LEMMA – Document de travail

DT 2026-04

Cooperation via Universalization in Asymmetric Social Dilemmas

Alberto Grillo

Université Paris Panthéon-Assas, LEMMA

LEMMA EA 4442 - 4 rue Blaise Desgoffe 75006 Paris

Cooperation via Universalization in Asymmetric Social Dilemmas

Alberto Grillo *

Université Paris-Panthéon-Assas, LEMMA

May 27, 2026

Abstract

This paper studies cooperation by moral agents guided by a principle of universalization (so-called *Kantian* behavior). Each agent chooses the strategy that maximizes her utility in the hypothetical case where all other agents behaved like her. I focus on social dilemmas with asymmetries, where defining alike behaviors can be difficult. I consider the strategies of different players to be alike if, when played jointly, they induce an equal or proportional division of the surplus relative to the uncooperative outcome. I show that universalizing agents can broadly coordinate on an efficient profile of alike strategies. Moreover, unlike other theories of universalization, the framework guarantees that the outcome Pareto-dominates the Nash equilibrium, ensuring that everyone benefits from cooperation.

*Contact: alberto.grillo@assas-universite.fr I thank Nicolas Gravel, Victor Hiller, Emmanuel Lagrée, Enrico Salonia, and seminar participants at AMSE (Marseille), LEMMA (Paris), University Alberto Hurtado (Santiago, Chile), the 2025 Game Theory workshop at the University of Barcelona, and the 2025 Lisbon Meetings in Game Theory and Applications for their helpful comments.

1 Introduction

Social dilemmas involve a conflict between individual and collective interests. In the classic Prisoner's dilemma, rational agents choose not to cooperate, thereby preventing the Pareto-efficient outcome. If the game is repeated, cooperation can be sustained by the threat of punishing defections in future interactions. However, many experiments show that cooperation often emerges even in the absence of such a dynamic incentive.[1] A line of research has thus sought to explain cooperation by appealing to moral reasoning, either explicitly or as an implicit driver of a social preference.[2]

In this paper, I revisit cooperation under a notion of morality that commands agents to *universalize* their behavior, i.e. to consider what would happen if the others behaved like them. A principle of universalization appears across many moral codes, both religious and civic, often in line with the *golden rule* of treating others as one would want to be treated. Many economists call such reasoning *Kantian*. In social dilemmas, universalization reflects a morality distinct from altruism, as it underpins the duty to do one's part in promoting cooperation rather than a direct concern for others' welfare.

Laffont (1975) first discussed the consequences of universalization in an economy where identical agents can contribute to a public good. By assuming that everyone contributes the same amount as themselves, agents provide the efficient level of the public good even in the absence of an appropriate tax. The symmetric positions guarantee, on the one hand, that each agent chooses the same contribution when thinking that the others would do the same, and, on the other hand, that everyone benefits equally from the efficient provision. Unlike in Laffont, however, many social dilemmas involve asymmetry among agents, which successful cooperation must take into account.

Applying universalization in asymmetric settings requires defining how different agents can behave in the same way. This paper proposes a criterion for considering agents' strategies as *alike* and shows that, if agents agree on this criterion, they can rely on universalization to coordinate on an efficient and mutually beneficial outcome. The starting point is that any judgment of alikeness presupposes a metric for comparison. I base this metric on the standard measure of the benefits from cooperation, namely the surplus relative to the uncooperative outcome. The

[1] See e.g. Roth (1988), Andreoni (1995), Ledyard (1995), Henrich et al. (2001), Fehr et al. (2002), Chaudhuri (2011).

[2] For a recent survey on social preferences, see Fehr & Charness (2025). Models of moral behavior include Brekke et al. (2003), Levitt & List (2007), Kaplow & Shavell (2007), Ellingsen & Mohlin (2025), in addition to those discussed in the literature section, which are more closely related to this paper.

criterion is therefore grounded in a bargaining framework and evaluates strategies according to the division of the surplus induced by their joint play.

As a first step, I define strategies to be alike if they induce an equal division of the surplus, i.e. an equal gain to all agents. In this case, I show that universalization allows agents to coordinate on a profile of alike strategies that Pareto-dominates the Nash equilibrium, from which the surplus is calculated. The resulting payoffs correspond to the egalitarian solution of the associated bargaining game, which is Pareto-efficient under regular conditions on the set of feasible payoffs.

The previous result marks a departure from the existing literature. Indeed, while both Bilodeau & Gravel (2004) and Roemer (2015, 2019) already argued that universalization can yield efficiency even outside the symmetric environment, the resulting outcome in their models does not always Pareto-dominate the Nash equilibrium. In my view, this is an important limitation for the claim that universalization enables cooperation, insofar as the essence of cooperation lies in its mutual benefits for the agents involved. In my framework, instead, all agents are better off by collectively following universalization than at the Nash equilibrium.

As a second result, I show that universalizing agents can still coordinate on a cooperative outcome if strategies are considered alike whenever they induce a division of the surplus in any fixed proportions. Specifically, the analysis yields the following generalization. Consider any strategy profile that is efficient and Pareto-superior to the Nash equilibrium in a social dilemma. By definition, such a profile induces a division of the cooperative surplus in some given proportions. I show that universalizing agents can coordinate on such a profile, if they regard as alike any strategies that induce payoff gains (relative to the Nash equilibrium) in those same proportions.

From a methodological perspective, however, defining alike strategies based on the payoff consequences that they generate rather than on the strategies themselves comes with a cost, since payoff consequences differ across utility functions representing the same preferences. In particular, interpreting alike strategies as those yielding an equal division of the surplus requires interpersonal comparisons of utilities. Instead, the interpretation in terms of a proportional division to the one induced by a given profile requires a cardinal notion of utility and that the profile is itself invariant to cardinal transformations, but does not require any interpersonal comparison. I relate these remarks to the difficulty of defining a bargaining solution based only on ordinal and non-comparable preferences.

The remainder of the paper is organized as follows. I end the introduction by reviewing the previous literature and by proposing a motivating example. Section 2 lays out the technical framework and defines all concepts. Section 3 gives the argument for interpreting alike behavior through the lens of the associated bargaining game, presents the main results, and reconsiders the motivating example. Section 4 concludes.

1.1 Review of the literature

Following Laffont (1975), so-called Kantian rules of behavior have been discussed by Sugden (1984), Bordignon (1990), and Bergstrom (1995).[3] In an important line of work, Alger & Weibull (2013, 2016) and Alger et al. (2020) showed that, under evolutionary dynamics with incomplete information and assortative matching, the only stable preferences are those of a *Homo Moralis* agent, who maximizes a weighted average of her own payoff and the payoff that would result if everybody else acted likewise. Yet, the Homo Moralis framework has not been formulated in asymmetric games: in that case, the authors simply study the ex-ante symmetric game in which all players have the same probability of occupying each role.

Bilodeau & Gravel (2004) first argued that universalizing agents in asymmetric positions can reach an efficient outcome in public good games. Specifically, they showed that, under some technical conditions, it is possible to partition the strategies of different players into classes of equivalent contributions such that each agent contributes her share of the Lindahl equilibrium, when thinking that the others contribute equivalently. The authors do not specify which contributions count as equivalent, other than those corresponding to the Lindahl equilibrium, but prove that such a partition exists. As is well known, the Lindahl equilibrium is efficient, but does not always Pareto-dominate the Nash equilibrium. My analysis is inspired by Bilodeau & Gravel (2004) in the framing of the problem. However, by adding a precise meaning for alike behavior, I argue that universalization is not restricted to yielding the Lindahl equilibrium as an outcome and is fully compatible with the attainment of mutual benefits.

Roemer (2010, 2015, 2019) has proposed a broader theory of *Kantian Optimization*. He extends universalization to asymmetric games in a way that applies beyond public good games whenever strategy sets are continuous. The idea is to shift the focus from same strategies to same deviations from a given strategy profile. Roemer (2015, 2019) conceives same deviations as

[3] In economics, such a Kantian approach implies that agents choose their action based on the consequences produced by universalization; yet, this perspective is closer to rule-consequentialism than to Kant's categorical imperative. The contributions by Sen (1977) and Harsanyi (1980) are also highly relevant to the discussion of moral behavior in economics.

either by the same proportion or by the same additive term. A multiplicative (additive, resp.) Kantian equilibrium is then defined as a profile from which no agent would want to deviate, if all other agents deviated by the same proportional (additive, resp.) factor. The main result of Roemer's theory is that, in games in which the effect of the externality is monotone, Kantian equilibria are Pareto-efficient. As with the Lindahl equilibrium, however, a Kantian equilibrium need not Pareto-dominate a Nash equilibrium. More fundamentally, whether deviations by the same multiplicative (additive, resp.) factor can be reasonably regarded as equivalent depends on the payoff functions: in general, proportional deviations from a profile may well result in very different consequences for the agents involved. My approach shares with Roemer the idea of starting from a precise definition of equivalent behavior, but adresses the previous points by defining equivalence through a comparison in the space of payoffs rather than in the space of strategies.

In parallel work, Salonia (2025) develops an axiomatic model characterizing a preference for universalization. Notably, he also interprets same behavior as the one inducing the same utility consequences, which he defines as equal sacrifice from a maximum attainable payoff. However, these utility consequences are calculated separately for each individual, by fixing the opponent's strategy, and thus the joint play of such strategies need not yield the same utility consequences. As a result, under this interpretation, universalization neither guarantees efficiency nor the attainment of mutual benefits.

Overall, such a variety of approaches may simply indicate that universalization indeed admits multiple interpretations: the one proposed here is particularly appealing as a grounding principle for cooperation in social dilemmas.[4]

1.2 Motivating example

Consider a game of public good provision, of the kind studied by Bergstrom et al. (1986). There are n agents, each endowed with exogenous income ω_i. Each agent can contribute an amount $z_i \in [0, \omega_i]$ to the public good, whose total provision is the sum of contributions, i.e. $Z = \sum_j z_j$. The income not allocated to the public good is used for the consumption of a private good $x_i = \omega_i - z_i$. Each agent has standard (increasing and quasi-concave) preferences for the public and the private goods, represented by a utility function $u_i(Z, x_i)$. Let us analyze the following

[4]To conclude the literature review, several papers have applied the universalization framework, in either the Homo Moralis or the Kantian equilibrium variant, to different domains, namely climate change (Grafton et al., 2017), tax competition (Eichner & Pethig, 2020), voting (Alger & Laslier, 2022), and vaccination (De Donder et al., 2025). Van Leeuwen & Alger (2024) and Benabou et al. (2024) provide recent evidence of deontological behavior in experiments.

specification of the game.

Example 1. *Consider the game described above, with $n = 2$, incomes $\omega_1 = 1$ and $\omega_2 = 3$, and the same utility function for both agents given by*

$$u_i(Z, x_i) = Z \cdot x_i \quad \forall i \in \{1, 2\}$$

The game in Example 1 has a unique Nash equilibrium, namely $z_1 = 0$, $z_2 = 1.5$ where player 2 contributes half of her income and player 1 free-rides. The total provision of the public good is $Z = 1.5$ and the equilibrium payoffs are $u_1 = 1.5$ and $u_2 = 2.25$. Clearly, the Nash equilibrium is inefficient, because agents do not take the reciprocal externality into account. The Pareto frontier of the game is indeed characterized by the condition that $Z \geq 2$, and specifically $Z = 2$ if $z_1 > 0$.[5]

Can agents overcome the collective action problem if they are guided by universalization? An answer to this question requires specifying what it means for each agent to universalize her behavior. Note first that if each agent simply chose the contribution that she would want both agents to make, then the two agents would contribute different amounts. That is, the asymmetry in incomes would raise a problem of consistency between their reasoning and the outcome that they reach under this reasoning. Universalization, in this case, could not serve as the basis of a coordinated cooperative play.

The alternative is to find a convincing metric according to which different contributions of the two agents can be considered alike (or equivalent). In the context of this public good game, for example, one may be tempted to consider contributions that are in the same proportion of income as equivalent. Suppose then that each agent chose the fraction of income that she would want both agents to contribute. In this case, both agents would consistently choose a contribution equal to half of their income. This corresponds to $z_1 = 0.5$, $z_2 = 1.5$, which ensures an efficient provision of the public good ($Z = 2$) and yields payoffs $u_1 = 1$ and $u_2 = 3$. Yet, player 1 is worse off at this outcome than at the Nash equilibrium: she could then reasonably question whether cooperating is in her interest, as opposed to not cooperating.

It turns out that neither Bilodeau and Gravel's nor Roemer's frameworks can improve on the previous concern. Indeed, both the Lindahl equilibrium and the multiplicative Kantian equilibrium of the game coincide with the previous profile $z_1 = 0.5$, $z_2 = 1.5$ (with Lindahl prices $p_1 = 0.25$ and $p_2 = 0.75$ for the Lindahl equilibrium), while an additive Kantian equilibrium

[5] All results concerning this example are proven in Appendix A.

does not exist. As such, at least in this example, the existing frameworks of universalization appear inadequate with respect to the premise of cooperation for mutual benefit. Figure 1 shows these different equilibria in the space of feasible payoffs.

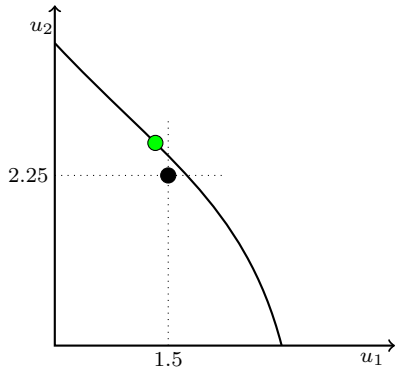

Figure 1: In black the Nash equilibrium, in green the Lindahl and multiplicative Kantian equilibrium.

How can agents adopt a universalizing stance to coordinate on achieving mutual benefits? The following sections propose a framework to do so.

2 Framework

Consider a static game $(N, (S_i)_{i\in N}, (u_i)_{i\in N})$, where N is the set of players, S_i is player i's strategy set, and $u_i : \prod_{j\in N} S_j \to \mathbb{R}$ her payoff function. Following standard notation, for any player i, denote S_{-i} the set $\prod_{j\in N/\{i\}} S_j$ and denote s_{-i} an element of this set, i.e. a vector of $n-1$ strategies, one for each of the other players. I restrict attention to pure strategies.[6]

2.1 Likeness relations

I call *likeness relation* the relation between strategies of different agents which enables them to apply the thought-experiment of universalization. Formally, a likeness relation is a finitary relation $\mathcal{R}$ on the sets $S_1 \times ... \times S_n$, which defines whether any n strategies $(s_1, ..., s_n)$, one per player, are *alike*. Only strategies in vectors of n components can be (or not) alike according to some $\mathcal{R}$: in particular, if $n \geq 3$, one cannot tell whether two strategies s_i for player i and s_j for player j are alike without specifying some strategies for the remaining $n-2$ players.

[6]This is for expositional clarity. The framework could be extended to the case where mixed strategies are allowed, and one could then judge mixed strategies as alike according to the expected division of the surplus that they induce when played jointly. Yet, doing so would add technical difficulty without yielding substantive insight.

For any player i, a likeness relation $\mathcal{R}$ induces a binary relation from the set S_i to the set S_{-i}, which maps strategies s_i into vectors s_{-i} of alike strategies, one for each of the other players. Denote $\mathcal{R}_i$ the induced binary relation for player i. Denote $S_i^{\mathcal{R}}$ the domain of definition of $\mathcal{R}_i$, i.e. the subset of S_i given by all strategies with at least one vector of alike strategies. Denote $S_{-i}^{\mathcal{R}}$ the codomain of definition of $\mathcal{R}_i$, i.e. the subset of S_{-i} given by vectors of strategies that are alike to some strategy s_i.

Example 2. *Consider a three-player game. Strategy sets are* $S_1 = \{a_1, b_1, c_1\}$, $S_2 = \{a_2, b_2\}$, *and* $S_3 = \{a_3, b_3\}$. *An example of a likeness relation on* $S_1 \times S_2 \times S_3$ *is*

$$\mathcal{R} = \Big\{(a_1, a_2, a_3), (b_1, b_2, a_3), (c_1, a_2, a_3)\Big\}$$

This relation contains three vectors of alike strategies: e.g. strategies a_1 *for player* 1, a_2 *for player* 2, *and* a_3 *for player* 3 *are alike.*

For player 1, *the domain of definition of the induced* $\mathcal{R}_1$ *is* $S_1^{\mathcal{R}} = \{a_1, b_1, c_1\}$ *and the codomain of definition is* $S_{-1}^{\mathcal{R}} = \{(a_2, a_3), (b_2, a_3)\}$.

For player 2, *the domain of definition of the induced* $\mathcal{R}_2$ *is* $S_2^{\mathcal{R}} = \{a_2, b_2\}$ *and the codomain of definition is* $S_{-2}^{\mathcal{R}} = \{(a_1, a_3), (b_1, a_3), (c_1, a_3)\}$.

For player 3, *the domain of definition of the induced* $\mathcal{R}_3$ *is* $S_3^{\mathcal{R}} = \{a_3\}$ *and the codomain of definition is* $S_{-3}^{\mathcal{R}} = \{(a_1, a_2), (b_1, b_2), (c_1, a_2)\}$.

It is important to stress that, for a given likeness relation $\mathcal{R}$ and a player i, $\mathcal{R}_i$ may not map all strategies and may not be a function. That is, there may exist strategies s_i without any vector s_{-i} of alike strategies, as well as strategies with more than one vector of alike strategies. I classify likeness relations as follows.

Definition 1. *An induced binary relation* $\mathcal{R}_i$ *has* ***full domain*** *if all strategies* s_i *have at least one vector of alike strategies. That is, if* $S_i^{\mathcal{R}} = S_i$. *By extension, a finitary relation* $\mathcal{R}$ *has full domain (on all its coordinates) if, for all* i, *the induced* $\mathcal{R}_i$ *has full domain.*

Definition 2. *An induced relation* $\mathcal{R}_i$ *is* ***functional*** *(or a function) if all strategies* $s_i \in S_i^{\mathcal{R}}$ *have a unique vector of alike strategies. By extension, a finitary relation* $\mathcal{R}$ *is functional if, for all* i, *the induced* $\mathcal{R}_i$ *is functional.*

A functional $\mathcal{R}$ induces injective functions, because if all induced $\mathcal{R}_i$ are functions, then they must all be injective.

Example 3. *Continuing from Example 2:*
$\mathcal{R}_1$ *has full domain and is functional;*
$\mathcal{R}_2$ *has full domain but is not functional, because it maps* a_2 *to both* (a_1, a_3) *and* (c_1, a_3)*;*
$\mathcal{R}_3$ *does not have full domain, because it does not map* b_3 *to anything, and it is not functional, because it maps* a_3 *to* (a_1, a_2)*,* (b_1, b_2)*, and* (c_1, a_2)*.*
Hence, the likeness relation $\mathcal{R}$ *neither has full domain, nor is functional.*

The following definition is inspired by Bilodeau and Gravel (2004).

Definition 3. *A finitary relation* $\mathcal{R}$ *is* ***tight*** *if it has full domain and is functional.*

A tight $\mathcal{R}$ cannot exist in a game in which the cardinality of players' strategy sets is not the same for all players. If all players have the same strategy set, there exists an intuitive tight $\mathcal{R}$, namely the one defining the same strategies as alike.

Example 4. *If* $S_i = S$ *for all* i*, the relation* $\mathcal{R} = \left\{(s, ..., s)_{s \in S}\right\}$ *is tight.*

Given a likeness relation $\mathcal{R}$, a universalizing agent envisions the scenario in which her choice of any strategy results in the choice of alike strategies by the other players. This scenario exists and is unique for every player if $\mathcal{R}$ is tight. If it is not tight, one should clarify what is envisioned by player i whenever (i) $\mathcal{R}_i$ does not have full domain and she plays a strategy which does not have any vector of alike strategies, or (ii) $\mathcal{R}_i$ is not functional and she plays a strategy which has more than one vector of alike strategies. To answer (i), I assume that universalization restricts the choice of each player i within $S_i^{\mathcal{R}}$, i.e. only among the strategies with at least one vector of alike strategies. This means that a player cannot play a strategy that has no alike strategies for the other players. To answer (ii), I define the following selection mechanism.

Definition 4. *Given a binary relation* $\mathcal{R}_i$*, a* ***selection*** σ *is a function from* $S_i^{\mathcal{R}}$ *to* $S_{-i}^{\mathcal{R}}$ *which, viewed as a binary relation, is a subset of* $\mathcal{R}_i$*.*

That is, for any strategy $s_i \in S_i^{\mathcal{R}}$, σ selects one and only one vector of alike strategies. Clearly, if $\mathcal{R}_i$ is functional, it allows for only one selection, which coincides with $\mathcal{R}_i$. For any strategy s_i, I refer to the selected strategies as the corresponding alike strategies: agent i envisions that if she played s_i, all other players would play their corresponding alike strategies in the vector $\sigma(s_i) \in S_{-i}^{\mathcal{R}}$.

2.2 Optimization problem

Given a likeness relation $\mathcal{R}$ and a set of selections $\{\sigma_i\}_{i \in N}$, each agent maximizes her utility under the constraint that all other players play their corresponding alike strategy. That is,

agent i solves

$$\max_{s_i \in S_i^{\mathcal{R}}} u_i\left(s_i, \sigma(s_i)\right) \tag{1}$$

Denote s_i^* a solution to the optimization problem for player i. Denote $s^* = (s_1^*, ..., s_n^*)$ a profile of solutions, one for each player. I am interested in whether, for any appealing likeness relation and set of selections, the profile s^* satisfies standard properties, namely Pareto-efficiency, Pareto-dominance with respect to the the Nash equilibrium, and a plausible notion of consistency of players' behavior.

2.3 Consistency

I propose the following notion of consistency.

Definition 5. *A profile $(s_1^*, ..., s_n^*)$ of solutions to the problem in (1) is **consistent** under the likeness relation $\mathcal{R}$ and the set of selections $\{\sigma_i\}_{i \in N}$ if*

- $(s_1^*, ..., s_n^*) \in \mathcal{R}$ *and*
- *for every player i, $\sigma(s_i^*) = s_{-i}^*$*

The first part of the definition requires that, if all players consider that the others behave like them, the result of their interaction is a profile of strategies that are indeed alike. I interpret it as a requirement that, if players agree on which of their strategies are alike, they can coordinate on how to play the game via universalization. The second part adds the requirement that the resulting profile must coincide with the outcome envisioned by all players, when each considers playing her strategy in the profile. This implies that players' coordination is based on a correct understanding of each other's behavior. If the likeness relation is functional, the second requirement is always verified. Instead, if the likeness relation is not functional, a strategy profile can satisfy the first part of the definition but not the second, as shown in Example 5 (appendix B).

Let me now stress one intuitive point. In a symmetric game,[7] any profile of solutions $(s^*, ..., s^*)$, i.e. composed of the same solution for all players, is consistent under the tight relation $\mathcal{R} = \{(s, ..., s)_{s \in S}\}$. In this case, consistency holds trivially, because players are in identical positions and choose the same optimal strategy.[8] Instead, if the strategy sets are the

[7] A game is symmetric if $\forall i$, $S_i = S$ and for any permutation π of the players, $u_{\pi(i)}(s_1, ..., s_n) = u_i(s_{\pi(1)}, ..., s_{\pi(n)})$.

[8] To clarify, even in a symmetric game for the tight relation $\mathcal{R} = \{(s, ..., s)_{s \in S}\}$, if problem (1) has multiple solutions, the profiles of solutions in which some players play different solutions are not consistent. Only the profiles composed of the same solution for all players are consistent. If problem (1) has a unique solution, then there is a unique profile of solutions, which is consistent.

same but the game is not symmetric, it may well be the case that no profile of solutions is consistent under the tight relation $\mathcal{R} = \{(s, ..., s)_{s \in S}\}$, as in Example 6 (appendix B).

2.4 Implementation

The next definition sets out the object of the following analysis.

Definition 6. *A profile $(s_1, ..., s_n)$ is **implementable** by universalizing agents under a given likeness relation $\mathcal{R}$ if there exists a set of selections $\{\sigma_i\}_{i \in N}$ such that:*

- *for every i, s_i is a solution to the problem in (1), and*
- *the profile is consistent, according to Definition 5.*

Consider now a social dilemma, by which I mean any game with a unique (or focal) Nash equilibrium that is inefficient. I refer to any profile that is efficient and Pareto-superior to this Nash equilibrium as a cooperative outcome, in the sense of a potential result of cooperation among agents. The question is which cooperative outcomes, if any, are implementable under a compelling likeness relation. My view is that any compelling generalization of universalization cannot abstract from clarifying its content, namely what behaving alike means for agents in asymmetric positions. In the next section, I advance a specific proposal in this direction.

3 A proposal for $\mathcal{R}$

I argue for generalizing universalization in social dilemmas along two principles. The first holds that the alikeness of agents' behavior should be judged by the alikeness of the payoff consequences that the behavior produces. This is a welfarist principle, in that only the payoffs resulting from players' strategies, not the strategies themselves, are relevant for considering the strategies alike. The welfarist perspective does not by itself specify which consequences count as alike, but suggests shifting the focus from the set of strategies that players may choose to the set of feasible payoffs that these strategies generate.

The second principle calls for treating any social dilemma as a bargaining game over the generated surplus relative to the uncooperative outcome. This principle reflects a contractarian view of cooperation, as the outcome of an agreement on the division of the surplus. I take the Nash equilibrium payoff to be the disagreement point of the bargaining game, thereby interpreting the uncooperative outcome as the case in which players fail to reach an agreement.

This formulation, often associated with Buchanan (1975), builds on the contractarian tradition of taking the *state of nature*, where each agent pursues her self-interest, as the benchmark with respect to which the social contract is negotiated. The implication of the second principle is that the relevant consequences for the judgment of alikeness, in a social dilemma, concern the division of the surplus among the players.

My proposal, therefore, is to consider players' strategies alike if they induce a division of the surplus that is equivalent. Clearly, it remains to clarify what it means for the division of the surplus to be equivalent. I first interpret an equivalent division as an equal division and then as a proportional division. In the second case, the judgment of equivalence is made with respect to a reference profile from which the proportions are calculated. I discuss the implications from these two interpretations in turn.

3.1 Equal division of the surplus

The simplest interpretation of an equivalent division of the surplus is the equal division. This interpretation is embedded in the following likeness relation, which I call the equal-gain relation.

Definition 7. *The **equal gain (EG)** relation – denoted $\mathcal{R}^{EG}$ – is the likeness relation according to which strategies $(s_1, ..., s_n)$ are alike if and only if they give all players an equal payoff gain with respect to the Nash equilibrium payoff $(u_1^N, ..., u_n^N)$.*

The EG relation is implicitely defined by the following system of $n-1$ equations:

$$u_1(s_1, ..., s_n) - u_1^N = u_2(s_1, ..., s_n) - u_2^N = ... = u_n(s_1, ..., s_n) - u_n^N \tag{2}$$

The counterfactual reasoning of an agent universalizing behavior under $\mathcal{R}^{EG}$ corresponds to asking: "which strategy would I take if all other players took the strategies that, together with mine, would result in the same payoff gain for everyone?"

The $\mathcal{R}^{EG}$ relation is non-empty by construction, since it contains at least the Nash equilibrium profile. How many other profiles belong to it depends on the specifics of the game in terms of strategy space and payoff functions. In general, the $\mathcal{R}^{EG}$ relation need not have full domain or be functional and, therefore, it may require specifying agents' selections $\{\sigma_i\}_{i \in N}$. But by construction, whenever the induced binary relation is not functional for some player, all possible selections have the following character of impartiality: for any strategy s_i of a player i

and for any two selections $\sigma(s_i)$ and $\sigma'(s_i)$, the difference

$$u_j(s_i, \sigma(s_i)) - u_j(s_i, \sigma'(s_i))$$

has the same value for all players $j \in N$. This means that, if a player has a higher payoff under one selection than under another, the payoff of all other players is also higher under the same selection. Therefore, the selection of the vector of alike strategies does not conflict with the cooperative objective, in the sense that no player can envision a profile that is advantageous for herself but disadvantageous for another player (in terms of hypothetical consequences).

Under this first interpretation, universalization allows agents to coordinate on an outcome that guarantees mutual benefits and is efficient under regular conditions of the set of feasible payoffs, namely comprehensiveness and the fact that its upper boundary coincides with the Pareto frontier. By definition, a set $\mathcal{X}$ in a real space is comprehensive if $x \leq y$ (i.e. each component in x weakly smaller than the associated component in y) and $y \in \mathcal{X}$ implies $x \in \mathcal{X}$. Comprehensiveness implies that the upper boundary of the set is continuous.[9] The fact that the upper boundary coincides with the Pareto frontier rules out any flat horizontal or vertical part on the upper boundary.

Proposition 1. *The profiles implementable under $\mathcal{R}^{EG}$ are those whose payoffs correspond to the egalitarian solution of the associated bargaining game, in which the Nash equilibrium payoff is the disagreement point. These profiles Pareto-dominate the Nash equilibrium unless they coincide with it, and are efficient if the set of feasible payoffs is comprehensive and its upper boundary coincides with the Pareto frontier.*

In the following proof, the first claim relies on a strong graphical intuition; the second on the fact that, if the set of feasible payoffs is comprehensive, the egalitarian solution lies on its upper boundary.

Proof of Proposition 1. In their optimization problem, agents are constrained to choose strategies such that the resulting payoff lies on the line that passes through the Nash equilibrium payoff and has a direction vector with all components equal to one.[10] Since these components are all strictly positive, all agents maximize their utility at the same point on the line, the one with the highest coordinates within the set of feasible payoffs. For any strategy profile whose

[9] The upper boundary of a set $\mathcal{X}$ in a real space is the set of all $x \in \mathcal{X} : \nexists\, y \in \mathcal{X}, y > x$.

[10] This is the line passing through $(u_1^N, ..., u_n^N)$ along which all coordinates increase equally, i.e. the 45° line if $n = 2$.

payoff corresponds to such a point, there must exist a set of selections for which all agents choose their corresponding strategies. Hence, those profiles are implementable. The point with the highest coordinate on the line corresponds to the egalitarian solution of the associated bargaining problem, in which the Nash equilibrium payoff is taken as disagreement point, and any corresponding strategy profile Pareto-dominates the Nash equilibrium by construction unless it coincides with it. If the set of feasible payoffs is comprehensive, its upper boundary must be crossed by the previous line. The egalitarian solution lies then on the upper boundary and, if the upper boundary coincides with the Pareto frontier, any corresponding profile is efficient. □

I do not wish to present the previous result as predictive of behavior, as in practice people may balance the moral principle against other considerations. Yet, this interpretation does connect to the observation that an equal division often appears as a focal point in experiments (Dawes et al., 2007; Agranov et al., 2025). In my framework, the egalitarian perspective is applied only to the cooperative surplus, not to the total welfare of players. An important aspect thus concerns whether the uncooperative outcome, from which the surplus is calculated, can be regarded as a fair benchmark for the subsequent division of the surplus. The more this is the case, the more one can expect the equal division of the surplus to acquire focality for the involved agents.

3.2 Proportional division of the surplus

In Proposition 1, the claim that the considered profiles are consistent, hence implementable, relies only on the fact that, in the space of payoffs, the constraint of universalization is represented by a line whose direction vector has positive components. Specifically, when such a constraint imposes equal gains, these components are all equal to one. But any other line whose direction vector has positive components would yield an analogous result (that any profile corresponding to the highest point on the line is implementable). Any such line corresponds to a constraint imposing gains in some fixed proportions between the agents. Hence, even if one defines strategies to be alike whenever the surplus is divided in any fixed proportions, then again an implementable profile exists, which Pareto-dominates the Nash equilibrium unless it coincides with it, and which is efficient if the set of feasible payoffs is comprehensive and its upper boundary coincides with the Pareto frontier.[11]

[11] This is even more generally true for any curve whose tangent vector has strictly positive components at every point.

The previous consideration also implies that any efficient profile that Pareto-dominates the Nash equilibrium is implementable under a corresponding likeness relation that treats strategies as alike if they divide the surplus in the same proportions as the profile. This change of perspective offers the following alternative interpretation. Suppose that players agree to view a specific (efficient and Pareto-superior to the Nash equilibrium) profile as the *target* of their cooperation. If the game is discrete, for example, such a profile could be the unique profile that exhausts the gains from cooperation. In a continuous game, instead, the profile could be focal for its correspondence to one (or more) bargaining solution. In the presence of a target, an equivalent division of the surplus can be interpreted relatively to the division induced at the target. This means considering players' strategies alike if, when played jointly, they yield payoff gains in the same proportions as those at the target. I call such a likeness relation an $\hat{s}$-target relation.

Definition 8. *Given a profile $(\hat{s}_1, ..., \hat{s}_n)$ whose payoff is $(\hat{u}_1, ..., \hat{u}_n)$ and a Nash equilibrium payoff $(u_1^N, ..., u_n^N)$, the $\hat{s}$-**target** relation – denoted $\mathcal{R}^{\hat{s}}$ – is the likeness relation according to which strategies $(s_1, ..., s_n)$ are alike if and only if there exists $\lambda \in \mathbb{R}$ such that*

$$\begin{pmatrix} u_1(s_1, ..., s_n) \\ u_2(s_1, ..., s_n) \\ \vdots \\ u_n(s_1, ..., s_n) \end{pmatrix} = \begin{pmatrix} u_1^N \\ u_2^N \\ \vdots \\ u_n^N \end{pmatrix} + \lambda \begin{pmatrix} \hat{u}_1 - u_1^N \\ \hat{u}_2 - u_2^N \\ \vdots \\ \hat{u}_n - u_n^N \end{pmatrix} \tag{3}$$

The condition in (3) states that, for every player i, the generated surplus $u_i(s_1, ..., s_n) - u_i^N$ must be a proportion λ of the surplus $\hat{u}_i - u_i^N$ that the player would obtain at the target profile. A $\mathcal{R}^{\hat{s}}$ relation is again non-empty by construction, since it contains at least the Nash equilibrium profile (for which $\lambda = 0$) and the target profile ($\lambda = 1$). Like the $\mathcal{R}^{EG}$ relation, a $\mathcal{R}^{\hat{s}}$ relation need not have full domain or be functional, and all possible selections have the same previous character of impartiality.

The counterfactual reasoning of an agent universalizing behavior under $\mathcal{R}^{\hat{s}}$ corresponds to asking: "which strategy would I take if all other players took the strategies that, together with mine, bring everyone at the same (utility) distance from our cooperative target?" Framed in this way, the intuitive result is that each agent would want to cover the whole distance to the target, in order to fully exploit the gains from cooperation.

Proposition 2. *In a social dilemma, any efficient profile $\hat{s} = (\hat{s}_1, ..., \hat{s}_n)$ which Pareto-dominates the Nash equilibrium is implementable under $\mathcal{R}^{\hat{s}}$, the corresponding $\hat{s}$-target relation.*

The proof of the result relies on the same strong graphical intuition as before.

Proof of Proposition 2. In their optimization problem under $\mathcal{R}^{\hat{s}}$, agents are constrained to choose strategies such that the resulting payoff lies on the line connecting $(u_1^N, ..., u_n^N)$ and $(\hat{u}_1, ..., \hat{u}_n)$ in the set of feasible payoffs. Since the direction vector $(\hat{u}_1 - u_1^N, ..., \hat{u}_n - u_n^N)$ of such a line has strictly positive components, all agents maximize their utility at the point with the highest coordinates on the line, namely $(\hat{u}_1, ..., \hat{u}_n)$. Hence, there must exist a set of selections for which all agents choose the strategies in the efficient profile $\hat{s}$. □

The appeal of $\mathcal{R}^{\hat{s}}$ as a likeness relation ultimately hinges on the appeal of the profile $\hat{s}$ as a cooperative target. In a given game, this can be of course a matter of opinions. Moreover, the agreement on the target must already presuppose some coordination among the agents. Yet, the simple takeaway is that any efficient outcome granting mutual benefits to all agents can result from the application of universalization, insofar as agents take the distance from the profile (in utility terms) to be the relevant metric for comparison in their universalization process. This is in stark contrast with the existing universalization frameworks in the literature, which, when applied to social dilemmas, do not guarantee mutual benefits.

3.3 Underlying assumptions of cardinality and comparability

I should stress that constructing the likeness relation indirectly in the space of feasible payoffs, rather than directly in the space of strategies, comes at a methodological cost. Indeed, in the $\mathcal{R}^{EG}$ relation, the requirement for payoff gains to be meaningfully equal is that players' utilities are to some degree inter-personally comparable. The precise notion of comparability that is required depends on the specifics of the game with respect to the Nash equilibrium payoff. In the class of games in which the (unique or focal) Nash equilibrium gives every player the same payoff, ordinal comparability of players' utilities is sufficient. The formal statement is that, in this case, $\mathcal{R}^{EG}$ is invariant to any common strictly monotone transformation of players' utility functions, since equality of payoff gains translates into equality of payoffs. Instead, in the general case in which the Nash equilibrium payoff does not give all players the same payoff, cardinal comparability is required. In this case, $\mathcal{R}^{EG}$ is invariant only to common affine transformations of players' utility functions of the type $t(u_i) = a\,u_i + b$, since equality of payoff gains requires defining a cardinal structure on which the gains are measured.

Instead, the construction of a $\mathcal{R}^{\hat{s}}$ relation does not require inter-personal comparability, as long as utilities have a cardinal meaning. In particular, if $\hat{s}$ is a strategy profile corresponding to a bargaining solution that satisfies scale-invariance (such as the Nash or the Kalai-Smorodinsky solutions), then the set $\mathcal{R}^{\hat{s}}$ is also scale-invariant: it does not change for any affine transformation $t_i(u_i) = a_i\, u_i + b_i$ of players' utilities. In this case, cardinality is encoded in the axiom of scale-invariance which restricts admissible transformations to affine functions, but comparability is not required as the transformations can be different across players.

The previous observations raise the question of whether one can define a likeness relation that is invariant to any ordinal and independent transformations of utilities. If it were to follow from the same principles stated at the beginning of Section 3, this would require defining an ordinal meaning for an equivalent division of the surplus. If the social dilemma is a simple discrete game, one could hope to base such a definition only on Pareto-improvements: e.g. if all profiles that Pareto-dominate the Nash equilibrium can be ordered in terms of Pareto-improvements, one could possibly regard the corresponding strategies as alike. Yet, in a social dilemma in which strategy sets are continuous and the set of feasible payoff is compact, being able to do so may be very challenging. This is strictly related to the question of existence of a (non-trivial) ordinal bargaining solution, which has been proved to be impossible in the case of two agents (Shapley, 1969).

3.4 Reconsidering Example 1

Let us reassess the public good game described in the introduction. Recall that the inefficient Nash equilibrium of the game is $z_1 = 0$, $z_2 = 1.5$, yielding payoffs $u_1 = 1.5$, $u_2 = 2.25$, while the Pareto frontier is characterized by the condition $Z = 2$ if $z_1 > 0$. In the space of payoffs, this condition is represented by the straight line $u_2 = 4 - u_1$, for $u_1 \in [0, 2]$, as shown in Figure 2. The value of u_1 at the point on the frontier at which $u_2 = u_2^N$ is $u_1 = 1.75$, while the value u_2 at the point on the frontier at which $u_1 = u_1^N$ is $u_2 = 2.5$. Hence, the set of feasible payoffs that are greater than the Nash equilibrium payoff for both players is symmetric. It follows that, in the associated bargaining problem, all standard bargaining solutions identify the same pair of payoffs as solution, namely $u_1 = 1.625$, $u_2 = 2.375$, at which the cooperative surplus is shared equally. There is a unique profile of contributions in the game corresponding to the previous pair of payoffs, namely $z_1 = 0.1875$, $z_2 = 1.8125$. Consider then choosing a profile of contributions $\hat{s} = (\hat{z}_1, \hat{z}_2)$ as a reasonable cooperative target. If one were to justify the choice of such a profile by its correspondence with any well-established bargaining solution, the only

candidate would be the profile $\hat{s} = (0.1875, 1.8125)$. Clearly, the corresponding $\mathcal{R}^{\hat{s}}$ relation coincides with the equal-gain relation:

$$\mathcal{R}^{EG} = \Big\{(z_1, z_2) : (z_1 + z_2)(1 - z_1) - 1.5 = (z_1 + z_2)(3 - z_2) - 2.25\Big\} \tag{4}$$

The condition above defines an equation of second degree in both z_1 and z_2, whose solutions z_2 as a function of z_1 are

$$z_2 = 1 \pm \frac{1}{2}\sqrt{1 + 8z_1 + 4z_1^2} \tag{5}$$

Figure 3 represents these solutions in the space of admissible contributions $z_1 \in [0, 1]$ and $z_2 \in [0, 3]$.

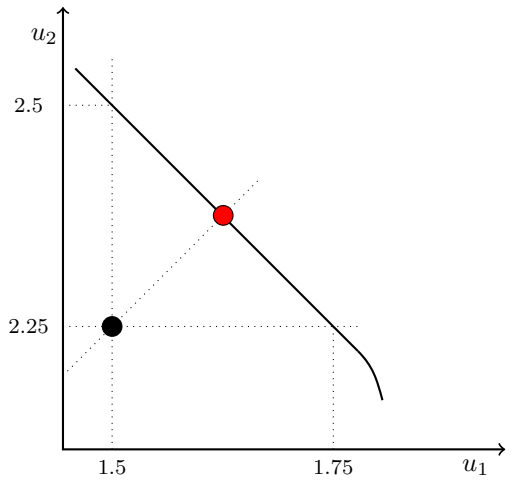

Figure 2: The set of feasible payoffs that Pareto-dominate the Nash equilibrium and the egalitarian solution (in red).

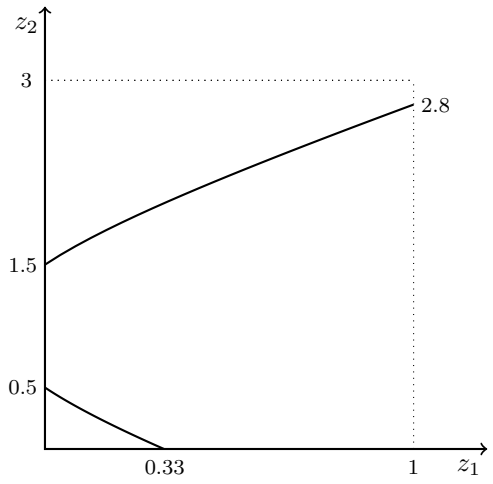

Figure 3: The $\mathcal{R}^{EG}$ likeness relation, as defined by the condition in (4).

The resulting likeness relation neither has full domain nor is functional. On the one hand, strategies $z_2 \in (0.5, 1.5) \cup (2.8, 3]$ do not have any corresponding alike strategy z_1. On the other hand, all strategies $z_1 \in [0, 0.33]$ have two corresponding alike strategies z_2. In particular, the strategy $z_1 = 0.1875$ has two alike strategies for player 2, namely $z_2 = 1.8125$ and $z_2 = 0.1875$. Hence, the efficient profile $z_1 = 0.1875, z_2 = 1.8125$ is implemented under $\mathcal{R}^{EG}$ only if, for player 1, the likeness relation selects $\sigma(0.1875) = 1.8125$ as the corresponding alike strategy. But the simple argument for taking this selection is that all players' payoffs are higher than under the alternative selection. Indeed, for all $z_1 \in [0, 0.33]$, the selection given by the bigger solution z_2 in (5) yields a higher payoff to both players than the selection given by the smaller

solution. Under such a selection, player 2's alike contributions are given by the function

$$z_2 = 1 + \frac{1}{2}\sqrt{1 + 8z_1 + 4z_1^2} \quad \text{for} \quad z_1 \in [0, 1]$$

while, for player 2, player 1's alike contributions are given by the function

$$z_1 = -1 + \frac{1}{2}\sqrt{7 - 8z_2 + 4z_2^2} \quad \text{for} \quad z_2 \in [0, 0.5] \cup [1.5, 2.8]$$

The associated optimization problems for the two universalizing agents are

$$\begin{aligned} &\max_{z_1 \in [0,1]} && \left(z_1 + 1 + \frac{1}{2}\sqrt{1 + 8z_1 + 4z_1^2}\right)\left(1 - z_1\right) \\ &\max_{z_2 \in [0,0.5]\cup[1.5,2.8]} && \left(z_2 - 1 + \frac{1}{2}\sqrt{7 - 8z_2 + 4z_2^2}\right)\left(3 - z_2\right) \end{aligned}$$

and their solutions are indeed given by the consistent profile $z_1 = 0.1875$, $z_2 = 1.8125$.

4 Conclusion

The principle of universalization is an important pillar of moral reasoning. Economists have explored behavior driven by universalization because of its potential to overcome social dilemmas and, I believe, because it is easily modeled as the solution to an optimization problem. Yet, while the meaning and implications of universalization are intuitive in symmetric games, the extension to asymmetric settings is more complicated. In this paper, I have interpreted the universalization principle through the lens of the associated bargaining game, by considering the behavior of different agents as alike according to the resulting division of any generated surplus. Under this view, and unlike other approaches in the literature, universalizing agents facing social dilemmas are able to coordinate on a cooperative outcome that guarantees mutual benefits with respect to the Nash equilibrium.

I conclude by arguing that the framework developed in the paper can also be relevant to the literature on reciprocity. Considerable evidence shows that for most people reciprocity motives are strong, and social norms are often perceived as conditional, meaning that individuals are willing to comply to the extent that others also comply (Ostrom, 1998; Fehr & Gächter, 2000; Malmendier et al., 2014). There is an analogy between universalization and reciprocity, in that both are grounded in a benchmark against which agents evaluate their actions. The difference

is stark in content: universalization concerns the hypothetical case in which others behave like the self, while reciprocity deals with the self's response to the observed or anticipated behavior of others. Yet, both motivations rely on some standard of comparison for agents' behavior. Clarifying how this standard is set in the presence of asymmetry is important both normatively, for understanding the formation of social and moral norms, as well as positively, for assessing the consequences when these norms are violated.

References

Agranov, M., Ali, S. N., Bernheim, B. D., & Palfrey, T. R. (2025). *Strategic complexity promotes egalitarianism in legislative bargaining* (Tech. Rep.). National Bureau of Economic Research.

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

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

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

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

Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, 891–904.

Benabou, R., Falk, A., & Henkel, L. (2024). *Ends versus means: Kantians, utilitarians, and moral decisions* (Tech. Rep.). National Bureau of Economic Research.

Bergstrom, T. (1995). On the evolution of altruistic ethical rules for siblings. *The American Economic Review*, 58–81.

Bergstrom, T., Blume, L., & Varian, H. (1986). On the private provision of public goods. *Journal of Public Economics*, *29*(1), 25–49.

Bilodeau, M., & Gravel, N. (2004). Voluntary provision of a public good and individual morality. *Journal of Public Economics*, *88*(3-4), 645–666.

Bordignon, M. (1990). Was kant right?: voluntary provision of public goods under the principle of unconditional commitment. *Economic Notes: Monte dei Paschi di Siena*(3), 342–372.

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

Buchanan, J. M. (1975). *The limits of liberty: Between anarchy and leviathan* (No. 714). University of Chicago press.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental economics*, *14*(1), 47–83.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796.

De Donder, P., Llavador, H., Penczynski, S. P., Roemer, J. E., & Vélez-Grajales, R. (2025). Nash versus kant: A game-theoretic analysis of childhood vaccination behavior. *Journal of Economics*, *145*(2), 97–128.

Eichner, T., & Pethig, R. (2020). Kant–nash tax competition. *International Tax and Public Finance*, *27*(5), 1108–1147.

Ellingsen, T., & Mohlin, E. (2025). A model of social duties. *Journal of Political Economy*.

Fehr, E., & Charness, G. (2025). Social preferences: fundamental characteristics and economic consequences. *Journal of Economic Literature*, *63*(2), 440–514.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human nature*, *13*(1), 1–25.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of economic perspectives*, *14*(3), 159–182.

Grafton, R. Q., Kompas, T., & Van Long, N. (2017). A brave new world? kantian–nashian interaction and the dynamics of global climate change mitigation. *European Economic Review*, *99*, 31–42.

Harsanyi, J. C. (1980). Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and decision*, *12*(2), 115–133.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American economic review*, *91*(2), 73–78.

Kaplow, L., & Shavell, S. (2007). Moral rules, the moral sentiments, and behavior: toward a theory of an optimal moral system. *Journal of Political Economy*, *115*(3), 494–514.

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

Ledyard, J. O. (1995). Public goods: A survey of experimental research. *The Handbook of Experimental Economics*, 111-194.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, *21*(2), 153–174.

Malmendier, U., te Velde, V. L., & Weber, R. A. (2014). Rethinking reciprocity. *Annual Review of Economics*, *6*(1), 849–874.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, american political science association, 1997. *American political science review*, *92*(1), 1–22.

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

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

Roemer, J. E. (2019). *How we cooperate: a theory of kantian optimization*. Yale University Press.

Roth, A. E. (1988). Laboratory experimentation in economics: A methodological overview. *The Economic Journal*, *98*(393), 974–1031.

Salonia, E. M. (2025). A foundation for universalisation in games.

Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & public affairs*, 317–344.

Shapley, L. S. (1969). Utility comparison and the theory of games. *La Décision: Aggregation et Dynamique des Ordres de Preference*, 251–263.

Sugden, R. (1984). Reciprocity: the supply of public goods through voluntary contributions. *The Economic Journal*, *94*(376), 772–787.

Van Leeuwen, B., & Alger, I. (2024). Estimating social preferences and kantian morality in strategic interactions. *Journal of Political Economy Microeconomics*, *2*(4), 665–706.

A Analysis of Example 1

1. Nash equilibrium:

The best-response of agent 1 is $z_1 = \max\{\frac{1-z_2}{2}, 0\}$; the best-response of agent 2 is $z_2 = \frac{3-z_1}{2}$. Hence, there is a unique Nash equilibrium given by $z_1 = 0$ and $z_2 = \frac{3}{2}$. The utility levels at the Nash equilibrium are $u_1 = \frac{3}{2}$ and $u_2 = \frac{9}{4}$.

2. Pareto-optimal allocations:

The Pareto-optimal allocations are the solutions of

$$\max \quad u_1(Z, x_1) \quad \text{subject to} \quad u_2(Z, x_2) = \bar{u} \quad \text{and the budget constraints}$$

$$\Rightarrow \quad \max_{z_1, z_2} \quad (z_1 + z_2)(1 - z_1) \quad \text{subject to} \quad (z_1 + z_2)(3 - z_2) = \bar{u}$$

$$\Rightarrow \quad \max_{z_2} \quad \frac{\bar{u}}{3 - z_2}\left(1 - \frac{\bar{u}}{3 - z_2} + z_2\right)$$

The first order condition for an interior solution yields $z_2 = \frac{6-\bar{u}}{2}$ and thus

$$z_1 + z_2 = \frac{\bar{u}}{3 - z_2} = 2$$

3. Same absolute contributions:

Suppose both players universalize their choice of an absolute contribution: they set $z_1 = z_2$ in their optimization problem. Then agent 1 solves

$$\max_{z_1} \quad 2z_1(1 - z_1)$$

and chooses $z_1 = \frac{1}{2}$. Agent 2 solves

$$\max_{z_2} \quad 2z_2(3 - z_2)$$

and chooses $z_2 = \frac{3}{2}$. The two contributions are not the same.

4. Same contributions as a fraction of income:

Suppose both players universalize their choice of a contribution as a fraction α of their income:

they both set $z_1 = \alpha$ and $z_2 = 3\alpha$. Agent 1 solves

$$\max_{\alpha} \quad 4\alpha(1-\alpha)$$

and chooses $\alpha = \frac{1}{2}$. Agent 2 solves

$$\max_{\alpha} \quad 4\alpha(3-3\alpha)$$

and also chooses $\alpha = \frac{1}{2}$. The utility levels at the allocation $z_1 = \frac{1}{2}$, $z_2 = \frac{3}{2}$ are $u_1 = 1$ and $u_2 = 3$; hence, agent 1 is worse off than at the Nash equilibrium.

5. Lindahl equilibrium:

The Lindahl equilibrium is given by an amount of public good Z and personalized prices p_1, p_2 for the public good such that $p_1 + p_2 = 1$ (the marginal cost of producing the public good) and agents demand the same level of public good, i.e.

$$\arg\max_{Z} \; Z \cdot x_1 \text{ subject to } p_1 Z + x_1 = 1 \quad = \quad \arg\max_{Z} \; Z \cdot x_2 \text{ subject to } p_2 Z + x_2 = 3$$

The solution is given by

$$p_1 = \frac{1}{4} \quad , \quad p_2 = \frac{3}{4} \quad , \quad Z = 2$$

Hence, the corresponding individual contributions are $z_1 = p_1 Z = \frac{1}{2}$ and $z_2 = p_2 Z = \frac{3}{2}$, i.e. they coincide with the previous allocation at which agent 1 is worse off than at the Nash equilibrium.

6. Implications from Bilodeau and Gravel (2004):

The public good game satisfies the assumptions of Bilodeau and Gravel's framework. Their result is that there exists an invertible function $z_2 = f(z_1)$ defining equivalent contributions for the two agents such that, under the universalization contraints, they would choose their corresponding Lindahl contributions. Denoting the Lindahl contributions by z_1^L and z_2^L, this means that the function f must be such that:

$$\begin{aligned} z_2^L &= f(z_1^L) \\ z_1^L &= \arg\max_{z_1} \quad (z_1 + f(z_1))(1 - z_1) \\ z_2^L &= \arg\max_{z_2} \quad (f^{-1}(z_2) + z_2)(3 - z_2) \end{aligned}$$

Bilodeau and Gravel do not provide a method to derive such a function in a general public good game. Yet, in this example, given the Lindahl contributions $z_1^L = \frac{1}{2}$, $z_2^L = \frac{3}{2}$, we can easily verify that one such function is

$$z_2 = 3z_1$$

7. Roemer's multiplicative Kantian equilibrium:

A multiplicative Kantian equilibrium is defined as a pair of contributions from which no agent wants both agents to deviate by any (positive) scalar factor. Denoting α the scalar factor, this corresponds to solving:

$$\begin{aligned} \arg\max_{\alpha} \quad & (\alpha z_1 + \alpha z_2)(1 - \alpha z_1) \quad = \quad 1 \\ \arg\max_{\alpha} \quad & (\alpha z_1 + \alpha z_2)(3 - \alpha z_2) \quad = \quad 1 \end{aligned}$$

Namely, if from the allocation (z_1, z_2) agents were to optimize by choosing a scalar deviation α, they would set $\alpha = 1$, i.e. they would not deviate. Solving the two first order conditions with respect to α and then imposing $\alpha = 1$ yields the following system of two equations in the two unknowns z_1, z_2:

$$\begin{aligned} (z_1 + z_2)(1 - 2z_1) \quad &= \quad 0 \\ (z_1 + z_2)(3 - 2z_2) \quad &= \quad 0 \end{aligned}$$

The only strictly positive multiplicative Kantian equilibrium is the allocation $z_1 = \frac{1}{2}$, $z_2 = \frac{3}{2}$, which coincides with the Lindahl equilibrium (by construction, the zero vector $z_1 = z_2 = 0$ is also always a multiplicative Kantian equilibrium).

8. Roemer's additive Kantian equilibrium:

An additive Kantian equilibrium is defined as a pair of contributions from which no agent wants both agents to deviate by any additive factor. Denoting β the additive factor, this corresponds to solving:

$$\begin{aligned} \arg\max_{\beta} \quad & (z_1 + \beta + z_2 + \beta)(1 - z_1 - \beta) \quad = \quad 0 \\ \arg\max_{\beta} \quad & (z_1 + \beta + z_2 + \beta)(3 - z_2 - \beta) \quad = \quad 0 \end{aligned}$$

Namely, if from the allocation (z_1, z_2) agents were to optimize by choosing an additive deviation β, they would set $\beta = 0$, i.e. they would not deviate. Solving the two first order conditions with respect to β and then imposing $\beta = 0$ yields the following system of two equations in the

two unknowns z_1, z_2:

$$\begin{aligned} 2(1-z_1)-(z_1+z_2) &= 0 \\ 2(3-z_2)-(z_1+z_2) &= 0 \end{aligned}$$

which has a unique solution $z_1 = 0$, $z_2 = 2$. This is not an interior solution, since $z_1 = 0$. Hence, to be consistent with the logic of Kantian optimization, we need to check that agent 2 does not want to reduce her contribution by an additive factor, assuming that agent 1 would stick with a zero contribution. Clearly, agent 2 does want to do that, since her best-response to $z_1 = 0$ is $z_2 = \frac{3}{2}$. Hence, the previous profile is not an additive Kantian equilibrium and none exists.

B Additional examples (to be erased after peer-review)

Example 5. *Consider the following* 2×2 *game, with player* 1 *choosing the row and player* 2 *the column (the payoffs are omitted when irrelevant):*

	a_2	b_2
a_1	*2,2*	$3,1$
b_1	*1,3*	$\cdot$

Consider the (non-functional) likeness relation $\mathcal{R} = \{(a_1, a_2), (a_1, b_2), (b_1, a_2)\}$. *Suppose that selections are:*

- *for player* 1*:* $\sigma(a_1) = b_2$ *and* $\sigma(b_1) = a_2$
- *for player* 2*:* $\sigma(a_2) = b_1$ *and* $\sigma(b_2) = a_1$

Then, solving their optimization problems, player 1 *plays* a_1 *and player* 2 *plays* a_2. *Strategies* (a_1, a_2) *are alike according to* $\mathcal{R}$ *but, given the selections, they do not satisfy the second part of Definition 5.*

Example 6. *Consider the following* 2×2 *game, with player* 1 *choosing the row and player* 2 *the column (the payoffs are omitted when irrelevant):*

	a_2	b_2
a_1	*3,1*	$\cdot$
b_1	$\cdot$	*1,3*

Given the tight likeness relation $\mathcal{R} = \{(a_1, a_2), (b_1, b_2)\}$, *player* 1 *plays* a_1 *and player* 2 *plays* b_2, *but strategies* (a_1, b_2) *are not alike according to* $\mathcal{R}$.