_i}\left(c\right) u_i\left(c\right) \geq \sum_c \rho_{\alpha_i',\alpha_i'}\left(c\right) u_i\left(c\right).
\end{aligned}$$

By equation 2, for all α, α'

$$\begin{aligned}
\alpha_i \succsim_i \alpha_i' \iff &\left(1-\lambda\right) \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{\alpha_i,a_{-i}}\left(c\right) u\left(c\right) + \lambda \sum_c \rho_{\alpha_i,\alpha_i}\left(c\right) u\left(c\right) \geq\\
&\left(1-\lambda\right) \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{\alpha_i',a_{-i}}\left(c\right) u\left(c\right) + \lambda \sum_c \rho_{\alpha_i',\alpha_i'}\left(c\right) u\left(c\right),
\end{aligned}$$

from which the result obtains. □

Proof of Proposition 4. By vNM, Non-degeneracy, Extended consequentialism and Theorem 1, the ranking $\succsim_i^{\phi}$ is represented by $\sum_{a_{-i}} \mu\left(a_{-i}\right) \sum_c \phi_{\alpha_i, a_{-i}}(c)\, u_i^{\phi}(c)$. Fix a profile $\left(\alpha_i, \alpha_{-i}\right)$. By Multiplicative Universalisation Counterfactual, for all numbers r, r', it holds that $r \cdot \alpha_i \succsim_i^A r' \cdot \alpha_i \iff \rho_{r \cdot \alpha_i, r \cdot \alpha_{-i}} \succsim_i \rho_{r' \cdot \alpha_i, r' \cdot \alpha_{-i}}$, and therefore

$$\begin{aligned} r \cdot \alpha_i \succsim_i^A r' \cdot \alpha_i & \iff \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \phi_{r \cdot \alpha_i, a_{-i}}(c)\, u_i^{\phi}(c) \geq \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \phi_{r' \cdot \alpha_i, a_{-i}}(c)\, u_i^{\phi}(c) \\ & \iff \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{r \cdot \alpha_i, r \cdot \alpha_{-i}}(c)\, u_i^{\phi}(c) \geq \sum_{a_{-i}} \mu_i\left(a_{-i}\right) \sum_c \rho_{r' \cdot \alpha_i, r' \cdot \alpha_{-i}}(c)\, u_i^{\phi}(c) \\ & \iff \sum_c \rho_{r \cdot \alpha_i, r \cdot \alpha_{-i}}(c)\, u_i^{\phi}(c) \geq \sum_c \rho_{r' \cdot \alpha_i, r' \cdot \alpha'_{-i}}(c)\, u_i^{\phi}(c), \end{aligned}$$

from which the result obtains. □

Proof of Corollary 2. A mixed action α_i is maximal according to a *MHK* preference $\succsim_i^A$ relative to the profile $\left(\alpha_i, \alpha_{-i}\right)$ if, for all r

$$\sum_c \rho_{\alpha_i, \alpha_{-i}}(c)\, u_i(c) \geq \sum_c \rho_{r\alpha_i, r\alpha_{-i}}(c)\, u_i(c). \tag{16}$$

If equation 16 holds for all i, then $\left(\alpha_i, \alpha_{-i}\right)$ constitutes a *MKE*. □

I now prove a version of Proposition 5 that holds for all equal sacrifice rules.

Proof of Proposition 5. I employ the notation of Appendix A. 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

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. 5

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302. 1, 2, 3, 5, 12, 16

Alger, I., & Weibull, J. W. (2016). Evolution and kantian morality. *Games and Economic Behavior*, *98*, 56–67. 5

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory*, *185*, 104951. 5

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199–205. 2, 4, 8

Baccelli, J., & Mongin, P. (2021). Can redescriptions of outcomes salvage the axioms of decision theory? *Philosophical Studies*, 1–28. 24

Battigalli, P. (1996). Comment on Mariotti (1996). *The Rational Foundations of Economic Behaviour*, 149–154. 11

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. 4, 26

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*(1), 1–35. 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. 11

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. 5

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. 24

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. 7

Chen, D. L., & Schonger, M. (2022). Social preferences or sacred values? Theory and evidence of deontological motivations. *Science Advances*, *8*(19). 4, 10

Daley, B., & Sadowski, P. (2017). Magical thinking: A representation result. *Theoretical Economics*, *12*(2), 909–956. 4, 16

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*. 5

Dekel, E., & Siniscalchi, M. (2015). Epistemic game theory. In *Handbook of Game Theory with Economic Applications* (Vol. 4, pp. 619–702). Elsevier. 14

Dewatripont, M., & Tirole, J. (2024). The Morality of Markets. *Journal of Political Economy*, 000–000. 5

Dierks, K., Alger, I., & Laslier, J.-F. (2024). Does universalization ethics justify participation in large elections? *TSE Working Paper*. 5

Dizarlar, A., & Karagözoğlu, E. (2023). Kantian equilibria of a class of Nash bargaining games. *Journal of Public Economic Theory*, *25*(4), 867–891. 5

Ellingsen, T., & Mohlin, E. (2024). A model of social duties. *Working Paper*. 5

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley. 3, 9

Fleurbaey, M. (2008). *Fairness, responsibility, and welfare*. Oxford University Press. 3

Fleurbaey, M. (2019). Economic theories of justice. *Annual Review of Economics*, *11*, 665–684. 2

Fleurbaey, M., Kanbur, R., & Snower, D. J. (2024). An Analysis of Moral Motives in Economic and Social Decisions. *Working Paper*. 5

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, *1*(1), 60–79. 7

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. 4, 6, 11

Grillo, A. (2022). Ethical Voting in Heterogenous Groups. *Working Paper*. 5

Hammond, P. J. (1998). Subjective expected utility. *Handbook of utility theory*, *1*, 213–271. 11

Harsanyi, J. C. (1953). Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*, *61*(5), 434–435. 23

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*. 5

Juan-Bartroli, P., & Karagözoğlu, E. (2024). Moral preferences in bargaining. *Economic Theory*, 1–24. 5

Kalai, E., & Smorodinsky, M. (1975). Other solutions to Nash's bargaining problem. *Econometrica: Journal of the Econometric Society*, 513–518. 23

Laffont, J.-J. (1975). Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, *42*(168), 430–437. 3, 13

Laslier, J.-F. (2022). Universalization and altruism. *Social Choice and Welfare*, 1–16. 5, 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. 1

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, *92*(368), 805–824. 5

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley. 2, 4

Maniquet, F. (2019). Comments on John Roemer's first welfare theorem of market socialism. *Review of Social Economy*, *77*(1), 56–68. 23

Mariotti, M. (1995). Is Bayesian rationality compatible with strategic rationality? *The economic journal*, *105*(432), 1099–1109. 11

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. 1, 10

Mill, J. S. (1885). *Principles of political economy*. D. Appleton. 3, 20, 21

Perea, A. (2012). *Epistemic game theory: Reasoning and choice*. Cambridge University Press. 14

Perea, A. (2024). *From Decision Theory to Game Theory: Reasoning about Decisions of Others*. Work in Progress. 11

Ponthiere, G. (2023). Epictetusian rationality. *Economic Theory*, 1–44. 5

Ponthiere, G. (2024). Stoicism and the Tragedy of the Commons. *GLO Discussion Paper*. 5

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. 1, 5

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, *127*, 45–57. 5

Roemer, J. E. (2019). *How we cooperate*. Yale University Press. 2, 3, 5, 6, 7, 12, 14, 15, 16, 17, 22

Rohde, K. I. (2010). A preference foundation for Fehr and Schmidt's model of inequity aversion. *Social Choice and Welfare*, *34*(4), 537–547. 11

Sarkisian, R. (2017). Team Incentives under Moral and Altruistic Preferences: Which Team to Choose? *Games*, *8*(3), 37. 5

Sarkisian, R. (2021a). Optimal Incentives Schemes under Homo Moralis Preferences. *Games*, *12*(1), 28. 5

Sarkisian, R. (2021b). Screening Teams of Moral and Altruistic Agents. *Games*, *12*(4), 77. 5

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications. 2, 8

Sen, A. (1973). Behaviour and the Concept of Preference. *Economica*, *40*(159), 241–259. 2

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. 23

Shi, Y. (2024). Endogenous Social Minimum. *Working Paper*. 5

Sobrado, E. M. (2022). Taxing moral agents. *CESifo working paper series 9867*. 5

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

Thomson, W. (2013). Game-theoretic analysis of bankruptcy and taxation problems: Recent advances. *International Game Theory Review*, *15*(03), 1340018. 25

Thomson, W. (2019). *How to divide when there isn't enough*. Cambridge University Press. 25

van Leeuwen, B., & Alger, I. (2024). Estimating social preferences and Kantian morality in strategic interactions. *Journal of Political Economy Microeconomics*. 1, 3, 10

von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton University Press. 8

Young, H. P. (1988). Distributive justice in taxation. *Journal of Economic Theory*, *44*(2), 321–335. 3, 20, 21