Universalization and altruism

Jean-François Laslier

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2021 – 30

Universalization and altruism

Jean-François Laslier

JEL Codes: C73; D63; D67.
Keywords: ethics, games, evolution, altruism, universalization, Kant, Homo Moralis.

Funded by a French government subsidy managed by the ANR under the framework of the Investissements d'avenir programme reference ANR-17-EURE-001

Universalization and altruism*

Jean-François Laslier†

May 15, 2021

Abstract

To any normal form game, we associate the symmetric two-stage game in which, in a first stage, the roles to be played in the base game are randomly assigned. We show that any equilibrium of the κ-universalization of this extended game is an equilibrium of the base game played by altruistic players ("ex ante Homo Moralis is altruistic"), and that the converse is false. The paper presents the implications of this remark for the philosophical nature of ethical behavior (Kantianism behind the veil of ignorance implies but is stronger than altruism) and for its evolutionary foundations.

Keywords: ethics, games, evolution, altruism, universalization, Kant, Homo Moralis
JEL Codes: C73; D63; D67.

1 Introduction

What Harsanyi (1955, 1980) calls "act utilitarianism" is the behavior rule that requires a decision-maker to choose the action that maximizes the total welfare of all involved players, or, in a less extreme form, to assign a positive weight $\alpha > 0$ to the other players' welfare (Becker 1974, Lindbeck and Weibull 1988). We will use the explicit phrase "utilitarian altruism" or simply "altruism".

The principle of universalization, as used by Laffont (1975), requires the decision maker to chose the action that would maximize her own welfare if this action was chosen by all players, or, in a less extreme form termed Homo Moralis by Alger and Weibull (2013), if a fraction $\kappa > 0$ of them would do so.

*I thank Ingela Alger, Jörgen Weibull and members of e-Michs seminar at the Chaire Hoover in Louvain. I acknowledge the EUR grant ANR-17-EURE-0001.

†CNRS, Paris School of Economics, 48 Bd. Jourdan, 75014 Paris, France.

Both theories have been proposed as models of ethics. In their definitions, each captures a different ingredient of what is commonly recognized as ethical behavior: the altruistic concern and the universalization principle. Still, they seem very different at first sight: the definition of Homo Moralis only refers to the player's personal satisfaction and does not require inter-personal comparison of utilities whereas utilitarian altruism considers their sum.

They do not use the same framework: Utilitarian altruism requires that the utility level of the others can be measured with the same stick as the individual's own, but the actions available to different players do not need to be the same. On the other hand, Homo Moralis requires the strategy sets to be identical for all players (to some extent this is unavoidable in order to capture a notion of universalization) but the utility of different players are never added or even compared. In order to understand the relation between the two notions, one needs to spell them precisely in a compatible framework.

Previous work that compared altruism and Homo Moralis have either focused on the specific case of symmetric games, in which case the two ideas directly apply (Alger and Weibull 2017), or considered games extended to symmetric two-stage games in which nature first assigns their roles to the players: see the sections 6.1 in Alger and Weibul (2013) and the following papers Alger and Weibull (2020), Alger and Laslier (2020, 2021), Miettinen et al. (2020) and van Leeuwen and Alger (2021). The restriction to symmetric games is severe because the forms of strategic interactions that are interesting from the point of view of ethics go beyond symmetric games; so the present paper follows the second path. As will be seen, the idea of considering the symmetric extension of a possibly non-symmetric game is, at the cost of a slight complexity, fruitful. Philosophically it corresponds to a form of "veil of ignorance" that is familiar in theoretical ethics, and it has a natural interpretation in the evolutionary theory as the study of intra-species externalities.

From the philosophical point of view, morality should not be reduced to the implementation of actions; Kant's imperative (Kant 1785) is not about actions but about "maxims" for action, where a maxim is a device that leads to different choices in a variety of circumstances (Braham and van Hees 2020). A game-theoretical model that captures this notion will be used in this paper: like a Kantian maxim, a strategy will be the list of actions to be undertaken in the various roles a player can have in a multi-player game.

Likewise, Harsanyi (1980) distinguishes act utilitarianism (at work given the sit-

uation) from "rule" utilitarianism, in which morally right actions are defined in two steps, as the correct implementation of a right general rule. Our approach therefore captures this feature of Harsanyi's ethical theory: rule utilitarianism rather than act utilitarianism, although we will have to come back in the discussion on that point. Another similarity with Harsanyi is that the present work is based on the same notion of collective utility as sum of individual utilities, in the manner of Bentham, and that these individual utilities are measured in a way that is compatible with a probabilistic treatment of uncertainty, in the manner of Von Neumann. A point of departure is that, for Harsanyi, what is played behind the veil of ignorance, at the stage where the society's moral code is chosen, is a cooperative game, while we stick to the non-cooperative framework.

The Homo Moralis model fits well with that idea of "maxims" or "rules" because of its evolutionary interpretation: the same genotype is at work to determine behavior under many (if not all) circumstances. To give a simple two-role example, male-female interactions involve actors of two types within one species. The evolutionary argument is that even when playing an asymmetric conflict, individuals belonging to one evolutionary unit (in Biology: an animal species) are guided by their genotype, which is the same whatever role they happen to have in the conflict. As a man, my genotype contains what I would do if I were a woman.

The consequences of this remark were first noticed by Selten (1980): the equilibrium conditions for an asymmetric conflict played by members of the same species turn out to be more demanding than those required in the "same" game that would be played by members of different species. The point revealed to be of particular importance for the study of signal games and the evolutionary theory of language (Kim and Sobel 1992, Wärneryd 1993, Schlag 1994, Banerjee and Weibull 2000, Laslier 2003). The technical point made in the present paper is another instance of the same phenomenon.

From Rousseau (1755) to Binmore (1994) there is a long intellectual tradition that intends to lay naturalistic foundations to the social contract. This line of thought has been revivified by modern evolutionary approaches, both empirical and theoretical (see Maynard Smith 1982, Skyrms 1996, de Waal 1996, Nowak and Sigmund 2005, Sidanus and Kurzban 2013). Where does the universalization principle stand in this picture?

This principle is in general presented as a logical consistency requirement, that ends up calling "rational " commitment the satisfaction of rights and obligations to

act. We already mentioned Kant and Harsanyi on this point. The formal structure of universalization as a consistency requirement is presented by Gravel et al. (2000), with various applications in Economics, beyond the question of the individual choice of action. More recently, Roemer (2019) presented his "theory of Kantian optimization" as an attempt to explain cooperation that is based on self-interest (cooperation is different from solidarity) but that is not driven by the standard non-cooperative logic of isolated maximization, and even requires to step out of the standard model.

Such theories may have good descriptive power (*How* do we cooperate?) and psychological appeal, but It is not clear to what extent these views are in contradiction with a naturalistic foundation, or may be sustained by evolutionary justifications (Laslier 2020). What is clear is that Biology tells a story in which we, humans, are similar even to our enemies, and similar in particular to those friends or enemies that we often meet. Drawing the consequences of this remark, Alger and Weibull (2013) showed that a form of partial universalization logically follows. This paper will show that this partial universalization in turn implies a form of partial altruism: ex ante Homo Moralis is altruistic.

The paper is organized in five sections. After this introduction, Section 2 defines the main notions: α-altruism, symmetric extension of a normal form game, and κ-universalization. The results are stated in Section 3: partially universalized equilibria are partially altruistic but some partially altruistic equilibria may not be partially universalized. Section 4 makes several clarification remarks: We first detail the particular case of a base game that is itself symmetric, we also discuss the idea that the extended game itself might be played by altruistic players, and this discussion allows to further explain the relation of the present work with Harsanyi's distinction between "act" and "rule" utilitarianism. Section 5 tackles some points about the empirical content of the previous developments.

2 Definitions

2.1 The base game

Since the two-role example conveys exactly the point we want to make, this article will, for simplicity, be entirely written about two-player games. Extension to any number of players causes no difficulty.

All through the paper, we consider u a two-player normal-form game

$$(u_1, u_2) : A_1 \times A_2 \to \mathbb{R}^2.$$

The strategies of this base game will be called "actions". Note that this framework is general and that a particular case maybe the case of "mixed strategies", where the sets A_1 and A_2 are simplices.

2.2 α-altruism

The modified utilitarian altruistic game of parameter α, denoted $u^{(\alpha)}$, has the same strategy sets, A_1 and A_2 but the modified payoffs:

$$\begin{aligned} u_1^{(\alpha)} &= u_1 + \alpha u_2, \\ u_2^{(\alpha)} &= u_2 + \alpha u_1. \end{aligned}$$

A Nash equilibrium of $u^{(\alpha)}$ will be called an α-altruistic equilibrium.

Notice that, since we restrict attention to the case of two players, we obtain for $\alpha = 1$ that both players have the same modified payoffs $u_1 + u_2$. Their common objective is the collective welfare as measured by the Bentham sum of utilities.[1]

2.3 Ex ante symmetrized game

The symmetric game associated to u is the two-stage game in which, in the first stage, Nature assigns at random each player to one of the two roles "1" and "2" to play the base game u in the second stage, with the players then knowing their roles. The normal form of the extended game is therefore the two-player game $\widetilde{u}$ in which the two players have the same strategy set, that is the product $A_1 \times A_2$, and the payoffs are:

$$\begin{aligned} \widetilde{u}_1(a_1, a_2, a_1', a_2') &= \tfrac{1}{2} u_1(a_1, a_2') + \tfrac{1}{2} u_2(a_1', a_2), \\ \widetilde{u}_2(a_1, a_2, a_1', a_2') &= \tfrac{1}{2} u_1(a_1', a_2) + \tfrac{1}{2} u_2(a_1, a_2'). \end{aligned}$$

Here, a strategy (a_1, a_2) is a pair of conditional actions of the form "a_1 in role 1 and a_2 in role 2." Note that the extended game is by definition symmetric: writing $s = (a_1, a_2)$ and $s = (a_1', a_2')$, it comes $\widetilde{u}_1(s, s') = \widetilde{u}_2(s', s)$. We will thus simply write $\widetilde{u}$ for $\widetilde{u}_1$.

[1] Notice also that the alternative specification $u_i = a u_i + (1 - a) u_j$ is equivalent once setting $a = \alpha/(1 - \alpha)$.

2.4 κ-universalization

Since the game $\widetilde{u}$ is symmetric, it is possible to apply the Homo Moralis trick, κ-universalization, to this game. Let κ be a morality parameter, the payoff to be considered is now

$$\widetilde{u}^{[\kappa]}(s,s') = (1-\kappa)\widetilde{u}(s,s') + \kappa\widetilde{u}(s,s),$$

which writes here:

$$\begin{array}{rcl}\widetilde{u}^{[\kappa]}(a_1,a_2,a_1',a_2') & = & (1-\kappa)\widetilde{u}(a_1,a_2,a_1',a_2') + \kappa\widetilde{u}(a_1,a_2,a_1,a_2) \\ & = & \frac{1-\kappa}{2}(u_1(a_1,a_2') + u_2(a_1',a_2)) + \frac{\kappa}{2}(u_1(a_1,a_2) + u_2(a_1,a_2)).\end{array}$$

A symmetric equilibrium of $\widetilde{u}^{[\kappa]}$ is a composed strategy (a_1^*,a_2^*) such that the maximization of $\widetilde{u}^{[\kappa]}(a_1,a_2,a_1^*,a_2^*)$ with respect to (a_1,a_2) is obtained at $(a_1,a_2) = (a_1^*,a_2^*)$. Such (a_1^*,a_2^*) would deserve to be called a "κ-universalized equilibrium" of the original game u. We will also refer to it as "ex ante HM equilibrium".

3 Equilibria

Both ideas (altruism and universalization) participate to most conceptions of morality, and we will try to make precise the relation between them.

3.1 Universalization implies utilitarian altruism

Our main result can be stated now.

Proposition 1. *Let u be a two player game and let (s^*,s^*) be a symmetric equilibrium of the κ-universalised extended game $\widetilde{u}^{[\kappa]}$. Then s^* is an equilibrium of $u^{(\alpha)}$ for $\alpha = \kappa$.*

Proof. Write $s^* = (a_1^*,a_2^*)$. The definition of HM equilibrium entails the maximization of the real function $\widetilde{u}^{[\kappa]}(a_1,a_2,a_1^*,a_2^*)$ with respect to the two independent variables a_1 and a_2. Because the two variables are independent, this implies maximization with respect to each of them, the other being fixed, meaning that the maximum of $\widetilde{u}^{[\kappa]}(a_1,a_2^*,a_1^*,a_2^*)$ with respect to the variable a_1 is obtained at $a_1 = a_1^*$ (and likewise for a_2^*). Write the ex ante payoff $\widetilde{u}$ as a function of the variable a_1:

$$\begin{array}{rcl}\widetilde{u}^{[\kappa]}(a_1,a_2^*,a_1^*,a_2^*) & = & \frac{1-\kappa}{2}(u_1(a_1,a_2^*) + u_2(a_1^*,a_2^*)) + \frac{\kappa}{2}(u_1(a_1,a_2^*) + u_2(a_1,a_2^*)) \\ & = & \frac{1}{2}u_1(a_1,a_2^*) + \frac{1-\kappa}{2}u_2(a_1^*,a_2^*) + \frac{\kappa}{2}u_2(a_1,a_2^*).\end{array}$$

One can notice that the term $u_2(a_1^*, a_2^*)$ does not involve the variable a_1, and that the factor 1/2 does not matter. Therefore the player is choosing a_1 to maximize

$$u_1(a_1, a_2^*) + \kappa u_2(a_1, a_2^*)$$

that is exactly the altruistic payoff $u_1^{(\alpha)}(a_1, a_2^*)$ for $\alpha = \kappa$. The same thing holds for a_2. We conclude that any symmetric equilibrium of $\widetilde{u}^{[\kappa]}$ is an equilibrium of $u^{(\kappa)}$. □

The converse might not be true because an equilibrium of $\widetilde{u}^{[\kappa]}$ must be robust to *joint* deviations in a_1 and a_2, as will be seen now.

3.2 Universalization demands more than utilitarian altruism

To see that point, one needs a game where Nash equilibrium can be destabilized by joint deviations. So take a prisoner dilemma as the base game:

$$u = \begin{pmatrix} (c, c) & (0, c+\varepsilon) \\ (c+\varepsilon, 0) & (1, 1) \end{pmatrix}$$

with $c > 1$ and $\varepsilon > 0$.[2] The only Nash equilibrium of u ("defection") yields payoff 1 to both players.

The altruistic payoffs are:

$$u^{(\alpha)} = \begin{pmatrix} ((1+\alpha)c,\ (1+\alpha)c) & (\alpha(c+\varepsilon),\ c+\varepsilon) \\ (c+\varepsilon,\ \alpha(c+\varepsilon)) & (1+\alpha,\ 1+\alpha) \end{pmatrix}$$

and one can see that:

- "Defection" is an $u^{(\alpha)}$ equilibrium iff $1 + \alpha \geq \alpha(c+\varepsilon)$, that is $\alpha \leq \frac{1}{c+\varepsilon-1}$.
- "Cooperation" is an $u^{(\alpha)}$ equilibrium iff $(1+\alpha)c \geq c+\varepsilon$, that is $\alpha \geq \frac{\varepsilon}{c}$.

Depending on the values of the parameters c and ε, it is possible or not that, for intermediate value of α the two equilibria co-exist, or not.

We now turn to the extended game $\widetilde{u}^{[\kappa]}$. Because of the symmetries in u we only need to write down the following values for the payoffs, where cc, cd, dc, dd have the

[2] Remark that the base game in this example has symmetric payoffs. This is only for the sake of simplicity that we take this example; we treat u as a standard two-player game that must not be confused with its associated two-stage extension $\widetilde{u}$ in which a player ex ante chooses the actions he or she play in the two roles.

obvious meaning for cooperation (c) and defection (d) and, for instance, $\widetilde{u_1}(cc|cd)$ denotes the payoff for a player who always cooperate while his opponent cooperates in the first role and defects in the second role.

$$\begin{aligned}
\widetilde{u}(cc|cc) &= c \\
\widetilde{u}(cd|cc) &= c+\varepsilon/2 \\
\widetilde{u}(dd|cc) &= c+\varepsilon \\
\widetilde{u}(cc|cd) &= c/2 \\
\widetilde{u}(cd|cd) &= (c+\varepsilon)/2 \\
\widetilde{u}(dc|cd) &= (c+1)/2 \\
\widetilde{u}(dd|cd) &= (c+\varepsilon+1)/2 \\
\widetilde{u}(cc|dd) &= 0 \\
\widetilde{u}(cd|dd) &= 1/2 \\
\widetilde{u}(dd|dd) &= 1
\end{aligned}$$

To check whether "Defection" (dd) is a κ-universalized equilibrium, we compute the following payoffs:

$$\begin{aligned}
\widetilde{u}^{[\kappa]}(cc|dd) &= (1-\kappa)\cdot 0+\kappa\cdot c \\
\widetilde{u}^{[\kappa]}(cd|dd) &= (1-\kappa)\cdot 1/2+\kappa\cdot(c+\varepsilon)/2 \\
\widetilde{u}^{[\kappa]}(dd|dd) &= 1
\end{aligned}$$

and find the two conditions:

$$\begin{aligned}
1 \geq \kappa c &\iff \kappa \leq \tfrac{1}{c} \\
1 \geq (1-\kappa+\kappa c+\kappa\varepsilon)/2 &\iff \kappa \leq \tfrac{1}{c+\varepsilon-1}.
\end{aligned}$$

The second of these two conditions was the condition for u^{α}, as seen in the theoretical part, and the other adds a constraint that is the binding one if $\varepsilon < 1$.

To check whether "Cooperation" (cc) is an κ-moral equilibrium, we compute the following payoffs:

$$\begin{aligned}
\widetilde{u}^{[\kappa]}(cc|cc) &= c \\
\widetilde{u}^{[\kappa]}(cd|cc) &= (1-\kappa)\cdot(c+\varepsilon/2)+\kappa\cdot(c+\varepsilon)/2 \\
\widetilde{u}^{[\kappa]}(dd|cc) &= (1-\kappa)\cdot(c+\varepsilon)+\kappa\cdot 1
\end{aligned}$$

and find the two conditions:

$$\begin{array}{rcl} c \geq -\kappa c/2 + c + \varepsilon/2 & \iff & \kappa \geq \frac{\varepsilon}{c} \\ c \geq \kappa(1-c-\varepsilon) + c + \varepsilon & \iff & \kappa \geq \frac{\varepsilon}{c+\varepsilon-1} \end{array}$$

The first of these two conditions was the condition for u^{α}, the other adds a constraint that, again, is the binding one if $\varepsilon < 1$.

For a numerical example, take $c = 2$ and $\varepsilon = 1/2$. For cooperation to be sustained by altruism, the condition is $\alpha \geq \varepsilon/c = 1/4$ but ex ante Homo Moralis requires the two conditions $\kappa \geq 1/4$ and $\kappa \geq \frac{\varepsilon}{c+\varepsilon-1} = 1/3$, so that for for α and κ between 1/4 and 1/3, cooperation is sustained in $u^{(\alpha)}$ and not in $\widetilde{u}^{[\kappa]}$.

Likewise, with the same example, one finds that Defection is an equilibrium for $\alpha \leq 2/3$ but requires $\kappa \leq 1/2$.

This example confirms that the conditions for universalization are strictly more stringent that the ones for the altruistic utilitarian model, the reason being that ex ante HM stability involves joint deviations that are not considered for altruistic equilibrium.

4 Further remarks

4.1 Symmetric base games and ex post universalization

The base game u considered here is a multi-player game. It is possible that u itself be a symmetric game, that is $A_1 = A_2 = A$ and for all $a_1, a_2 \in A$,

$$u_1(a_1, a_2) = u_2(a_2, a_1) \tag{1}$$

(Such is the case is the example of the previous section.) In the case of a two-player symmetric game, we may consider that the actions available to the players "are the same" and it is thus formally possible to consider the two-player game played by Homo Moralis players. This game deserves to be noted $u^{[\kappa]}$, its has action set $A_1 \times A_2 = A^2$ and is defined by:

$$\begin{array}{rcl} u_1^{[\kappa]}(a_1, a_2) & = & (1-\kappa)u_1(a_1, a_2) + \kappa u_1(a_1, a_1), \\ u_2^{[\kappa]}(a_1, a_2) & = & (1-\kappa)u_2(a_1, a_2) + \kappa u_2(a_2, a_2). \end{array} \tag{2}$$

This $u^{[\kappa]}$ is a two-player game with action space A^2 should not be confused with $\widetilde{u}^{[\kappa]}$, whose strategy space is larger (in the symmetric case, it is $A^2 \times A^2$). As previously explained, $\widetilde{u}^{[\kappa]}$ describes *ex ante Homo Moralis*, who chooses an action plan behind a veil of ignorance; but $u^{[\kappa]}$ describes *ex post Homo Moralis*, who knows what is her role in the game. What is the relation between these two games ? The following proposition shows that, for symmetric equilibria, being an equilibrium of $u^{[\kappa]}$ is a necessary condition for $\widetilde{u}^{[\kappa]}$.

Proposition 2. *Let u be a two player symmetric game and let (s^*, s^*) be a symmetric equilibrium of the κ-universalised extended game $\widetilde{u}^{[\kappa]}$ whose strategy is itself symmetric: $s^* = (a^*, a^*)$ for some action a^*. Then (a^*, a^*) is an equilibrium of $u^{[\kappa]}$ the κ-universalization of u.*

Proof. Consider a symmetric equilibrium of $\widetilde{u}^{[\kappa]}$, that is s^* such that (s^*, s^*) is a Nash equilibrium of $\widetilde{u}^{[\kappa]}$. Write $s^* = (a_1^*, a_2^*)$. The payoff $\widetilde{u}^{[\kappa]}(s, s^*)$ can be written as follows:

$$\begin{aligned}
\widetilde{u}^{[\kappa]}(a_1, a_2, a_1^*, a_2^*) &= \tfrac{1-\kappa}{2}[u_1(a_1, a_2^*) + u_2(a_1^*, a_2)] + \tfrac{\kappa}{2}[u_1(a_1, a_2) + u_2(a_1, a_2)] \\
&= \tfrac{1-\kappa}{2}[u(a_1, a_2^*) + u(a_2, a_1^*)] + \tfrac{\kappa}{2}[u(a_1, a_2) + u(a_2, a_1)].
\end{aligned}$$

Suppose moreover that the equilibrium strategy s^* is itself symmetric: $s^* = (a^*, a^*)$. Then:

$$\begin{aligned}
\widetilde{u}^{[\kappa]}(a_1, a_2, a^*, a^*) &= \tfrac{1-\kappa}{2}[u_1(a_1, a^*) + u_2(a^*, a_2)] + \tfrac{\kappa}{2}[u_1(a_1, a_2) + u_2(a_1, a_2)] \\
&= \tfrac{1-\kappa}{2}[u(a_1, a^*) + u(a_2, a^*)] + \tfrac{\kappa}{2}[u(a_1, a_2) + u(a_2, a_1)].
\end{aligned}$$

The maximum of $\widetilde{u}^{[\kappa]}(a_1, a_2, a^*, a^*)$ with respect to the pair (a_1, a_2) being reached at $a_1 = a_2 = a^*$ implies that the same maximum is reached under the constraint $a_1 = a_2 = a$. Isolating such variable a it comes:

$$\begin{aligned}
\widetilde{u}^{[\kappa]}(a, a, a^*, a^*) &= \tfrac{1-\kappa}{2}[u_1(a, a^*) + u_2(a^*, a)] + \tfrac{\kappa}{2}[u_1(a, a) + u_2(a, a)] \\
&= \tfrac{1-\kappa}{2}[u(a, a^*) + u(a, a^*)] + \tfrac{\kappa}{2}[u(a, a) + u(a, a)] \\
&= (1-\kappa)u(a, a^*) + \kappa u(a, a) \\
&= u^{[\kappa]}(a, a^*)
\end{aligned}$$

It follows that $a = a^*$ is maximizing $u^{[\kappa]}(a, a^*)$, meaning that (a^*, a^*) is a Nash equilibrium of $u^{[\kappa]}$. We conclude that any doubly symmetric equilibrium $(s^*, s^*) = ((a^*, a^*), (a^*, a^*))$ of $\widetilde{u}^{[\kappa]}$ defines a symmetric equilibrium (a^*, a^*) of $u^{(\kappa)}$. □

Using again the prisoner's dilemma as an example, the reader can check that the converse is not true, a counter-example being found for $c = 2$ and $\varepsilon = 3/2$.

4.2 Extended games played by altruistic players

Coming back to the general setting of possibly non symmetric base games, this section will consider the game that deserves to be note $(\widetilde{u})^{(\alpha)}$, that is the extended game $\widetilde{u}$ played by α-altruistic players. Writing:

$$\begin{aligned}
(\widetilde{u})^{(\alpha)}(s,s') &= (\widetilde{u})^{(\alpha)}(a_1,a_2,a'_1,a'_2) \\
&= \tfrac{1}{2}\Big(u_1(a_1,a'_2)+u_2(a'_1,a_2)\Big)+\alpha\cdot\tfrac{1}{2}\Big(u_2(a_1,a'_2)+u_1(a'_1,a_2)\Big) \\
&= \tfrac{1}{2}\Big(u_1(a_1,a'_2)+\alpha u_2(a_1,a'_2)\Big)+\tfrac{1}{2}\Big(u_2(a'_1,a_2)+\alpha u_1(a'_1,a_2)\Big) \\
&= \tfrac{1}{2}\Big(u_1^{(\alpha)}(a_1,a'_2)\Big)+\tfrac{1}{2}\Big(u_2^{(\alpha)}(a'_1,a_2)\Big)
\end{aligned}$$

one can see that:

$$(\widetilde{u})^{(\alpha)} = \widetilde{u^{(\alpha)}}.$$

Things are very simple in that case with respect to Nash equilibrium. A symmetric equilibrium (s^*,s^*) of the extended game defines a strategy s^* that is a pair of actions $s^* = (a_1^*,a_2^*)$ which forms a Nash equilibrium of the game $u^{(\alpha)}$: the veil of ignorance has simply no effect here. Altruism behind the veil of ignorance is just altruism in all cases.

One might be tempted to call (α-partial) *rule* utilitarianism an equilibrium strategy of the game $(\widetilde{u})^{(\alpha)}$ in which altruistic players chose "behind the veil of ignorance" courses of actions suited for all circumstances, by contrast with the partial *act* utilitarianism captured by $u^{(\alpha)}$. According to the remark above, such an interpretation would have the strange consequence of making spurious the distinction between the two forms of utilitarianism.

But this would be stretching Harsanyi's notion of rule utilitarianism too far. As defined for instance in Section II of Harsanyi (1992), rule utilitarianism is a socially accepted set of constraints that are decided by a cooperative process and are flexible enough to avoid the "intolerably burdensome negative implementation effects" of full act utilitarianism.

Since the cooperative process by which an utilitarian rule emerges is left unspecified by Harsanyi, it seems that, to be true to his idea, what is presented here might be better described as a form of *act* utilitarianism. It is a mild form of act utilitarian-

ism, thanks to the parameter $\alpha \in [0,1]$, but the decision process that we use remains non-cooperative and hardly captures what Harsanyi meant to say.

5 Empirics

A number of formulas have been proposed, that complete the standard Homo Economicus individualistic utility in order to increase the descriptive power of utilitarian theory. Such formulas are of current use in Behavioral Economics and in particular in Experimental Economics, and it is a challenge to determine which of these theories best render the behavior of the participants to laboratory experiments who are faced with various social dilemma games.

Such statistical exercise, that incorporates the Homo Moralis model to the tool kit of experimental game theory, has been performed by Miettinen et al. (2020) and van Leeuwen and Alger (2021). These studies demonstrate the remarkable descriptive power of the Homo Moralis model in these experiments. In order to do so, they need to use symmetric settings, as explained in the introduction, and they do so by considering two-stage symmetric extensions of two-player games.

In practice, these studies have participants playing either sequential version of already symmetric games (like the prisoners' dilemma) or small sequential games, like the dictator game played in the "strategy" form: "What would you do as a first mover? What would you do as a second mover in the various possible circumstances ?" Moreover, these papers do not rely on equilibrium assumptions (which would not be justifiable in laboratory settings) and therefore have to fit the models as good responses to beliefs about the others' strategies, beliefs which, in turn have to be elicited. This makes it difficult to inform the point made in the present paper with these laboratory experiments.

In a different vein, an interesting empirical counterpoint to the present work is provided by Oprea et al. (2011) who precisely test in the laboratory the finding of evolutionary game theory that distinguishes single-population from multi-population dynamics on the basis of their ability to materialize joint deviations. They use Hawk-Dove games, in which only joint deviations are able to stabilize the interior equilibrium. This possibility exists in a single-population, but not in standard multi-population models. Using a nice continuous-time protocol for implementing both mono- and multi-population dynamics, the empirical findings confirm the evolutionary-based predictions: convergence toward the interior equilibrium

for single-population dynamics, and converge towards boundary equilibria in multi-population dynamics. Such experiments show at work the strength of Selten's theoretical remark that is the root of the present paper.

References

[1] Ingela Alger and Jean-François Laslier (2020) "Homo moralis goes to the voting booth: coordination and information aggregation" working paper halshs-03031118.

[2] Ingela Alger and Jean-François Laslier (2021) "Homo moralis goes to the voting booth: a new theory of voter turnout" working paper halshs-03152172.

[3] Ingela Alger and Jörgen Weibull (2013) "Homo moralis: preference evolution under incomplete information and assortative matching" *Econometrica* 81: 2269—2302.

[4] Ingela Alger and Jörgen Weibull (2017) "Strategic behavior of moralists and altruists" *Games* 8: 38; doi:10.3390/g803003.

[5] Ingela Alger and Jörgen Weibull (2020) "Morality: evolutionary foundations and policy implications" in *The State of Economics, the State of the World*, edited by Kaushik Basu, David Rosenblatt, and Claudia Sepulveda, MIT Press, pp. 395—443.

[6] Abhijit Banerjee and Jörgen Weibull (2000) "Neutrally stable outcomes in cheap-talk coordination games" *Games and Economic Behavior* 32: 1—24.

[7] Gary S. Becker (1974) "A theory of social interactions" *Journal of Political Economy* 82(6): 1063—1093

[8] Ken Binmore (1994) *Playing Fair: Game theory and the social contract*. MIT Press.

[9] Matthew Braham and Martin van Hees (2020) "*Kantian* Kantian optimization" *Erasmus Journal for Philosophy and Economics* 13(2): 30—42.

[10] Nicolas Gravel, Jean-François Laslier, and Alain Trannoy (2000) "Consistency between tastes and values: a universalization approach" *Social Choice and Welfare* 17: 293—320.

[11] John Harsanyi (1955) "Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility" *Journal of Political Economy* 63: 309—321.

[12] John Harsanyi (1980) "Rule utilitarianism, rights, obligations and the theory of rational behavior" *Theory and Decision* 12: 115—133.

[13] John Harsanyi (1992) "Game and decision theoretic models in ethics" in: *Handbook of Game Theory* edited by R.J. Aumann and S. Hart, Elsevier, volume 1, pp. 669—707.

[14] Immanuel Kant (1785) *Grundlegung zur Metaphysik der Sitten*. Translation: Mary Gregor and Jens Timmermann (2011) *Groundwork of the Metaphysics of Morals: A German-English Edition*. Cambridge University Press.

[15] Yong-Gwan Kim and Joel Sobel (1992) "An evolutionary approach to pre-play communication" *Econometrica* 63: 1181—1194.

[16] Jean-Jacques Laffont (1975) "Macroeconomic constraints, economic efficiency, and ethics: An introduction to Kantian economics" *Economica* 42: 430—437.

[17] Jean-François Laslier (2003) "The evolutionary analysis of signal games" in *Cognitive Economics*, edited by Paul Bourgine and Jean-Pierre Nadal, Springer, pp. 281—291.

[18] Jean-François Laslier (2020) "Do Kantians drive others to extinction?" *Erasmus Journal for Philosophy and Economics* 13(2): 98—108. doi: 10.23941/ejpe.v13i2.501.

[19] Boris van Leeuwen and Ingela Alger (2021) "Estimating social preferences and Kantian morality in strategic interactions" working paper.

[20] Assar Lindbeck and Jörgen Weibull (1988) "Altruism and time consistency: the economics of Fait Accompli" *The Journal of Political Economy* 96(6): 1165—1182.

[21] John Maynard Smith (1982) *Evolution and the Theory of Games*, Cambridge University Press.

[22] Topi Miettinen, Michael Kosfeld, Ernst Fehr, and Jörgen Weibull (2020) "Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions" *Journal of Economic Behavior and Organization* 173: 1—25.

[23] Martin A. Nowak and Karl Sigmund (2005) "Evolution of indirect reciprocity" *Nature* 437: 1293—1295.

[24] Ryan Oprea, Keith Henwood, and Daniel Friedman (2011) "Separating the Hawks from the Doves: evidence from continuous time laboratory games" *Journal of Economic Theory* 146(6): 2206—2225.

[25] John Roemer (2019) *How We Cooperate. A Theory of Kantian Optimisation*, Yale University Press.

[26] Jean-Jacques Rousseau (1755) *Discours sur l'origine et les fondements de l'inégalité parmi les hommes*. Reprinted in : *Ecrits politiques*, 1992, Le livre de Poche.

[27] Reinhard Selten (1980) "A note on evolutionary stable strategies in asymmetric animal conflicts" *Journal of Theoretical Biology* 84: 93—101.

[28] Karl Schlag (1994) "When does evolution lead to efficiency in communication games?" Discussion paper B-299, Friedrich Wilhelms University of Bonn.

[29] Jim Sidanus and Robert Kurzban (2013) "Toward an evolutionary informed political psychology" in *The Oxford Handbook of Political Psychology* edited by L. Huddy, D. O. Sears, and J. S. Levy, Oxford University Press, pp. 205—236.

[30] Brian Skyrms (1996) *Evolution of the Social Contract*, Cambridge University Press.

[31] Franz de Waal (1996) *Good Natured: The origins of right and wrong in humans and other animals*, Cambridge University Press.

[32] Karl Wärneryd (1993) "Cheap talk, coordination and evolutionary stability" *Games and Economic Behavior* 5: 532—546.