A Foundation for Universalisation in Games

Enrico Mattia Salonia[*]

September 22, 2023

Abstract

In revealed preference theory, observed choice is interpreted as revealing tastes over the outcomes of the decision. Nevertheless, if a moral principle prescribes an act for reasons unrelated to its consequences, the inference drawn regarding preferences is misleading. In this paper, I study the behaviour of deontological decision makers who follow the moral principle of universalisation. I develop a decision theory for agents who value the impact of their choice in determining a counterfactual outcome they envision. Hence, the choice of action reveals a preference for counterfactual outcomes. I propose a unifying model based on my theory, inspired by the equal sacrifice principle. It can be specified to obtain the most prominent models of universalisation, compare them, highlight and arguably overcome their limitations.

> *"It is both possible and eminently desirable to treat decision theory, ethics, and game theory as special cases of the same general theory of rational behavior"*
>
> Harsanyi (1986)

1 Introduction

That agents evaluate actions according to their consequences is an ingrained assumption in economics (Arrow, 1951). Consequentialism entails a strong link between how decision makers rank actions and their associated consequences. The best action results in the most desirable

[*] Toulouse School of Economics. mattia.salonia@tse-fr.eu. I would like to thank Ingela Alger, Matteo Broso, Annalisa Costella, Franz Dietrich, Philippe De Donder, Mostapha Diss, Peter Hammond, Pau Juan Bartroli, Giacomo Rubbini, François Salanié, and Karine Van Der Straeten for helpful discussions and comments. I also thank worshop participants at TSE, 10^{th} Warwick PhD Conference, 2022 Young Economist Symposium, XVI GRASS Workshop, 1^{st} NSEF Workshop, the 18th Doctorissimes Conference, 12th Lisbon Meetings, 34th Stony Brook Conference on Game Theory, ESAM 2023. 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).

consequence. In light of this connection between action and outcome, economists can use preferences for one or the other as a primitive of a decision problem.

The assumption of consequentialism is not as restrictive as it appears. It is possible to expand the domain of consequences to consider a plethora of rationales for behaviour, including selfishness, pro-social attitudes and even emotions. There is, however, a category of decision makers that cannot be dealt with under the assumption that the desirability of an action is tied to the associated outcome. I refer to these as deontological agents. Deontological agents do not relate the optimality of an action to the goodness of its outcome, thus breaking the relation between the two. They are non-consequentialists. As Sen (1973, p. 251) argues, in the prisoners' dilemma,"... *the choice of non-confession follows not from calculations based on [a] welfare function, but from following a moral code of behaviour suspending the rational calculus. And it is this difference that is inimical to the revealed preference approach to the study of human behaviour.*"[1]

Fleurbaey (2019) acknowledges that economics has remained impervious to comprise non-consequentialist motivations. The classical decision-theoretic frameworks of Anscombe & Aumann (1963) and Savage (1972) illustrate this resistance. In these models, the agent ranks acts, functions from uncertain states to outcomes, according to the induced consequences. It is impossible to rank acts considering a criterion that does not evaluate their consequences without relabeling such a notion trivially, e.g. by including the chosen action in the description of consequences. The question is whether deontology can be reconciled with the consequentialist approach of standard decision theory without resorting to ad hoc solutions.

In this paper, I attempt to offer a positive answer for a specific deontological motivation, namely, a preference for universalisation. In symmetric games, an agent with preferences for universalisation considers optimal the action that he would prefer if implemented by everyone else as well.[2] As with other deontological criteria, optimality is not related to the goodness of consequences. Contrary to pro-social attitudes, universalisation cannot be rationalised by considering any feature of the material outcome. To carry on with the example of Sen (1973, p. 251): "*The prisoners' non-confession will be quite easy to put within the framework of revealed preference if it were the case that they had so much concern for the sufferings of each other that they would choose non-confession on grounds of joint welfare of the two. The problem arises precisely because that is not being assumed.*"

I maintain that non-consequentialist decision criteria must be studied by taking as a primitive a ranking over actions, not consequences. I employ the decision-theoretic framework under uncertainty of Luce & Raiffa (1957), where the object of choice is an element of the action set, not an outcome. The model developed here deals with a class of moral maxims that ranks actions based on their consequences in counterfactual scenarios. The focus of this paper is the universalisation counterfactual "what would happen were everyone behaving as I do?". I make explicit that each action induces both a material outcome, as in the standard model, and a counterfactual outcome, what would occur under a different realisation of an uncertain state

[1] In the words of Herbert Gintis at the International Meeting in Experimental and Behavioral Social Science (2018): "*Decision theory is wrong because it does not take into account non-consequentialist moral behaviour.*"

[2] For a discussion about the notion of universalisation in economics and Kantian philosophy, I direct the reader to the Erasmus Journal for Philosophy and Economics book symposium on Roemer (2019). I do not employ the term "Kantian", where possible, to avoid taking a stance on this relation.

of the world. The counterfactual consequence expands the outcome domain.

The first result of this paper, Proposition 1, traces the requirement for the ranking over actions of an agent to be equivalent to a ranking over couples of material and counterfactual consequences. The necessary and sufficient condition is a property I label Extended Consequentialism, requiring that two actions resulting in the same distribution of material and counterfactual consequences must be ranked equally. If a decision maker behaves as a consequentialist, with respect to an expanded domain of outcomes that does not exclusively pertain to its material features, his ranking over actions relates to a ranking over the extended consequences. The second result, Theorem 1, provides a functional representation of preferences over actions under the standard von Neumann & Morgenstern (vNM) axioms. Extended Consequentialism implies that the decision maker is not sensitive to the correlation structure between the lotteries of material and counterfactual consequences. Using an argument from the literature on conjoint measurement (Fishburn, 1970), I show that the lack of preference for correlation guarantees that material and counterfactual concerns are aggregated additively.

Applications of Theorem 1 provide a decision-theoretic rationale for Homo Moralis preferences for universalisation à la Alger & Weibull (2013) and the various definitions of Kantian Equilibrium by Roemer (2019), both constituting a generalisation of the classical model by Laffont (1975). Since the axioms are testable, I provide guidance for future empirical investigations in this literature. I comment on the difference between my foundation for Kantian Equilibrium and that of Roemer. I argue that his distinction between optimisation protocol and preferences is inconsistent with revealed preference theory, but also not needed.

I develop a novel concept of universalisation inspired by the equal sacrifice principle (Young, 1988). Consider an agent with any aim. Given a profile of actions, he evaluates a deviation by considering the outcome that would obtain if his opponents also deviated to induce an equivalent difference in aim satisfaction, that is, an equal sacrifice. I show that this preference is equivalent to Homo Moralis and Simple Kantian Equilibrium in symmetric games. By contrast, its definition does not rely on the label of actions or requires the veil of ignorance of Homo Moralis to be defined in asymmetric contexts.

I conclude by defending my modelling choices from a methodological perspective and elaborating on the implications of my results for the nature preferences for universalisation.

Related literature. This paper builds on an observation by Battigalli et al. (2017). The authors argue that Luce & Raiffa (1957)'s framework for choice under uncertainty constitutes a natural way to interpret mixed actions, ubiquitous in game theory, and show under which conditions it is equivalent to the more tractable model of Anscombe & Aumann (1963). The requirement guaranteeing equivalence of the two theories is an assumption of consequentialism. The agent 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 outcomes. The model in this paper can be viewed as a two-dimensional generalisation of Battigalli et al. (2017), except that I assume the existence of objective probability.

A second decision-theoretic work on consequentialism is by Hammond (1988) and subsequent papers. Hammond studies the implications of assuming a form of consequentialism in extensive forms. He shows that it implies expected utility maximisation. Contrary to his work, I do not elaborate on consequentialism and expected utility in dynamic settings.

Gilboa & Schmeidler (2003) provide axioms for context-dependent preferences in games modelled as decision problems. They study collections of agents' 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 actions and obtain an expected utility representation in games. I discuss in detail the relation between my work and theirs in the motivating example in Section 1.1.

A concept related to universalisation is magical thinking, studied from a decision-theoretic point of view by Daley & Sadowski (2017). An agent exhibits magical thinking if he expects the probability the opponent selects a specific action to increase if he chooses that action. They provide axioms equivalent to assuming agents exhibit magical thinking in symmetric games. Magical thinking and universalisation are different from a decision-theoretic perspective. An agent with preferences for universalisation does not believe he affects opponents' choice.

Bradley & Stefánsson (2017) build a decision-theoretic framework where agents care about counterfactual realisations of lotteries. Contrary to this paper, they rely on the Jeffrey (1990) desirability decision theory and a modal logic semantic, although they also study which assumptions are needed to obtain an equivalence between the maximisation of desirability and expected utility. A discussion on consequentialism is absent in their work.

Chen & Schonger (2022) develop a decision-theoretic model to guide an experimental design testing for the presence of deontological preferences. They argue that, to identify deontology from choice, subjects must face the possibility that their actions will not be implemented or observed by the experimenter. The intuition is that deontological agents care about the act itself, not its consequences. Their model has a different interpretation compared with mine. In their experiment, subjects knew that there was an objective probability that their action would not have been implemented, whereas in this paper, there is no such possibility.

In this paper, I provide a decision-theoretic foundation for various models of universalisation. The two main frameworks are Homo Moralis preferences by Alger & Weibull (2013, 2016); Alger et al. (2020) and Kantian Equilibrium by Roemer (2010, 2015, 2019).

Homo Moralis, in two-player games, 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 utility functions, Homo Moralis is the only evolutionary stable one in games featuring incomplete information about preferences and assortativity in the matching 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 agents are Kantian rather than Nash optimisers, they behave following the Kantian Equilibrium logic, i.e. choose the action that maximises their payoff if adopted by everyone else as well. Alger & Weibull change preferences and Roemer changes the equilibrium concept, when compared with standard selfish-Nash agents. I comment on the relation between these two models in the body of the paper.

There are several papers in economics investigating universalisation and other deontological motivations. Some of these study moral attitudes or their relation with pro-social preferences, as Dewatripont & Tirole (2022), Fleurbaey et al. (2021), Kordonis (2020), Laslier (2022) and Long (2020). Others are applications in economic environments, including bargaining (Dizarlar & Karagözoğlu, 2023), contract theory (Sarkisian, 2017, 2021a,b), public goods (Brekke et al., 2003), taxation (Eichner & Pethig, 2020; Sobrado, 2022), vaccination (De Donder et al., 2021) and voting (Alger & Laslier, 2021a,b; Grillo, 2021).

1.1 A MOTIVATING EXAMPLE

I concisely illustrate the contribution of this paper through an example. I show the issues arising with a deontological preference if the outcome domain is exclusively the material consequence of action. Then, I discuss the solution I propose and how it relates to the existing literature.

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

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

Assume the row player has beliefs μ, highlighted in blue. He conjectures his opponent will play ℓ or m with probability $\frac{1}{2}$. By choosing a mixed action, the row player can induce any distribution over outcomes 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 outcome $(0,0)$. Now, consider a second scenario where the same decision maker has beliefs p', in red in the table, according to which his opponent plays m or r with probability $\frac{1}{2}$. The feasible set of lotteries over outcomes is the same as before. Actions m and r induce the midpoint between $(0,0)$ and $(1,1)$ whereas ℓ induces the sure outcome $(0,0)$. 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 preference reversal in the second scenario, thus violating the weak axiom of revealed preference. Therefore, it is not possible to obtain a rational preference relation from observed choice, regardless of any physical feature of the game at hand. This impossibility does not occur for consequentialist preferences defined on the distribution of material rewards, such as selfishness, altruism, inequity aversion, or maximin. Nevertheless, there are functional forms for preferences for universalisation in the literature, meaning that these represent an order over an object that is different from the material outcome.

Let me distil two points from this example. First, it might be argued that the outcome domain is too narrow, and hence, the impossibility I showed is artificial. Second, the analysis hinges on the description of the problem. It would be sufficient to relabel the actions in the game to "coordinate on (ℓ,ℓ)", "anti-coordinate on (m,r)" and "anti-coordinate on (r,m)". In the next paragraphs, I discuss how I deal with these two points.

As for the first, that the mere description of the material outcome, in this example monetary rewards, does not contain enough information to incorporate all the relevant features of a decision problem is not new. Context-dependent preferences by Gilboa & Schmeidler (2003) can easily handle such an issue. The authors derive an expected utility representation for preferences ranking couples of action and state realisation, 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 $(\ell, \cdot)$ is preferred to both $(m, \cdot)$ and $(r, \cdot)$. Although this works, such a general framework is silent on the determinants of preferences and makes it difficult to test a particular hypothesis, such as the presence of deontological motivations. One solution is to have a less general framework that specifies the relevant features of the decision problem. For example, in psychological games, it is explicit that preferences depend on both material outcomes and players' beliefs (Geanakoplos et al., 1989; Battigalli & Dufwenberg, 2009). With preferences for reciprocity, agents return the favour to their previously altruistic opponents (Charness & Rabin, 2002). In these cases, beliefs or previous actions are part of the outcome. The primitives of these models can be elicited or observed. I take a similar route, obtaining a representation theorem for preferences for universalisation whose unique primitives are observable material features of the outcome. Such a modeling choice gives empirical guidance, as I discuss in Section 3. I defend again this methodological direction in Section 6.

The second point is that universalisation relies on the description of the game. Indeed, there are multiple instances where Roemer (2019) discusses how to change the label of actions to define and employ universalisation. The solution concept is sensitive to relabeling. In Section 5, I present a novel definition of universalisation, relying on the general theory, that is equivalent under any redescription of the game.

2 Preliminaries

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

Definition 1. *A **one-agent decision problem** is an ordered list $\mathcal{D} = (A, S, C, \mu, \rho, \phi)$, where:*

- *A is the finite set of actions;*
- *S is the finite set of uncertain states of the world;*
- *C is the finite set of consequences;*
- *$\mu \in \Delta(S)$ is a prior belief over states;*
- *$\rho : A \times S \to C$ is the material consequence function, mapping actions and state realisations to **material** consequences (ρ for real);*
- *$\phi : A \times S \to C$ is the counterfactual consequence function, mapping actions and state realisations to **counterfactual** consequences (ϕ for fictitious).*

Since A, S and C are finite, the sets of probability distributions $\Delta(A), \Delta(S)$ and $\Delta(C)$ are mixture sets (Herstein & Milnor, 1953). Hence, I can perform the standard mixing operations between their elements. The material consequence function maps actions and realisation of uncertain states of the world to consequences. It is related to the nature of the problem under study. The counterfactual consequence function is instead a primitive feature of the decision maker. It describes the link between his behaviour and a counterfactual consequence he envisions.

Each pair of pure action $a \in A$ and state realisation $s \in S$ induces a couple of material and counterfactual consequences $(c, c') = (\rho(a, s), \phi(a, s))$ where $c, c' \in C$. Given μ and ρ, the choice of a is equivalent to selecting a finite probability distribution over material consequences, which, with an abuse of notation, I denote $\rho(a, \mu) \in \Delta(C)$. For all consequences, $\rho(a, \mu)[c] = \mu(\{s \in S \mid \rho(a, s) = c\})$. Equivalently, an action induces a distribution $\phi(a, \mu) \in \Delta(C)$ of counterfactual consequences. Mixed actions $\alpha \in \Delta(A)$ induce lotteries over consequences $\rho(\alpha, \mu) \in \Delta(C)$ that are compositions of μ and the probability of realisation of pure actions in the support of the mix. I assume there exist enough mixed actions to induce every possible couple of distributions of consequences.[3] This is a standard richness assumption.

The agent is endowed with a ranking over mixed actions $\succsim^A \subset \Delta(A) \times \Delta(A)$, with symmetric $\sim^A$ and asymmetric $\succ^A$. He can delegate his choice to a random device that selects a pure action according to an objective probability distribution.

The consequence functions allow me to study under which conditions an action can be interpreted as a revealed preference for consequences. This link would be impossible to investigate in the standard frameworks of von Neumann & Morgenstern (2007) or Anscombe & Aumann (1963) and Savage (1972). The reason is the following. In the von Neumann & Morgenstern model, agents choose lotteries over consequences directly. The relation between action and outcome is omitted. This reduced form model is not explicit enough to study whether one values an action regardless of its consequences. As for Anscombe & Aumann and Savage, the objects of choice are (mixed) acts, functions from states to consequences, in a subjective probability framework. In the language of this paper, a Savage act is the section at A of the material consequence function, i.e. $\rho_a : S \to C$. This model is richer than von Neumann & Morgenstern, but still collapses the relation between action and outcome.

Assumptions. I assume $\succsim^A$ satisfies the standard von Neumann & Morgenstern requirements.

vNM. *The ranking $\succsim^A$ satisfies the vNM axioms if, for all actions $\alpha, \alpha', \alpha'' \in \Delta(A)$:*

1. *(**Weak Order**) it is complete and transitive;*
2. *(**Independence**) for all $\lambda \in (0, 1)$, $\alpha \succsim^A \alpha'$ if and only if $\lambda\alpha + (1 - \lambda)\alpha'' \succsim^A \lambda\alpha' + (1 - \lambda)\alpha''$;*
3. *(**Continuity**) the sets*

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

are closed.

The crucial axiom in my analysis imposes that the agent is indifferent between two actions that induce the same distributions of material and counterfactual consequences.

Extended Consequentialism. *The ranking $\succsim^A$ satisfies Extended Consequentialism if, for all actions $\alpha, \alpha' \in \Delta(A)$, if $(\rho(\alpha, \mu), \phi(\alpha, \mu)) = (\rho(\alpha', \mu), \phi(\alpha', \mu))$ then $\alpha \sim^A \alpha'$.*

[3] A sufficient condition for this to hold is that for each couple of consequences $(c, c') \in C \times C$ there exists an action $a \in A$ such that $(\rho(a, s), \phi(a, s)) = (c, c')$ for all $s \in S$.

Extended Consequentialism allows the agent to prefer an action to another even if these two induce the same distribution of material consequences. In models with consequentialist decision makers this possibility is ruled out. One could also be indifferent between actions that induce the same distribution of counterfactual consequences, but different material ones. As I show in Section 4, this is the case for agents who play according to the Kantian Equilibrium notion of Roemer (2019) and for the extreme case of Homo Moralis by Alger & Weibull (2013), Homo Kantiensis. I dub such a preference as purely deontological. Being purely consequentialist or deontological is consistent with Extended Consequentialism. In general, agents who care about both material and counterfactual consequences satisfy Extended Consequentialism without being purely consequentialist or deontological.

Extended Consequentialism is also a reduction condition. The agent cannot prefer a compound lottery that leads to the same distribution of consequences as a simple lottery. For this reason, I do not consider in my model compound lotteries over actions.

Lastly, Extended Consequentialism rules out preferences for the correlation structure between the two distributions,[4] a property that I use in the next section.

3 Functional Representation

From Actions to Consequences. My first result shows that the choice of action can be interpreted as revealing a preference for couples of distributions of material and counterfactual consequences if and only if Extended Consequentialism holds.

Proposition 1. *Assume the ranking $\succsim^A$ is a weak order. Then, $\succsim^A$ satisfies Extended Consequentialism if and only if there exists a weak order over couples of distributions of material and counterfactual consequences $\succsim^C$ such that, for all actions $\alpha, \alpha' \in \Delta(A)$,*

$$(\rho(\alpha,\mu), \phi(\alpha,\mu)) \succsim^C (\rho(\alpha',\mu), \phi(\alpha',\mu)) \iff \alpha \succsim^A \alpha' \ .$$

All proofs are in Appendix B. Proposition 1 shows that Extended Consequentialism is crucial to define a ranking over couples of material and counterfactual consequences linked to the primitive ranking over actions.

Since the decision maker cares about the counterfactual outcome of his choice, it is hard to empirically estimate deontological preferences via a variation of material features. Since the counterfactual consequence function is not observed, one viable option is to elicit a strict preference for an action compared with another that induces the same material consequence. Chen & Schonger (2022) opt for this identification choice, relying on a simple theoretical framework where agents have lexicographic preferences for "moral" actions. A second way consistent with Proposition 1 is to specify ϕ and structurally estimate a taste for specific counterfactual consequences. Alger & Van Leeuwen (2023) and Miettinen et al. (2020) take this route to investigate the presence of preferences for universalisation in lab experiments.

[4] A similar point has been made by Al-Najjar & Pomatto (2020), who show that the Pareto condition in social choice theory, close in spirit to Extended Consequentialism, rules out preferences for the distribution of aggregate risk. The axiom is also close to a requirement of preferences in conjoint measurement (Fishburn, 1970, p. 149).

Linear Aggregation. I show next that the vNM axioms on $\succsim^A$, together with Extended Consequentialism, imply that the agent must separate his tastes for material and counterfactual outcomes, i.e. he behaves as if he has two rankings in these dimensions. The same conditions guarantee that the two concerns are aggregated linearly.

It is easy to check that, due to Extended Consequentialism, continuity and independence of $\succsim^A$ imply that the induced ranking $\succsim^C$ satisfies the same properties. Hence, under these assumptions, $\succsim^C$ satisfies conditions analogous to the ones imposed on $\succsim^A$. Furthermore, as discussed in Section 2, Extended Consequentialism allows me to avoid worrying about compound lotteries. By standard results, $\succsim^C$ has an expected utility representation $U : \Delta(C) \times \Delta(C) \to \mathbb{R}$ with Bernoulli utility $u : C \times C \to \mathbb{R}$.[5]

Before going to Theorem 1, I introduce one more property of $\succsim^A$. For all distributions $\gamma \in \Delta(C)$, define $\gamma \succsim \gamma' \iff \alpha \succsim^A \alpha'$ for every $\alpha, \alpha' \in \Delta(A)$ such that $\rho(\alpha, s) = \phi(\alpha, s) = \gamma$ and $\rho(\alpha', s) = \phi(\alpha', s) = \gamma'$, an incomplete order over distributions in $\Delta(C)$.

Separability. *The ranking $\succsim^A$ satisfies Separability if the following two conditions hold:*

1. *if $\phi(\alpha, \mu) = \phi(\alpha', \mu)$, $\rho(\alpha, \mu) = \gamma$ and $\rho(\alpha', \mu) = \gamma'$ then $\gamma \succsim \gamma' \iff \alpha \succsim^A \alpha'$;*
2. *if $\rho(\alpha, \mu) = \rho(\alpha', \mu)$, $\phi(\alpha, \mu) = \gamma$ and $\phi(\alpha', \mu) = \gamma'$ then $\gamma \succsim \gamma' \iff \alpha \succsim^A \alpha'$.*

Separability imposes that each dimension must agree on the ordering of distributions.[6] The following result clarifies what it implies, together with vNM and Extended Consequentialism.

Theorem 1. *The ranking $\succsim^A$ satisfies Extended Consequentialism and vNM if and only if there exist functions $u^\rho, u^\phi : C \to \mathbb{R}$ such that, for all actions $a \in A$,*

$$\sum_{s\in S} \mu(s)\, u(\rho(a,s), \phi(a,s)) = \sum_{s\in S} \mu(s)\, u^\rho(\rho(a,s)) + \sum_{s\in S} \mu(s)\, u^\phi(\phi(a,s)) \tag{1}$$
$$U(\rho(a,\mu), \phi(a,\mu)) = U^\rho(\rho(a,\mu)) + U^\phi(\phi(a,\mu))$$

where capital letters denote expected utility.

Moreover, if Separability holds too, there exist a function $v : C \to \mathbb{R}$ and weights $\lambda^\rho, \lambda^\phi > 0$ such that for all $a \in A$,

$$\sum_{s\in S} \mu(s)\, u(\rho(a,s), \phi(a,s)) = \lambda^\rho \sum_{s\in S} \mu(s)\, v(\rho(a,s)) + \lambda^\phi \sum_{s\in S} \mu(s)\, v(\phi(a,s)) \tag{2}$$
$$U(\rho(a,\mu), \phi(a,\mu)) = \lambda^\rho V(\rho(a,\mu)) + \lambda^\phi V(\phi(a,\mu)) \quad .$$

The functions U^ρ, U^ϕ and V are unique up to positive affine transformations.

Theorem 1 states that the functional representation of $\succsim^C$, and hence of $\succsim^A$, aggregates material and counterfactual concerns linearly. No other form is consistent in this setting.

I do not derive how u^ρ, u^ϕ and v depend on consequences. One may have any taste regarding the outcome. This fact makes clear the difference between my exercise and, as an

[5] See Fishburn (1970, p. 107) or Kreps (1988, p. 46).

[6] I show the independence between Separability and Extended Consequentialism in Appendix B.

example, that of Rohde (2010). In that paper, the author puts conditions on the ranking over material outcomes, defined as the monetary rewards for all agents in a game, to obtain inequity aversion. In my framework, she studies the shape of u^ρ. My result implies nothing about this feature. It allows the agent, as an example, to both have preferences for universalisation and, say, inequity aversion. Then, in a game, the agent would select the action that, if implemented by everyone else, satisfies his inequity aversion objective. Therefore, Theorem 1 clarifies that pro-social preferences and deontological attitudes are not exclusive. On the contrary, these two can coexist.

In the next section, I provide foundations for game-theoretic models of universalisation building from Theorem 1.

4 Preferences for universalisation

I model complete information games as collections of one-agent decision problems as in Definition 1. I am explicit about the relation between profiles of actions and payoffs to linking the traditional definition to my framework. Each profile of actions leads to a couple of material and counterfactual consequences via the consequence functions. The domain of utility is then the cartesian product of these two dimensions. Restricting attention to two-player games suffices to convey my points.

Definition 2. *A* ***normal-form game*** *is an ordered list* $\mathcal{G} = \Big(I, C, \rho, \big(A_i, \succsim_i^A, \phi_i, \big)_{i\in I}\Big)$ *where:*

- $I = \{1, 2\}$ *is the set of players;*
- A_i *is the finite set of actions of player* i*;*
- C *is the finite set of consequences;*
- ρ *and* ϕ_i *map profiles of actions to material and counterfactual consequences* $\rho, \phi_i : A_i \times A_{-i} \to C$*, for all* $i \in I$*;*
- $\succsim_i^A$ *is the ranking on actions of player* i*.*

Any complete information game is a collection of decision problems. The space of uncertain states for each player is exclusively the set of the opponent's actions. Fixing $\mathcal{G}$, the associated decision problem for player i is $\mathcal{D}_i = (A_i, A_{-i}, C, \mu_i, \rho, \phi_i)$. Given a belief μ_i, each mixed action $\alpha_i \in \Delta(A_i)$ leads to a lottery over material consequences denoted with $\rho(\alpha_i, \mu_i) \in \Delta(C)$. The same holds for counterfactual consequences $\phi_i(\alpha_i, \mu_i)$. I abuse notation by omitting the index i for the function ρ, to make it clear that it is the same map for both players, except that from player i perspective the first argument is his strategy and the second is his belief.

In the following subsections, I derive sufficient conditions on the decision problem of each player leading to universalisation behaviour in the corresponding game.[7] Optimality in decision problems amounts to selecting the best mixed action according to the ranking $\succsim_i^A$.

[7] The link between individual decision problems and games has been thoroughly studied in Mariotti (1995) and Battigalli (1996). They show that, in a subjective expected utility framework, such a link is problematic. Since

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

I assume that both players satisfy vNM and Extended Consequentialism. Hence, optimality is equivalent to maximisation of expected utility in Equation 1.

In the following subsections, I study conditions under which various notions of universalisation in a game are equivalent to the optimality requirement in the related collection of decision problems. I start with the Simple Kantian Equilibrium concept by Roemer (2019), to later proceed with Homo Moralis preferences by Alger & Weibull (2013) and conclude with the Multiplicative Kantian Equilibrium by Roemer (2019). I accompany formal results with discussions on the interpretation of these concepts and the relation between them.

4.1 Simple Kantian Equilibrium

Assume the game $\mathcal{G}$ is symmetric, i.e. $A_1 = A_2 = A$, $u_1^\rho(\rho(a,a')) = u_2^\rho(\rho(a,a'))$ and $u_1^\phi(\phi_1(a,a')) = u_1^\phi(\phi_2(a,a'))$ for all $a, a' \in A$.[8] I can define the following.

Definition 4. *A mixed action $\alpha \in \Delta(A)$ constitutes a **Simple Kantian Equilibrium** (SKE) of the game $\mathcal{G}$ if, for each player i and action α',*

$$U_i^\phi(\rho(\alpha,\alpha)) \geq U_i^\phi(\rho(\alpha',\alpha')) \quad .$$

A mixed action constitutes a Simple Kantian Equilibrium if, in a counterfactual scenario where it is employed by both players, it leads to their payoff maximisation.

I show under which conditions optimality in decision problems $\mathcal{D}_i$ implies *SKE* actions in the game $\mathcal{G}$. First, players must exclusively consider counterfactual outcomes, i.e. be purely deontological. Following the discussion in Section 2, pure deontology entails that the decision maker is indifferent between two actions that induce the same counterfactual outcome.

Pure Deontology. *The ranking $\succsim_i^A$ satisfies Pure Deontology if, for all actions $\alpha, \alpha' \in \Delta(A)$, if $\phi_i(\alpha,\mu_i) = \phi_i(\alpha',\mu_i)$ then $\alpha \sim^A \alpha'$.*

The second requirement is a specification of the counterfactual consequence function. Both players must consider the universalisation thought experiment where their opponent behaves as they do. Since the object of choice is a mixed action, the counterfactual outcome they consider is the distribution of material outcomes induced when both choose the same α, i.e. $\phi_i(\alpha,\mu_i) = \rho(\alpha,\alpha)$ for all i, μ_i and α.

Proposition 2. *Assume $\phi_i(\alpha,\mu_i) = \rho(\alpha,\alpha)$ is the counterfactual consequence function for all players i, beliefs μ_i, actions α and that the ranking $\succsim_i^A$ satisfies Pure Deontology for all i. Then, if α^* is optimal in any decision problem $\mathcal{D}_i$, it is also optimal for $\mathcal{D}_{-i}$ and it constitutes a Simple Kantian Equilibrium of the game $\mathcal{G}$.*

eliciting beliefs from choice is not the focus of this paper, I assume for simplicity that the decision maker has a belief over his opponent action and impose simple conditions that lead to equilibrium behaviour when needed. The restriction to two players allows me to avoid worring about belief consistency across many players.

[8] I define symmetry this way because the domain of ρ and ϕ_i is the ordered product of own and opponent action for each player i.

Proposition 2, together with other similar statements in this section, only identifies sufficient conditions, as one might play according to *SKE* for any reason.

This result allows me to compare the foundation I offer for *SKE* with that of Roemer (2019). He argues that, contrary to other models in economics, he does not assume exotic preferences, but classical self-regarding attitudes.[9] What he varies, instead, is agents' "optimisation protocol", as he refers to it. He contrasts Nash optimisation with Kantian optimisation. The former, 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. However, 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 a profile of actions constitutes a Nash Equilibrium. Nevertheless, this is only vaguely related to the interpretation of the concept.[10] Outside contexts of long repeated interactions and adaptive dynamics, an action in a Nash Equilibrium profile is played by a rational agent holding correct conjectures about opponents' behaviour.[11] 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 agent 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, lotteries 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 decision maker or of an equilibrium concept. There is no empirical observation able to tell that two agents 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 Proposition 2 that there is no need to rely on informal arguments regarding how agents optimise. Behaviour consistent with *SKE* reveals a preference for more desirable counterfactual outcomes. Therefore, Roemer is correct in arguing that assuming agents 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*, agents have selfish preferences over material outcomes but the optimisation protocol is different from Nash generates confusion. As I showed in the motivating example, it is possible that an agent who plays according to *SKE* does not have a rational preference, and hence a utility representation, over material outcomes. I believe the closest reformulation of Roemer's point is that one can have preferences for universalisation even if the utility index for

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

[10] Battigalli et al. (2022) offers a thorough discussion on the interpretation of Nash Equilibrium.

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

material outcomes U^{ρ} is the same as that for counterfactuals U^{ϕ}. However, this equivalence does not imply the agent 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 preferences over actions.

Proposition 2 also offers a novel interpretation of mixed actions. From a decision-theoretic perspective, there is always a pure action in the set of best replies to probabilistic conjectures regarding opponents' behaviour. The equilibrium mixed action of player i can be interpreted as strategic uncertainty from player $-i$'s perspective. Nevertheless, one who plays a mixed *SKE* action has no interest in being difficult for one's opponents to predict. In his best reply set, there may be no pure actions. My result reveals that a rationale for employing mixed actions is the adherence to a deontological attitude.

One last point is that *SKE* does not require strategic stability or correctness of beliefs. There are no conditions on agents' conjectures regarding their opponents for an action to constitute a *SKE*. Hence, *SKE* is an entirely decision-theoretic concept which puts no requirements on interactive features of the game, captured by the belief μ_i. I cast a vote in favour of the removal of the label equilibrium.

4.2 Homo Moralis

In this section, I exploit the representation in Theorem 1 to obtain Homo Moralis preferences. In the context of two-player symmetric games, a Homo Moralis is defined as follows.

Definition 5. *Fix a profile of mixed actions $(\alpha_i, \alpha_{-i}) \in \Delta(A) \times \Delta(A)$. Player i is a* ***Homo Moralis*** *(HM) with morality parameter $\kappa \in [0,1]$ if he derives the payoff*

$$(1-\kappa)\, V_i\left(\rho\left(\alpha_i, \alpha_{-i}\right)\right) + \kappa\, V_i\left(\rho\left(\alpha_i, \alpha_i\right)\right) \quad .$$

A Homo Moralis maximises a convex combination between expected material payoff and the same payoff in a counterfactual circumstance where both decision makers play his action. Moreover, he employs the same function V for both lotteries of consequences. Contrary to *SKE*, *HM* is a specification of preferences, it does not require joint optimality.

A *HM* with $\kappa = 1$, Homo Kantiensis, maximises the same objective function as the one in the definition of *SKE*. Proposition 2 implies that satisfying Pure Deontology and having the same counterfactual consequence function of *SKE* results in behaving as a Homo Kantiensis. In the general formulation with an intermediate κ, Pure Deontology must be relaxed, but Extended Consequentialism still holds. The following result identifies the conditions under which optimality in decision problems leads to *HM* payoff maximisation.

Proposition 3. *Assume $\phi_i(\alpha_i, \mu_i) = \rho(\alpha_i, \alpha_i)$ is the counterfactual consequence function for player i, for all beliefs μ_i, actions α_i and that the ranking $\succsim_i^A$ satisfies Separability. Then, player i maximises Homo Moralis payoff.*

A *HM* is not only interested in the universalisation counterfactual, but trades off consequentialist and deontological motives. Since *HM* is partially strategic, he also cares about his opponent's action and thus his beliefs matter. For this reason, it would be misleading to derive universalisation by imposing the player believes his opponent behaves like him, i.e. putting assumptions on μ_i. 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. The discussion of the previous section on the nature of the preference and the interpretation of mixed actions extends to *HM*. In the generalisation of *HM* to multiplayer games (Alger & Weibull, 2016), the agent behaves as if his opponents will implement the same action with some probability. This preference can easily be handled by properly specifying the counterfactual consequence function, which would then put probability weights on profiles of actions that are mixtures of μ_i and α_i.

The reader might have noticed that both *SKE* and *HM* are only well-defined in games with common action sets. Alger & Weibull (2013) suggested a way to employ *HM* preferences in asymmetric games. They proposed 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 and Rawls. 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 the latter concept, postponing observations on *Ex-ante HM* to Section 5.

4.3 Multiplicative Kantian Equilibrium

I specialise again Theorem 1 to obtain Multiplicative Kantian Equilibrium from optimality in individual decision problems.[12] The solution concept is defined in games where the action space features a linear structure. It is employed when players can choose from the real line, but the extension of my results in such a framework 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 to multiple actions games is not trivial.

I remove the restriction to symmetric games, but I assume players have only 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 6. *A mixed action profile $(\alpha_i, \alpha_{-i}) \in \Delta(A_i) \times \Delta(A_{-i})$ constitutes a **Multiplicative Kantian Equilibrium** (MKE) of the game $\mathcal{G}$ if, for all players i and real number $r \geq 0$*

$$U_i^{\phi}\left(\rho\left(\alpha_i, \alpha_{-i}\right)\right) \geq U_i^{\phi}\left(\rho\left(r \cdot \alpha_i, r \cdot \alpha_{-i}\right)\right)$$

A profile of mixed actions constitutes a Multiplicative Kantian Equilibrium if everyone prefers it compared with any counterfactual scenario where both players deviate by the same multiplicative factor r.[13] The notion is well defined as I am restricting attention to two actions. A multiplicative deviation is equivalent to move weight from one action to the other.

[12] An analysis equivalent to the one in this section delivers similar results for Additive Kantian Equilibrium (Roemer, 2019).

[13] Notice that, according to this definition, $(0, 0)$ is always a *MKE*.

The difference between *MKE* and *SKE* is the counterfactual consequence function, illustrated in Figure 1. An agent 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 outcomes lie on the line connecting the origin and the reference profile, i.e. all the couples 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.

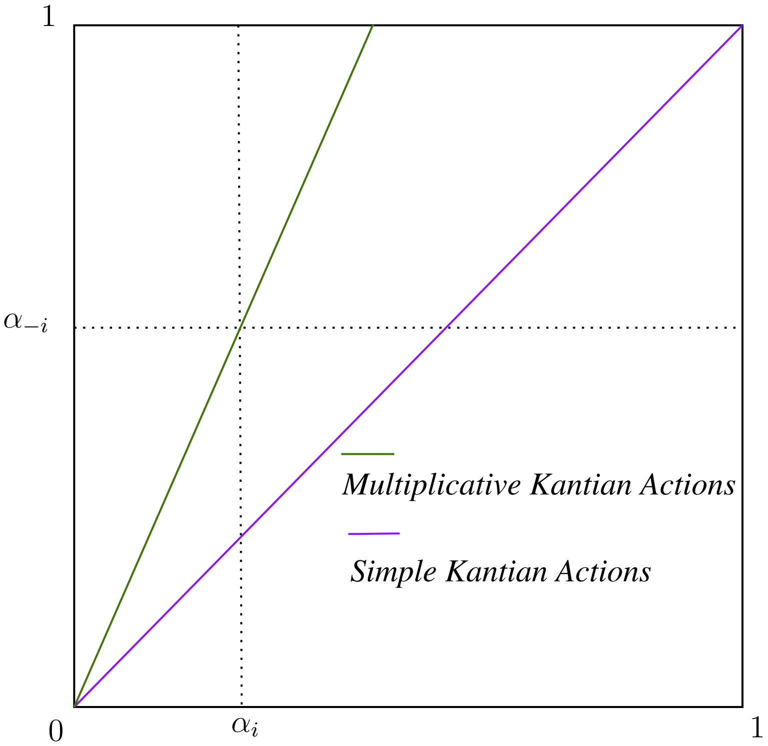

Figure 1: Counterfactual action profiles of Simple and Multiplicative Kantian Equilibria.

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 show that *MKE* can be reached if agents construct their counterfactual consequence functions from beliefs.

Fix an action α_i, a belief μ_i and consider the counterfactual $\phi_i\left(r \cdot \alpha_i, \mu_i\right) = \rho\left(r \cdot \alpha_i, r \cdot \mu_i\right)$ for all r. Player i takes (α_i, μ_i) as a special 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 β_1, 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_2, and (β_2, α_2) is his favourite profile compared to the counterfactuals on the line passing thorough it from the origin. Then, α_1 and β_2 are optimal actions, but do not necessarily constitute a *MKE*.

Proposition 4 links this counterfactual consequence function to *MKE*.

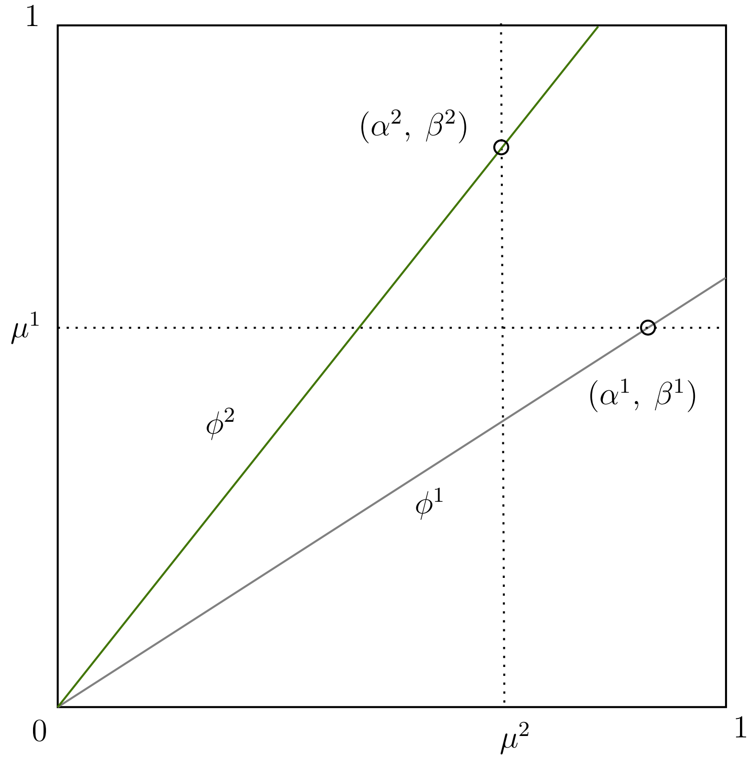

Figure 2: Multiplicative counterfactuals under different reference profiles.

Proposition 4. *Assume beliefs are $\mu_i = \alpha_{-i}$ for all players i and this is common knowledge. Assume counterfactuals are $\phi_i\left(r \cdot \alpha_i, \mu_i\right) = \rho\left(r \cdot \alpha_i, r \cdot \mu_i\right)$ for all players i, beliefs μ_i and multiplicative deviations r. Assume that $\succsim_i^A$ satisfies Pure Deontology for all i. Then, if α_i is optimal in decision problem $\mathcal{D}_i$ for all i, the profile (α_i, α_{-i}) constitutes a Multiplicative Kantian Equilibrium of the game $\mathcal{G}$.*

One of the main differences between *MKE* and *SKE*, highlighted by Proposition 4, is that the former, contrary to the latter, requires correctness of beliefs. It would be possible to assume that players have the same counterfactual consequence function, that this is common knowledge, and a similar result obtains without requirements on conjectures. Nevertheless, my approach is more general and it is reminiscent of the decision-theoretic requirements for Nash Equilibrium.

Contrary to *SKE*, *MKE* can be defined outside symmetric settings and allows agents 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. 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 *SKE* for purely deontological agents.

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 label of actions has a clear conceptual significance, as it is effort, contribution to a public good or use of a common resource.

I propose to use the relevant material outcome of the game as a currency. In game theory, this is usually players' *VnM* utility, but it can be any other index of wellbeing. Then "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 agents playing the standard prisoners' dilemma. Numbers are Bernoulli utilities for material outcomes.

1 \ 2	b	b'
a	$2,2$	$0,3$
a'	$3,0$	$1,1$

Row player 1 obtains his highest material reward in the outcome $(3,0)$, induced by the profile (a',b). Any mixed action α^k can be compared to a' in terms of the difference in expected material payoff $3-\left[2\alpha^k+3\left(1-\alpha^k\right)\right]=k$. The profile leading to the greatest material reward for column player 2 is instead (a,b'). As for player 2, his mixed actions β^k induces the difference $3-\left[2\beta^k+3\left(1-\beta^k\right)\right]=k$. Assume player 1, when picking any mix α^k, considers the counterfactual where 2 chooses β^k. He envisions the outcome that would obtain if his opponent chooses the action that, if considered a unilateral deviation from (a,b'), generates the same difference in material payoff. In this game, $\alpha^k=\beta^k=k$ for all k, therefore the counterfactual is identical to the one of *SKE* and *HM*. The actions that lead to the highest reward under this counterfactual evaluation are a for 1 and b for 2, i.e. $\alpha^k=\beta^k=1$, the same prediction of *SKE* and *HM* under proper relabeling 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*).[14] An agent who exhibits *ESU* first identifies the profile of actions implementing his preferred material outcome. Second, he evaluates each action considering the induced difference, or sacrifice, 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 sacrifice.

To ease the exposition, I here focus on equal absolute sacrifice (Young, 1988). In Appendix A, I consider general equal sacrifice problems. The results and arguments in this section hold for any equal sacrifice rule. Denote with $\left(\alpha_i^*,\alpha_{-i}^*\right)$ a profile that induces the maximal material outcome for player i. Actions leading to absolute sacrifice k satisfy:

$$U_i^\rho\left(\rho\left(\alpha_i^*,\alpha_{-i}^*\right)\right)-U_i^\rho\left(\rho\left(\alpha_i^k,\alpha_{-i}^*\right)\right)=k \tag{3}$$

[14] A fancier name is *Millian universalisation*.

An *ESU* counterfactual consequence function satisfies $\phi_i\left(\alpha^k,\mu_i\right)=\rho\left(\alpha_i^k,\alpha_{-i}^k\right)$ for any sacrifice k and belief μ_i. When choosing α^k, the agent 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^k are guaranteed to be unique, so it is possible that there are multiple equal sacrifice counterfactual consequence functions. The definition of α^k and β^k relies on the existence of a cardinal ranking over material outcomes U_i^ρ for all players i. Therefore, I already assumed vNM and Extended Consequentialism. By Theorem 1, preferences can be represented by Equation 1.

Definition 7. *Player i exhibits* ***Equal Sacrifice Universalisation*** *(ESU) if he derives the payoff $U_i^\phi\left(\rho\left(\alpha_i^k,\alpha_{-i}^k\right)\right)$ from choosing action α_i^k, as defined in Equation 3.*

Of course, it is possible to consider convex combinations between material and counterfactual payoffs, as in *HM*.

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 outcomes.

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

Asymmetry

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

Common Actions

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

Symmetry

Consider the game in the table on the left and assume throughout that $U_i^\rho=U_i^\phi$ for all i, i.e. Separability holds. The greatest achievable material payoff of both players is 2, in (a,b) and (a',b'). The action α^k of player 1 inducing a sacrifice of k solves $2-2\alpha^k=k$ and hence $\alpha^k=\frac{2-k}{2}$. The equivalent for player 2 is $2-(2-2\beta^k)=k$ which implies $\beta^k=\frac{k}{2}=1-\alpha^k$. *ESU* prescribes to choose k to maximise $U_i^\phi\left(\alpha^k,\beta^k\right)=U_i^\phi\left(\alpha^k,1-\alpha^k\right)$. The optimum is reached at $k=\frac{2}{3}$ with $\alpha^k=\beta^k=\frac{1}{2}$, 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 (b,b') 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 Homo Kantiensis. The issue is that it is meaningless to define "the same thing" as "the same action" in this scenario. The relabeling of actions from the first to the second table 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 the third table on the right, to make it symmetric. Now actions are interpred 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 relabeling of actions from the first to the third table amounts to changing any mixed action β^k to $1-\alpha^k$, which leads to $b=a'$ and $b'=a$ and switch columns. This is exactly the *ESU* counterfactual.

Now consider the difference between *ESU* and *Ex-Ante HM*. The latter, for morality parameter $\kappa=1$, prescribes player i to maximise $U_1^{\rho}(\rho(\alpha,\beta))+U_2^{\rho}(\rho(\beta,\alpha))$, leading to (a,b) or (a',b'). 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 equivalent 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 *ESU* is equivalent to *HM* and *SKE* in symmetric games. It holds for any equal sacrifice rule, as shown in the proof in Appendix B.

Proposition 5. *Assume the game $\mathcal{G}$ is symmetric. Then, there exists an ESU counterfactual consequence function equivalent to the HM and SKE one.*

A corollary of Proposition 5 is that if $\succsim_i^A$ satisfies Separability, *ESU* is equivalent to *HM*, while if $\succsim_i^A$ satisfies Pure Deontology, then *ESU* has the same objective function in the definition of *SKE*. The result may be interpreted as a 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 agent. 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*. It is clear that *Ex-ante HM* also relies on the same informational requirement, as it employs the veil of ignorance construct. If one player *VnM* utility is blown up by a positive affine transformation, *Ex-ante HM* would deliver a different prediction. Since it prescribes to maximise the sum of utilities, it will favour the satisfaction of the individual with the highest scale. 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. Since the label of actions is meaningless in most economic environments, I proposed to consider an index of wellbeing, *VnM* utility in this case, as the relevant universalisation currency, in contrast with previous conceptions.

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.

6 Conclusion

I have built a decision-theoretic model to account for a deontological preference for universalisation. I derived a representation theorem for preferences that evaluate counterfactual scenarios from a collection of standard axioms and highlighted the conceptual difference between deontological and pro-social attitudes. Then, I specified my model to account for the universalisation thought experiment. I showed that the general framework allows for unifying the two most prominent models of universalisation, namely Kantian Equilibrium and Homo Moralis. Lastly, I proposed a novel concept of universalisation, inspired by the equal sacrifice principle, that does not rely on the label of actions, is equivalent to the previous models under symmetry and can be defined in asymmetric games. My work sheds light on the conceptual underpinnings of universalisation, and thus guides empirical work and 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 outcomes to account for apparent paradoxes. Baccelli & Mongin (2021), among others, 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 adhered to this principle. I do not need to affect 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 agent cares about something else unrelated to the material features of the game. Expansion of the consequence domain is necessary. A second possibility is to include the chosen action in the description of the outcome. 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 theory does not rely on assuming that an action is optimal, but explains why, i.e. because it induces the preferred counterfactual consequence.

The last point regards the nature of preferences for universalisation. I denoted these as deontological and the literature refers to them as moral. Nevertheless, I show that universalisation satisfies consequentialism under an appropriate redefinition of consequences. What is then 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 wellbeing that results from it*". Universalisation is thus, strictly speaking, not a welfarist attitude, as the optimal action is unrelated to the distribution of wellbeing it induces. It may be welfarist in the evaluation of the counterfactual outcome, but, as I showed, this is not necessary.

[15] Binmore (1994, Ch. 4) and Sen (2017, Ch. 7) offer critical overviews of reasonable approaches to perform this exercise.

REFERENCES

Al-Najjar, N. I., & Pomatto, L. (2020). Aggregate risk and the Pareto principle. *Journal of Economic Theory*, *189*, 105084. 8

Alger, I., & Laslier, J.-F. (2021a). Homo moralis goes to the voting booth: A new theory of voter turnout. *Working Paper*. 4

Alger, I., & Laslier, J. F. (2021b). Homo moralis goes to the voting booth: Coordination and information aggregation. *Journal of Theoretical Politics*. 4

Alger, I., & Van Leeuwen, B. (2023). Estimating social preferences and kantian morality in strategic interactions. *Working Paper*. 8

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302. 3, 4, 8, 11, 14

Alger, I., & Weibull, J. W. (2016). Evolution and kantian morality. *Games and Economic Behavior*, *98*, 56–67. 4, 14

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: Genes, guns, and culture. *Journal of Economic Theory*, *185*, 104951. 4

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199–205. 2, 3, 7

Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica*, 404–437. 1

Baccelli, J., & Mongin, P. (2021). Can redescriptions of outcomes salvage the axioms of decision theory? *Philosophical Studies*, 1–28. 20

Battigalli, P. (1996). Comment on Mariotti (1996). *The Rational Foundations of Economic Behaviour*, 149–154. 10

Battigalli, P., Catonini, E., & De Vito, N. (2022). *Game theory: Analysis of strategic thinking.* 12

Battigalli, P., Cerreia-Vioglio, S., Maccheroni, F., & Marinacci, M. (2017). Mixed extensions of decision problems under uncertainty. *Economic Theory*, *63*(4), 827–866. 3

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*(1), 1–35. 6

Binmore, K. (1994). *Game Theory and the Social Contract: Playing fair*. MIT Press. 20

Bradley, R., & Fleurbaey, M. (2021). John Broome. *Conversations on Social Choice and Welfare Theory-Vol. 1*, 115–127. 20

Bradley, R., & Stefánsson, H. O. (2017). Counterfactual desirability. *The British Journal for the Philosophy of Science*, *68*, 485–533. 4

Brekke, K. A., Kverndokk, S., & Nyborg, K. (2003). An economic model of moral motivation. *Journal of public economics*, *87*(9-10), 1967–1983. 4

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. 20

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869. 6

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

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

De Donder, P., Llavador, H., Penczynski, S., Roemer, J. E., & Vélez, R. (2021). A game-theoretic analysis of childhood vaccination behavior: Nash versus Kant. *Working Paper*. 4

Dekel, E., & Siniscalchi, M. (2015). Epistemic game theory. In *Handbook of Game Theory with Economic Applications* (Vol. 4, pp. 619–702). Elsevier. 12

Dewatripont, M., & Tirole, J. (2022). The morality of markets. *Working Paper*. 4

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

Eichner, T., & Pethig, R. (2020). Kant–Nash tax competition. *International Tax and Public Finance*, *27*, 1108–1147. 4

Fishburn, P. C. (1970). *Utility theory for decision making*. New York: Wiley. 3, 8, 9

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

Fleurbaey, M., Kanbur, R., & Snower, D. J. (2021). An Analysis of Moral Motives in Economic and Social Decisions. *Working Paper*. 4

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and economic Behavior*, *1*(1), 60–79. 6

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. 3, 5

Grillo, A. (2021). Ethical Voting in Heterogenous Groups. *Working Paper*. 4

Hammond, P. J. (1988). Consequentialist foundations for expected utility. *Theory and decision*, *25*(1), 25–78. 3

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, *63*(4), 309–321. 14

Harsanyi, J. C. (1986). *Rational behaviour and bargaining equilibrium in games and social situations*. Cambridge University Press. 1

Herstein, I. N., & Milnor, J. (1953). An axiomatic approach to measurable utility. *Econometrica, Journal of the Econometric Society*, 291–297. 6

Jeffrey, R. (1990). *The Logic of Decision*. Chicago: The University of Chicago Press. 4

Kordonis, I. (2020). A Model for Partial Kantian Cooperation. In *Advances in Dynamic Games* (pp. 319–348). Springer. 4

Kreps, D. M. (1988). *Notes on the theory of choice*. Westview Press. 9

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

Laslier, J.-F. (2022). Universalization and altruism. *Social Choice and Welfare*, 1–16. 4, 19

Long, N. V. (2020). A dynamic game with interaction between Kantian players and Nashian players. In *Games in management science* (pp. 249–267). Springer. 4

Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley. 2, 3

Maniquet, F. (2019). Comments on John Roemer's first welfare theorem of market socialism. *Review of Social Economy*, *77*(1), 56–68. 19

Mariotti, M. (1995). Is Bayesian rationality compatible with strategic rationality? *The economic journal*, *105*(432), 1099–1109. 10

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. 8

Mill, J. S. (1885). *Principles of political economy*. D. Appleton. 17

Perea, A. (2012). *Epistemic game theory: Reasoning and choice*. Cambridge University Press. 12

Rawls, J. (1971). *A theory of justice*. Harvard university press. 14

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

Roemer, J. E. (2015). Kantian optimization: A microfoundation for cooperation. *Journal of Public Economics*, *127*, 45–57. 4

Roemer, J. E. (2019). *How we cooperate*. Yale University Press. 2, 3, 4, 6, 8, 11, 12, 14, 18

Rohde, K. I. (2010). A preference foundation for Fehr and Schmidt's model of inequity aversion. *Social Choice and Welfare*, *34*(4), 537–547. 10

Sarkisian, R. (2017). Team Incentives under Moral and Altruistic Preferences: Which Team to Choose? *Games*, *8*(3), 37. 4

Sarkisian, R. (2021a). Optimal Incentives Schemes under Homo Moralis Preferences. *Games*, *12*(1), 28. 4

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

Savage, L. J. (1972). *The foundations of statistics*. Dover Publications. 2, 7

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. 20

Sher, I. (2020). Normative aspects of kantian equilibrium. *Erasmus Journal for Philosophy and Economics*, *13*(2), 43–84. 19

Sobrado, E. M. (2022). Taxing moral agents. *CESifo working paper series 9867*. 4

Stovall, J. E. (2020). Equal sacrifice taxation. *Games and Economic Behavior*, *121*, 55–75. 20

Thomson, W. (2013). Game-theoretic analysis of bankruptcy and taxation problems: Recent advances. *International Game Theory Review*, *15*(03), 1340018. 23

Thomson, W. (2019). *How to divide when there isn't enough*. Cambridge University Press. 23

von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton University Press. 3, 7

Young, H. P. (1988). Distributive justice in taxation. *Journal of Economic Theory*, *44*(2), 321–335. 3, 17

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. As I always refer to utility for material outcomes in this section, the index ρ is omitted to avoid clutter. 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 agents and $x_i \in \mathbb{R}_{++}$ is the claim of agent 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 outcomes he can obtain, denoted $\overline{U_i}$. Therefore, $x_i = \overline{U_i}$ for all i. An **allocation rule** maps claims to awards $\pi : \mathbb{R}^2_{++} \to \mathbb{R}^2_+$. An **equal sacrifice function** is a continuous, strictly increasing and hence invertible function $R : \mathbb{R}_{++} \to \mathbb{R}$. The equal sacrifice allocation rule relative to the function R and sacrifice $k \in \mathbb{R}_+$ is

[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.

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

As an example, the equal loss rule $\pi_R(x_i, x_{-i}) = (x_i - k)_{i\in I}$ in the main text has $R(x_i) = x_i$ for all x_i. In a game, utilies depend on actions, so the above can be rewritten as

$$R^{-1}\left(R\left(U_i\left(\alpha_i, \alpha_{-i}\right)\right) - k\right)$$

Denote the profile of actions inducing the maximal expected utility for i with $\left(\alpha_i^*, \alpha_{-i}^*\right)$, i.e. $U_i\left(\alpha_i^*, \alpha_{-i}^*\right) = \overline{U_i}$. Then, action α_i^{Rk} induces sacrifice k relative to the function R if

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

A player exhibits equal sacrifice universalisation with respect to R if his counterfactual is $\phi_i\left(\alpha_i^{Rk}, \mu_i\right) = \rho\left(\alpha_i^{Rk}, \alpha_{-i}^{Rk}\right)$ for any belief μ_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_{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 Extended Consequentialism, imposed on a weak order over actions, implies the existence of a weak order $\succsim^C$ on couples of distribution of consequences. Consider the couples of mixed actions (α, β) and (α', β'), not necessarily distinct among themselves, that induce the same distributions of consequences $\gamma, \delta \in \Delta(C)$:

$$\begin{aligned}(\rho(\alpha,\mu), \phi(\alpha,\mu)) = (\rho(\beta,\mu), \phi(\beta,\mu)) = (\gamma, \delta) \quad ; \\ (\rho(\alpha',\mu), \phi(\alpha',\mu)) = (\rho(\beta',\mu), \phi(\beta',\mu)) = (\gamma', \delta') \quad .\end{aligned}$$

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 couples of distributions of consequences, define $(\gamma, \delta) \succsim^C (\gamma', \delta')$ if actions α, α' exist such that $(\rho(\alpha,\mu), \phi(\alpha,\mu)) = (\gamma, \delta)$, $(\rho(\alpha',\mu), \phi(\alpha',\mu)) = (\gamma', \delta')$ and $\alpha \succsim^A \alpha'$. I can restrict the domain of $\succsim^C$ to the product $\Delta(C) \times \Delta(C)$, as by Extended Consequentialism two actions that induce the same distributions with different correlation structures must be indifferent to the agent. By completeness and transitivity of $\succsim^A$, it follows that $\succsim^C$ is complete and transitive.

Second, I show that the existence of the weak order $\succsim^C$ implies that Extended Consequentialism is satisfied. I proceed by contrapositive, i.e. I prove that if Extended Consequentialism is violated then $\succsim^C$ is not a weak order. For the sake of the argument, assume that Extended Consequentialism does not hold. Then, there exist actions α and α' that induce the same distributions over consequences such that $\alpha \succ^A \alpha'$. If both α and α' induce the lotteries (γ, δ), then $(\gamma, \delta) \succ^C (\gamma, \delta)$, which breaks reflexivity and hence completeness of $\succsim^C$. □

Before proceedig to the proof of Proposition 1, I show the independence between Extended Consequentialism and Separability. First, Extended Consequentialism does not imply Separability, as the former only disciplines preferences between actions that induce the same marginal distributions, 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 marginal distibutions (γ, γ'). Assume $\alpha \succ^A \alpha'$, which is possible because Extended Consequentialism is not required. Separability only implies $\gamma \succsim \gamma$ and $\gamma' \succsim \gamma'$.

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

Part 1. Consider a point $(c_0, c_0') \in C \times C$ and define $u^\rho, u^\phi : C \to \mathbb{R}$ so that

$$u^\rho(c_0) + u^\phi(c_0') = u(c_0, c_0') \tag{4}$$

$$u^\rho(c) = u(c, c_0') - u^\phi(c_0') \ \ \forall \ c \in C \tag{5}$$

$$u^\phi(c') = u(c_0, c') - u^\rho(c_0) \ \ \forall \ c' \in C \tag{6}$$

Adding 5 and 6

$$u^\rho(c) + u^\phi(c') = u(c, c_0') + u(c_0, c') - u^\rho(c_0) - u^\phi(c_0')$$

By equation 4

$$u^\rho(c) + u^\phi(c') = u(c, c_0') + u(c_0, c') - u(c_0, c_0') \tag{7}$$

Now, consider the comparison between the distributions

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

These have the same marginals

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

By Extended Consequentialism, the actions inducing these two distributions must be indifferent.[17] The definition of $\succsim^C$, implies $u(c, c') + u(c_0, c_0') = u(c, c_0') + u(c_0, c')$. Thus, by equation 7

$$u^\rho(c) + u^\phi(c') = u(c, c')$$

Define $u_*^\rho(c, c') = u^\rho(c)$ and $u_*^\phi(c, c') = u^\phi(c')$ for all (c, c'). Then, for all probability distribution $p \in \Delta(C) \times \Delta(C)$, by denoting $\mathbb{E}_p[\cdot]$ the expectation with respect to p, I obtain

[17] Indifference between equivalent $50-50$ gambles is enough to perform this step. However, such a relaxation of Extended Consequentialism does not allow to prove Proposition 1 and hence define $\succsim^C$.

$$\begin{aligned}\mathbb{E}_p\left[u\left(c,c'\right)\right] =& \mathbb{E}_p\left[u_*^{\rho}\left(c,c'\right)+u_*^{\phi}\left(c,c'\right)\right]\\ &\mathbb{E}_p\left[u_*^{\rho}\left(c,c'\right)\right]+\mathbb{E}_p\left[u_*^{\phi}\left(c,c'\right)\right]\\ &\mathbb{E}_p\left[u^{\rho}\left(c\right)\right]+\mathbb{E}_p\left[u^{\phi}\left(c'\right)\right]\end{aligned}$$

Hence, for any action $a \in A$

$$\begin{aligned}\sum_{s\in S}\mu\left(s\right)u\left(\rho\left(a,s\right),\phi\left(a,s\right)\right) &= \sum_{s\in S}\mu\left(s\right)u^{\rho}\left(\rho\left(a,s\right)\right)+\sum_{s\in S}\mu\left(s\right)u^{\phi}\left(\phi\left(a,s\right)\right) \quad (8)\\ U\left(\rho\left(a,\mu\right),\phi\left(a,\mu\right)\right) &= U^{\rho}\left(\rho\left(a,\mu\right)\right)+U^{\phi}\left(\phi\left(a,\mu\right)\right)\end{aligned}$$

Since U is an expected utility, there exists equivalent representations $U' = qU + r$ with $q > 0$. Then

$$\begin{aligned}u'\left(\rho\left(a,\mu\right),\phi\left(a,\mu\right)\right) &= qu\left(\rho\left(a,\mu\right),\phi\left(a,\mu\right)\right)+r\\ &= qu^{\rho}\left(\rho\left(a,\mu\right)\right)+qu^{\phi}\left(\phi\left(a,\mu\right)\right)+r\end{aligned}$$

From which $u'^{\rho}\left(c\right) = qu^{\rho}\left(c\right)+r^{\rho}$ where $r^{\rho} = r + qu^{\phi}\left(c_0'\right) - u_{\phi}'\left(c_0'\right)$.

Part 2. If Separability holds, then $\gamma \precsim \gamma'$ if and only if both $\mathbb{E}_{\gamma}\left[u^{\rho}\right] \geq \mathbb{E}_{\gamma'}\left[u^{\rho}\right]$ and $\mathbb{E}_{\gamma}\left[u^{\phi}\right] \geq \mathbb{E}_{\gamma'}\left[u^{\phi}\right]$, for any $\gamma, \gamma' \in \Delta\left(C\right)$. Therefore, U^{ρ} and U^{ϕ} are two functions on an identical domain representing the same preferences and hence must be related by positive affine transformations so that $U^{\phi} = q^{\phi}U^{\rho} + \delta^{\phi}$ with $q^{\phi} > 0$. Fix $V = U^{\rho}$, then the representation follows with $\lambda^{\rho} = 1$ and $\lambda^{\phi} = q^{\phi}$. By the same reasoning of **Part 1.** of the proof, V is unique up to positive affine transformations. □

For Proposition 2, I first prove a preliminary lemma. Denote $\alpha_{(\gamma,\gamma')}$ the actions that induce the couple of lotteries over material and counterfactual outcomes (γ, γ').

Lemma 1. *Assume $\succsim^A$ is a weak order. Then, $\succsim^A$ satisfies Pure Deontology if and only if there exists a weak order over distributions of counterfactual consequences $\succsim^{\phi}$ such that, for all distributions $\delta, \delta' \in \Delta\left(C\right)$ and actions $\alpha_{(\delta,\cdot)}, \alpha_{(\delta',\cdot)} \in \Delta\left(A\right)$,*

$$\gamma \succsim^{\phi} \gamma' \iff \alpha_{(\delta,\gamma)} \succsim^{A} \alpha_{(\delta',\gamma')} \;\; .$$

Proof of Lemma 1. First, I prove that the hypotheses imply the existence of $\succsim^{\phi}$. It is enough to construct a ranking that is constant in the material outcome, as Pure Deontology puts no requirements in the evaluation of material consequences. The logic of this proof is similar to the one of Proposition 1.

Consider the couple of not necessary distinct actions (α, β) and (α', β') inducing the same distribution over counterfactual outcomes $\phi\left(\alpha,\mu\right) = \phi\left(\beta,\mu\right) = \gamma$ and $\phi\left(\alpha',\mu\right) = \phi\left(\beta',\mu\right) = \gamma'$. I assumed the action set is rich enough for these actions to exist. By Pure Deontology $\alpha \sim^{A} \beta$ and $\alpha' \sim^{A} \beta'$. Transitivity of $\succsim^{A}$ implies $\alpha \succsim^{A} \alpha' \iff \beta \succsim^{A} \beta'$. For all distributions of consequences, define $\gamma \succsim^{\phi} \gamma'$ if actions α, α' exist such that $\phi(\alpha, \mu) = \gamma$, $\phi(\alpha', \mu) = \gamma'$ and $\alpha \succsim^{A} \alpha'$. The ranking $\succsim^{\phi}$ inherits completeness and transitivity from $\succsim^{A}$.

As for the other direction of the statement, a procedure by contrapositive similar to the one of Proposition 1 works. □

Proof of Proposition 2. By Lemma 1, as all agents satisfy Pure Deontology, their optimal action induces the maximal element according to $\succsim_i^{\phi}$. Since $\succsim_i^{A}$ satisfies vNM, being maximal in $\succsim_i^{\phi}$ is equivalent to maximise expected utility. Expected utility of player i is then $\sum_{c \in C} \phi_i(\alpha, \mu_i)[c]\, u_i^{\phi}(c)$, where u_i^{ϕ} is the Bernoulli utility representation of $\succsim_i^{\phi}$ from Lemma 1. Hence, i objective function is

$$\begin{aligned}\sum_{c \in C} \phi_i(\alpha, \mu_i)[c]\, u_i^{\phi}(c) &= \sum_{c \in C} \rho(\alpha, \alpha)[c]\, u_i^{\phi}(c)\\ &= \sum_{a \in A}\sum_{b \in A} \alpha(a)\, \alpha(b)\, u_i^{\phi}(\rho(a, b))\\ &= U_i^{\phi}(\rho(\alpha, \alpha)).\end{aligned}$$

If α^* maximises this function, then it is also optimal for player $-i$ as $U_i^{\phi}(\rho(\alpha, \alpha)) = U_{-i}^{\phi}(\rho(\alpha, \alpha))$ for all α by symmetry and because players share the same counterfactual consequence function. □

Proof of Proposition 3. Since player i satisfies vNM, Extended Consequentialism and Separability, the requirements for the second part of Theorem 1 are met. Given mixed action α_i, he derives the following payoff

$$\lambda^{\rho} \sum_{c \in C} \rho(\alpha_i, \mu_i)[c]\, v_i(c) + \lambda^{\phi} \sum_{c \in C} \phi_i(\alpha_i, \mu_i)[c]\, v_i(c)$$

By assumption, $\phi_i(\alpha_i, \mu_i) = \rho(\alpha_i, \alpha_i)$. Hence, the payoff function becomes

$$\lambda^{\rho} \sum_{a \in A}\sum_{b \in A} \alpha_i(a)\, \mu_i(b)\, v_i(\rho(a, b)) + \lambda^{\phi} \sum_{a \in A}\sum_{b \in A} \alpha_i(a)\, \alpha_i(b)\, v_i(\rho(a, b))$$

$$\lambda^{\rho} V_i(\rho(\alpha, \mu_i)) + \lambda^{\phi} V_i(\rho(\alpha, \alpha)).$$

The relative weights pin down the morality parameter κ, i.e. $\frac{\lambda^{\rho}}{\lambda^{\rho}+\lambda^{\phi}} = \kappa \in [0, 1]$. □

Proof of Proposition 4. Since they all players satisfy Pure Deontology, by an argument similar to Proposition 2, they maximise counterfactual expected utility. The counterfactual consequence function for both of them is $\phi_i(\alpha_i, \mu_i) = \rho(r \cdot \alpha_i, r \cdot \mu_i) = \rho(r \cdot \alpha_i, r \cdot \alpha_{-i})$ for all r, since $\mu_i = \alpha_{-i}$ for all i. If α_i is optimal in $\mathcal{D}_i$, then

$$U_i^{\phi}(\rho(\alpha_i, \alpha_{-i})) \geq U_i^{\phi}(\rho(r \cdot \alpha_i, r \cdot \alpha_{-i}))$$

for all i and therefore (α_i, α_{-i}) constitutes a Multiplicative Kantian Equilibrium. □

Proof of Proposition 5. I employ the notation of Appendix A. Pick a profile implementing the maximal expected utility for material outcomes for player i denoted $\left(\alpha_i^*, \alpha_{-i}^*\right)$. An action α_i^{Rk} inducing sacrifice k for rule R satisfies the following:

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

Since the game is symmetric, the profile $\left(\alpha_i^*, \alpha_{-i}^*\right)$ also induces a maximal outcome for player $-i$ as $U_1^{\rho}\left(\rho\left(\alpha_i^*, \alpha_{-i}^*\right)\right) = U_2^{\rho}\left(\rho\left(\alpha_i^*, \alpha_{-i}^*\right)\right)$. Then, the condition for equal sacrifice of $-i$ is equivalent to the one of i

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

that implies $\alpha_i^{Rk} = \alpha_{-i}^{Rk}$ for every k. Then, the counterfactual consequence function of player i is $\phi_i\left(\alpha_i^k, \mu_i\right) = \rho\left(\alpha_i^k, \alpha_i^k\right)$. □