WORKING PAPERS

N° 17-69

Strategic Behavior of Moralists and Altruists

Ingela Alger and Jörgen W. Weibull

IAST
INSTITUTE for
ADVANCED
STUDY in
TOULOUSE

Strategic Behavior of Moralists and Altruists

INGELA ALGER*AND JÖRGEN W. WEIBULL†

August 3, 2017‡

ABSTRACT. In order to shed light on the complex and non-trivial effects of altruism and morality on equilibrium behavior and outcomes, we examine a few canonical strategic interactions between egoists, altruists and moralists. Altruists care not only about own material gains and losses but also about the material gains and losses of others, and a moralist cares about own material gains and losses and also to what would happen if others were to act like himself. Both altruism and morality may improve or worsen the equilibrium outcomes. In infinitely repeated interactions both altruism and morality may diminish the prospects of sustaining cooperation, and morality more so than altruism. In coordination games, however, morality can eliminate inefficient equilibria while altruism cannot.

Keywords: altruism, morality, *Homo moralis,* repeated games, coordination games

JEL codes: C73, D01, D03.

1. INTRODUCTION

Few humans are motivated solely by their private gains. Most have more complex motivations, usually including some moral considerations, a concern for fairness or an element of altruism or even spite or envy towards others. There can even be a concern for the well-being of one's peer group, community, country or even humankind. By contrast, for a long time almost all of economics was based on the premise of narrow self-interest, by and large following the lead of Adam Smith's *Inquiry into the Nature and Causes of the Wealth of Nations* (1776). But also Adam Smith himself thought humans in fact have more complex and often social concerns and motives, a theme developed in his *Theory of Moral Sentiments* (1759).[1] Philosophers still argue how

*Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, and Institute for Advanced Study inToulouse. ingela.alger@tse-fr.eu

†Stockholm School of Economics, Institute for Advanced Study in Toulouse, KTH Royal Institute of Technology, and Toulouse School of Economics

‡Support by Knut and Alice Wallenberg Research Foundation and by ANR-Labex IAST is gratefully acknowledged. We also thank Agence Nationale de la Recherche for funding (Chaire d'Excellence ANR-12-CHEX-0012-01 for I. Alger, and Chaire IDEX ANR-11-IDEX-0002-02 for J. Weibull).

[1] Edgeworth (1881) also included such concerns in his original model formulation (see Collard, 1975).

to reconcile the themes of these two books in the mind of one and the same author. Did Adam Smith change his mind between the first and second book? Or was his position in his second book to demonstrate that well-functioning markets would result in beneficial results for society at large even if all individuals were to act only upon their own narrow self interest?

In view of the overwhelming experimental evidence that only a minority of people behave in accordance with predictions based on pure material self interest, it appears relevant to ask whether and how alternative preferences affect outcomes in standard economic interactions. It is commonly believed that if an element of altruism or morality were added to economic agents' self-interest, then outcomes would improve for all. Presumably, people would not cheat when trading with each other, they would work hard even when not monitored or remunerated by way of bonus schemes. They would contribute to public goods, respect and defend the interests of others, and might even be willing to risk their lives to save the lives of others. While this has certainly proved to be right in some interactions,[2] this belief is not generally valid. For example, Lindbeck and Weibull (1988) demonstrate that altruism can diminish welfare among strategically interacting individuals engaged in intertemporal decision-making. The reason is that if interacting individuals are aware of each others' altruism, then even altruists will, to some extent, exploit each others' altruism, resulting in misallocation of resources. One prime example is under-saving for one's old age, in the rational expectation that others will help if need be. In this example everyone would benefit from commitment not to help each other; as this could induce intertemporally optimal saving. Likewise, Bernheim and Stark (1988) show that altruism may be harmful to long-run cooperation. There, the reason is that in repeated games between altruists, punishments from defection may be less harsh if the punisher is altruistic — just like a loving parent who cannot credibly threaten misbehavior by a child with even a mild punishment. Specifically, in repeated interactions the mere repetition of a static Nash equilibrium in the stage game has better welfare properties between altruists than between purely self-interested individuals, thus diminishing the punishment from defecting from cooperation. However, altruism also diminishes the temptation to defect in the first place, since defecting harms the other party. Bernheim and Stark (1988) show that the net effect of altruism may be to diminish the potential for cooperation in the sense that it diminishes the range of discount factors that enables cooperation as a subgame-perfect equilibrium outcome.

[2] Thus, Becker (1976) shows that an altruistic family head is beneficial for the rest of the family, even if other family members are selfish (see also Bergstrom, 1989). More recently, Bourlès, Bramoullé, and Perez-Richet (2017) show that altruism is beneficial for income sharing in networks. Regarding morality, Laffont (1975) shows how an economy with Kantian individuals achieves efficiency. More recently, Brekke, Kverndokk, and Nyborg (2003) show that a certain kind of moral concerns enhances efficiency in the private provision of public goods.

The aim of the present study is to examine strategic interactions between altruists, as well as between moralists, more closely, in order to shed light on the complex and non-trivial effects of altruism and morality on equilibrium behavior and the associated material welfare. By 'altruism' we here mean that an individual cares not only about own material welfare but also about the material welfare of others, in line with Becker (1974,1976), Andreoni (1988), Bernheim and Stark (1988), and Lindbeck and Weibull (1988). As for 'morality' we rely on recent results in the literature on preference evolution, which shows that among all continuous preferences, a certain class, called *Homo moralis* preferences, stands out as being particularly favored by natural selection (Alger and Weibull, 2013, 2016). A holder of such preferences maximizes a weighted sum of own material welfare, evaluated at the true strategy profile, and own material welfare, evaluated at hypothetical strategy profiles in which some or all of the other player's strategies have been replaced by the individual's own strategy.[3]

We examine the effects of altruism and such morality for behavior and outcomes in repeated interactions, as well as in static interactions with complementarities – coordination games. Some of the results may, at first sight, be counter-intuitive and surprising. For instance, we find that morality may be even worse than altruism when it comes to sustaining cooperation in infinitely repeated interactions. We also show similarities and differences between altruism and morality, the main difference between these two motivations being due to the fact that while the first is purely consequentialistic - that is only concerned with the resulting material allocations - the second is partly deontological - that is placing some weights directly on "duty" or the moral value of acts, to care about what is "the right thing to do" in the situation at hand.

Our study complements the literature that analyzes the effects of pro-social inclinations on the qualitative nature of equilibrium outcomes in a variety of economic interactions; see, e.g., Arrow (1973), Andreoni (1988, 1990), Bernheim (1994), Levine (1998), Akerlof and Kranton (2000), Fehr and Schmidt (1999), Bénabou and Tirole (2006), Alger and Renault (2007), Englmaier and Wambach (2010), and Dufwenberg et al. (2011). The analysis of coordination interactions is further related to the literature on norms; see, e.g., Young (1993), Kandori, Mailath, and Rob (1993), Sethi and Somanathan (1996), Bicchieri (1997), Lindbeck, Nyberg, and Weibull (1999), Huck, Kübler, and Weibull (2012), and Myerson and Weibull (2015).

In the next section we define the three classes of preferences that we study, and review some known results. We then turn to studying repeated interactions (Sections 3 and 4), and coordination games (Section 5), and finally conclude.

[3] This is certainly not the only way morality can be modeled. See Bergstrom (2009) for mathematical representations of several well-known moral maxims for pairwise interactions. See also Gauthier (1986), Binmore (1994), Bacharach (1999), Sugden (2003), and Roemer (2006).

2. Defining self-interest, altruism, and morality

We consider n-player normal-form games (for any $n \geq 1$) in which each player has the same set X of (pure or mixed) strategies, and $\pi\left(x, \boldsymbol{y}\right) \in \mathbb{R}$ is the *material payoff* to strategy $x \in X$ when used against strategy profile $\boldsymbol{y} \in X^{n-1}$ for the other players. By 'material payoff' we mean the tangible consequences of playing the game, defined in terms of the individual's monetary gains (or losses), or, more generally, his or her indirect consumption utility from these gains (or losses). We assume π to be *aggregative* in the sense that $\pi\left(x, \boldsymbol{y}\right)$ is invariant under permutation of the components of $\boldsymbol{y}$. The strategy set X is taken to be a non-empty, compact and convex set in some normed vector space.

We say that an individual is purely self-interested, or a *Homo oeconomicus* if he only cares about his own material payoff, so that his utility is

$$u\left(x_i, \boldsymbol{x}_{-i}\right) = \pi\left(x_i, \boldsymbol{x}_{-i}\right) \quad \forall\left(x_i, \boldsymbol{x}_{-i}\right) \in X^n.$$

An individual is an *altruist* if he cares about his own material payoff and also attaches a weight, his or her *degree of altruism* $\alpha \in [0,1]$, to the material payoffs to others, so that his utility is:

$$v\left(x_i, \boldsymbol{x}_{-i}\right) = \pi\left(x_i, \boldsymbol{x}_{-i}\right) + \alpha \cdot \sum_{j \neq i} \pi\left(x_j, \boldsymbol{x}_{-j}\right) \quad \forall\left(x_i, \boldsymbol{x}_{-i}\right) \in X^n. \tag{1}$$

Finally, an individual is a *Homo moralis* if he cares about his own material payoff and also attaches a weight to what his material payoff would be should others use the same strategy as him. Formally, the utility to a *Homo moralis* with degree of morality $\kappa \in [0,1]$ is

$$w\left(x_i, \boldsymbol{x}_{-i}\right) = \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa^m \left(1-\kappa\right)^{n-m-1} \pi\left(x_i, \tilde{\boldsymbol{x}}_{-i}^m\right), \tag{2}$$

where $\tilde{\boldsymbol{x}}_{-i}^m$ is a random $(n-1)$-vector such that with probability κ^m exactly $m \in \{0, ..., n-1\}$ of the $n-1$ components of $\boldsymbol{x}_{-i}$ are replaced by x_i, with equal probability for each subset of size m, while the remaining components of $\boldsymbol{x}_{-i}$ keep their original values. We observe that a *Homo oeconomicus* can be viewed as an altruist with degree of altruism $\alpha = 0$, and as a *Homo moralis* with degree of morality $\kappa = 0$.

Our purpose is to compare equilibria of interactions in which all individuals are altruists with interactions in which all individuals are moralists. We are interested both in the equilibrium behaviors as well as in welfare properties of these equilibria. We will use G^α to refer to the n-player game between altruists with common degree of altruism α, with payoff functions defined in (1), and Γ^κ to refer to the n-player game between *Homo moralis* with common degree of morality κ, with payoff functions defined in (2).

A few such comparisons already exist in the literature. First, consider interactions in which $X = \mathbb{R}$ and π is continuously differentiable. Then any symmetric Nash equilibrium strategy x^* in game G^α, for any $0 \leq \alpha < 1$, satisfies the first-order condition:

$$\left.\frac{\partial \pi\left(x_i, \boldsymbol{x}_{-i}\right)}{\partial x_i}\right|_{x_1=\ldots=x_n=x^*} + (n-1)\,\alpha \cdot \left.\frac{\partial \pi\left(x_i, \boldsymbol{x}_{-i}\right)}{\partial x_n}\right|_{x_1=\ldots=x_n=x^*} = 0 \qquad (3)$$

(by permutation invariance of π, all partial derivatives, with respect to any x_j with $j \neq i$, are identical). But (3) is also necessary for x^* to be a symmetric equilibrium strategy in the same interaction between moralists, Γ^κ for $\kappa = \alpha$ (Alger and Weibull, 2016).[4]

2.1. Public goods.

A simple public goods game will illustrate. Suppose that

$$\pi\left(x_i, \boldsymbol{x}_{-i}\right) = \left(x_i + \sum\nolimits_{j \neq i} x_j\right)^{1/2} - x_i^2.$$

The unique symmetric Nash-equilibrium contribution in the associated game between moralists, Γ^κ, coincides with that in the associated game between altruists, G^α, when the degree of altruism is the same as the degree of morality, $\alpha = \kappa$. Hence, the behavioral effects of morality and altruism are here indistinguishable. Figure 1 below shows the equilibrium contribution as a function of community size n, for different degrees of morality, with higher curves for higher degrees of morality.

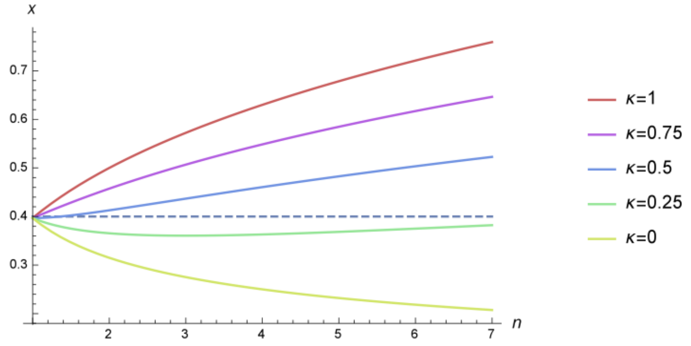

Figure 1: The unique Nash equilibrium contribution in the public-goods game for different degrees of morality.

[4] Second-order conditions may differ, however, so that the set of symmetric equilibria do not necessarily coincide. However, in the present public good example they do coincide. See also Bergstrom (1995) for an example for $\kappa = 1/2$ and $n = 2$.

We note the positive (negative) monotonicity with for high (low) degrees of morality, and the non-monotonicity for intermediate degrees (κ near 0.25).

2.2. Two-by-two games. Second, we briefly consider symmetric two-by-two games, with a_{ij} denoting the material payoff accruing to a player using pure strategy $i = 1, 2$ against pure strategy $j = 1, 2$. For mixed strategies, let $x, y \in [0, 1]$ denote the players' probabilities for using pure strategy 1. The expected material payoff from using mixed strategy x against mixed strategy y is bilinear:

$$\pi(x, y) = a_{11}xy + a_{12}x(1-y) + a_{21}(1-x)y + a_{22}(1-x)(1-y).$$

In such an interaction, an altruist's utility function is still bilinear:

$$\begin{aligned} v(x, y) &= a_{11}xy + a_{12}x(1-y) + a_{21}(1-x)y + a_{22}(1-x)(1-y) \qquad (4) \\ &\quad +\alpha \cdot [a_{11}xy + a_{12}y(1-x) + a_{21}(1-y)x + a_{22}(1-x)(1-y)], \end{aligned}$$

$$a_{11} + \alpha \cdot [a_{11}] > a_{21} + \alpha \cdot a_{12} \qquad (5)$$

while a *Homo moralis* has a utility function with quadratic terms:

$$\begin{aligned} w(x, y) &= (1-\kappa) \cdot [a_{11}xy + a_{12}x(1-y) + a_{21}(1-x)y + a_{22}(1-x)(1-y)] \quad (6) \\ &\quad +\kappa \cdot \left[a_{11}x^2 + (a_{12} + a_{21})x(1-x) + a_{22}(1-x)^2\right]. \end{aligned}$$

Depending on whether the sum of the diagonal elements of the payoff matrix, $a_{11} + a_{22}$, exceeds, equals, or falls short of the sum of the off-diagonal elements, $a_{12} + a_{21}$, the utility of *Homo moralis* is either strictly convex, linear, or strictly concave in his own mixed strategy, x. Hence, the set of symmetric equilibria of Γ^κ typically differs from that of G^α even when $\alpha = \kappa$.

As an illustration, consider a prisoner's dilemma with the first pure strategy representing "cooperate", that is, payoffs $a_{21} > a_{11} > a_{22} > a_{12}$. Assume also that $a_{11} + a_{22} > a_{12} + a_{21}$. Using the standard notation $a_{21} = T$, $a_{11} = R$, $a_{22} = P$, and $a_{12} = S$, it is easy to verify that "cooperation", that is, the strategy pair $(1, 1)$, is a Nash equilibrium in Γ^κ if and only if $\kappa \geq (T-R)/(T-P)$ and it is a Nash equilibrium in G^κ if and only if $\alpha \geq (T-R)/(R-S)$.[5] In this example, both altruism and morality help sustain cooperation, since in a game between two *Homo oeconomicus*, cooperation is not an equilibrium. We also note that altruism promotes cooperation more than does morality, since $R + P > T + S$ implies $(T-R)/(T-P) > (T-R)/(R-S)$.

We next turn to exploring unchartered territories, by studying repeated interactions and coordination, respectively. In both cases, we find that morality helps sustain socially efficient outcomes to a larger extent than altruism, and that sometimes pure self-interest is the best promoter.

[5] For a complete characterization of the set of symmetric equilibria in two-by-two games between moralists, see Alger and Weibull (2013).

3. Repeated prisoners' dilemmas

Consider an infinitely repeated prisoner's dilemma with material stage-game payoffs as above (i.e., $T > R > P > S$ and $R + P > T + S$), and a common discount factor $\delta \in (0, 1)$. The stage-game utilities to a row player with degree of altruism α are given in (4):

	C	D
C	$(1+\alpha)R$	$S + \alpha T$
D	$T + \alpha S$	$(1+\alpha)P$

If played by two equally altruistic individuals, grim trigger—that is, cooperate until someone defects, otherwise defect forever—if used by both players, constitutes a subgame perfect equilibrium that sustains cooperation forever if

$$(1+\alpha)R \;\geq\; (1-\delta)\cdot(T+\alpha S) \;+\; \delta(1+\alpha)P \tag{7}$$

and

$$\alpha < \frac{T-R}{R-S}. \tag{8}$$

The first inequality makes one-shot deviations from cooperation unprofitable for an altruist of degree α, while the second inequality makes one-shot deviations from defection unprofitable for such a player. The first inequality follows from the observation that if both players always cooperate, each gets utility $(1+\alpha)R$ in every period. If one player defects, he gets the temptation utility $T + \alpha S$ once, and then the punishment payoff $(1+\alpha)P$ forever thereafter. The second inequality follows from the observation in the preceding section that (D,D) is a Nash equilibrium in the stage game if and only if (8) holds. If the inequality in (8) is reversed, then (C,C) is a Nash equilibrium in the stage game, so then perpetual play of (C,C) is trivially a subgame perfect equilibrium.

In sum, perpetual cooperation is a subgame perfect equilibrium outcome if (7) holds, or, equivalently, if $\delta \geq \delta^A$, where

$$\delta^A = \frac{T - R - \alpha(R-S)}{T - P - \alpha(P-S)}.$$

We note that condition (8) is necessary and sufficient for δ^A to be strictly positive; indeed, it is only when cooperation is not an equilibrium of the stage game that it may be a challenge to sustain cooperation in the repeated interaction.

For a pair of purely self-interested players, the corresponding threshold value for the discount factor is

$$\delta^S = \frac{T-R}{T-P}.$$

It is easy to verify that $\delta^{A} < \delta^{S}$ for all $\alpha \in (0,1]$, implying that the grim trigger strategy enables cooperation more easily for a pair of altruists than for a pair of self-interested players. In this sense, altruism enhances cooperation.

Turning now to moralists, the stage-game utilities to a row player with degree of morality κ are given in (6):

	C	D
C	R	$(1-\kappa)\,S+\kappa R$
D	$(1-\kappa)\,T+\kappa P$	P

Comparison with the corresponding matrix for altruists reveals that while an altruist who defects internalizes the pain inflicted on the opponent, and is thus sensitive to the value S, a moralist who defects internalizes the consequence of his action should both choose to defect simultaneously, and is thus sensitive to the value P. Following the same logic as above, in a game between two equally moral individuals, perpetual cooperation is a subgame perfect equilibrium outcome if $\delta \geq \delta^{K}$, where

$$\delta^{K} = \frac{T-R-\kappa\,(T-P)}{T-P-\kappa\,(T-P)},$$

and

$$\kappa < \frac{T-R}{T-P}. \tag{9}$$

It is easy to verify that $\delta^{K} < \delta^{S}$ for all $\alpha \in (0,1]$, implying that the grim trigger strategy enables cooperation more easily for a pair of moralists than for a pair of self-interested players.

We finally compare a pair of moralists to a pair of altruists. A pair of moral players with common degree of morality κ can sustain cooperation by way of the grim trigger strategy more easily than a pair of altruistic players with common degree of altruism $\alpha = \kappa$ in the sense that $\delta^{K} < \delta^{A}$, if and only if

$$\kappa \cdot (T-P) > P-S. \tag{10}$$

In sum, the grim trigger strategy can sustain cooperation in an infinitely repeated prisoner's dilemma more easily if the players are altruists or moralists, since both altruism and morality reduces the temptation to defect. Moreover, if $P-S$ is small enough, i.e., if the pain inflicted by defecting on a cooperating opponent is small enough compared to the pain induced if both defected, then a pair of *Homo moralis* with degree of morality κ can sustain cooperation more easily than a pair of altruists with degree of altruism $\alpha = \kappa$.

4. Repeated sharing

The conclusion that it is easier for altruists and moralists than for self-interested individuals to sustain cooperation in an infinitely repeated game is far from general, however. This fact was pointed out for altruism by Bernheim and Stark (1988, section II.B). We first recapitulate their model (for the case when their parameter k is 1). We then carry through the same analysis for *Homo moralis*, and finally compare the two. The stage-game is the same as used by Bernheim and Stark, and represents sharing of consumption goods.

4.1. Altruism. The stage game is a two-player simultaneous-move game in which each player's strategy set is $X = [0, 1-\mu]$ for some small $\mu > 0$. If player 1 chooses $x \in X$ and player 2 chooses $y \in X$, payoffs are

$$v_1(x,y) = [x(1-y)]^\gamma + \alpha_1 \cdot [(1-x)y]^\gamma$$

and

$$v_2(x,y) = [y(1-x)]^\gamma + \alpha_2 \cdot [(1-y)x]^\gamma,$$

where $0 < \gamma < 1/2$. A necessary FOC for an interior NE is thus

$$\left(\frac{1-y}{y}\right)^\gamma = \alpha_1 \cdot \left(\frac{1-x}{x}\right)^{\gamma-1},$$

and likewise for player 2. Bernheim's and Stark's consider the symmetric case when $\alpha_1 = \alpha_2 = \alpha$, in which case the above FOC is their equation (16).[6] They use this to identify the following unique symmetric Nash equilibrium: $x = y = x^N$:

$$x^N = \min\left\{\frac{1}{1+\alpha}, 1-\mu\right\}.$$

They compare this with the unique symmetric Pareto optimum, $x^C = 1/2$, the solution of

$$\max_{x\in X} \quad [x(1-x)]^\gamma + \alpha \cdot [(1-x)x]^\gamma.$$

The equilibrium utility is $v^N = (1+\alpha) \cdot \left[x^N\left(1-x^N\right)\right]^\gamma$ and the Pareto optimal utility is $v^C = (1+\alpha) \cdot 4^{-\gamma}$. Bernheim and Stark consider an infinitely repeated play

[6] Bernheim and Stark instead use the utility specification

$$v = (1-\beta) \cdot [x(1-y)]^\gamma + \beta \cdot [(1-x)y]^\gamma,$$

with $\beta \in [0, 1/2]$. Hence our behavioral predictions coincide with theirs if one substitutes α by $\beta/(1-\beta)$.

of this stage game, with discount factor $\delta \in (0,1)$. They note that perpetual play of "cooperation", $\left(x^C, x^C\right)$, is sustained in subgame perfect equilibrium by the threat of (perpetual) reversion to $\left(x^N, x^N\right)$ iff $\delta \geq \delta_A$, where

$$\delta_A = \frac{v^D - v^C}{v^D - v^N}, \tag{11}$$

where v^D is the maximal utility from a one-shot deviation from cooperation, that is,

$$v^D = \max_{x \in X} \quad \frac{1}{2^\gamma}\left[x^\gamma + \alpha \cdot (1-x)^\gamma\right].$$

We find that

$$x^D = \min\left\{\frac{\alpha^{1/(\gamma-1)}}{1+\alpha^{1/(\gamma-1)}}, 1-\mu\right\}$$

In particular, for $\alpha = 1$, $x^D = 1/2$ and $v^D = 2 \cdot 4^{-\gamma} = v^C$. Hence, full altruists do not benefit from deviation and hence cooperation is sustainable irrespective of δ. As we will see, a discontinuity will appear in this respect when $\alpha \to 1$.

Bernheim and Stark proceed by considering a numerical example, namely, when $\mu = 0.01$ and $\gamma = 1/4$, and find that the lowest discount factor δ then needed to sustain cooperation is strictly *increasing* with α. Hence, they have an example in which altruism makes cooperation harder. We proceed in parallel with them by setting $\mu = 0.01$, $\gamma = 1/4$ and/ $\alpha > 0.05$. Then $x^N = 1/\left(1+\alpha\right)$,

$$v^N = \alpha^\gamma \left(1+\alpha\right)^{1-2\gamma},$$

and

$$x^D = \frac{1}{1+\alpha^{1/(1-\gamma)}}$$

for all α above approximately 0.05. See diagram below, where the upper dashed line marks $x^D = 1 - \mu = 0.99$.

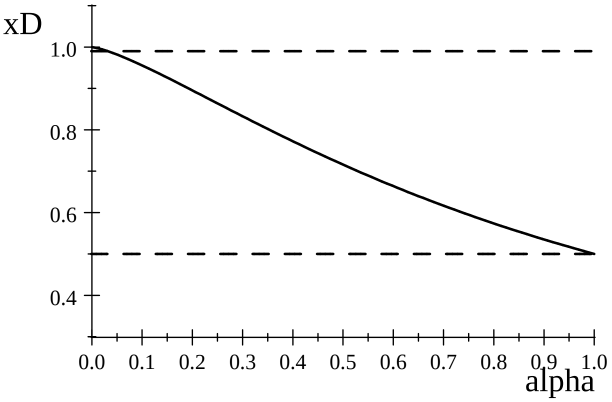

Figure 2: The optimal one-shot deviation for altruists in the repeated game.

For such α,

$$v^D = 2^{-\gamma}\alpha \cdot \left[1 + \alpha^{1/(\gamma-1)}\right]^{1-\gamma},$$

and

$$\delta_A = \frac{\left[1 + \alpha^{1/(1-\gamma)}\right]^{1-\gamma} - (1+\alpha)\cdot 2^{-\gamma}}{\left[1 + \alpha^{1/(1-\gamma)}\right]^{1-\gamma} - (1+\alpha)^{1-2\gamma}(2\alpha)^{\gamma}}.$$

The diagram below shows δ^A as a function of α when $\gamma = 1/4$, for $0.05 < \alpha < 1$. In particular, as $\alpha \to 1$, both the nominator and denominator in the definition tend to zero. By l'Hopital's rule, $\delta_A \to -\infty$ as $\alpha \to 1$.

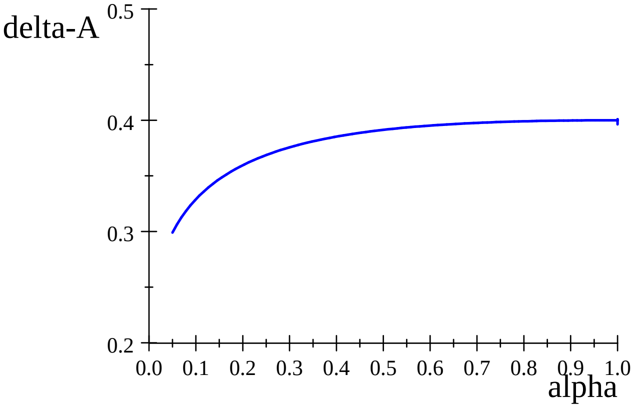

Figure 3: The critical discount factor for cooperation between altruists in the repeated game.

The above numerical results agree with those reported in Table 1 in Bernheim and Stark (1988), when keeping in mind that our altruism parameter α is a transformation of theirs (see footnote 5 above).

Comparing the above results with those for the special case of a pair of *Homo oeconomicus* ($\alpha = 0$), for whom cooperation can be sustained only if

$$\delta \geq \frac{(0.99/2)^{0.25} - 0.25^{0.25}}{(0.99/2)^{0.25} - 0.0099^{0.25}} \simeq 0.25,$$

we conclude that in this example altruism has an economically significant negative impact on the ability to sustain cooperation, since even a small degree of altruism, such as $\alpha = 1/9$, raises the discount factor needed for cooperation by 40%.

4.2. Stage-game morality. We begin by considering morality within the stage-game, just as Bernheim and Stark considered altruism within the stage-game. We then consider morality at the level of the repeated game. The stage game is again a two-player simultaneous-move game in which each player's strategy set is $X = [0, 1-\mu]$ for some small $\mu > 0$. If player 1 chooses $x \in X$ and player 2 chooses $y \in X$, payoffs are

$$w_1(x, y) = (1-\kappa_1) \cdot [x(1-y)]^\gamma + \kappa_1 \cdot [x(1-x)]^\gamma$$

and

$$w_2(x, y) = (1-\kappa_2) \cdot [y(1-x)]^\gamma + \kappa_2 \cdot [(1-y) y]^\gamma ,$$

where $0 < \gamma < 1/2$. A necessary FOC for an interior NE is thus

$$(1-\kappa_1) \cdot (1-y)^\gamma + \kappa_1 (1-x)^\gamma = \kappa_1 x^\gamma \cdot \left(\frac{1-x}{x}\right)^{\gamma-1} ,$$

and likewise for player 2. Suppose that $\kappa_1 = \kappa_2 = \kappa$, and consider first potential interior symmetric equilibria, $x = y = x^{NK} \in (0, 1-\mu)$, in which case the above FOC boils down to $x^{NK} = 1/(1+\kappa)$. More generally, the unique symmetric equilibrium strategy is

$$x^{NK} = \min\left\{\frac{1}{1+\kappa}, 1-\mu\right\}.$$

Comparing a pair of altruists with common degree of altruism α to a pair of moralists with common degree of morality $\kappa = \alpha$, we note that $x^N = x^{NK}$.

Henceforth, assume that the first term is the smallest, that is, $\kappa \geq \mu/(1+\mu)$. Then the Nash equilibrium utility is

$$w^N = \left[x^{NK}\left(1-x^{NK}\right)\right]^\gamma = \left[\frac{\kappa}{(1+\kappa)^2}\right]^\gamma$$

The unique symmetric Pareto optimum is still $x^C = 1/2$, and the Pareto optimal utility is $w^C = 4^{-\gamma}$.

Consider an infinitely repeated play of this stage game, with discount factor $\delta \in (0, 1)$. Perpetual "cooperation", play of $\left(x^C, x^C\right)$, is sustained in subgame perfect equilibrium by the threat of (perpetual) reversion to $\left(x^{NK}, x^{NK}\right)$ iff $\delta \geq \delta_K$, where

$$\delta_K = \frac{w^D - w^C}{w^D - w^N} \tag{12}$$

where w^D is the maximal utility from a one-shot deviation from cooperation, , that is,

$$w^D = \max_{x \in X} \quad (1-\kappa) \cdot (x/2)^\gamma + \kappa \cdot [(1-x) x]^\gamma .$$

We find that

$$x^{DK} = \min\{x^*, 1-\mu\} ,$$

where x^* is the unique solution to the fixed-point equation

$$x = \frac{1-\kappa + [2(1-x)]^{\gamma}\kappa}{1-\kappa+2[2(1-x)]^{\gamma}\kappa}$$

see diagram below, plotting the solution as a function of κ, for $\gamma = 1/4$ (and for $\kappa \geq 0.05$, since our aim is to compare with the solution under altruism for $\alpha \geq 0.05$).

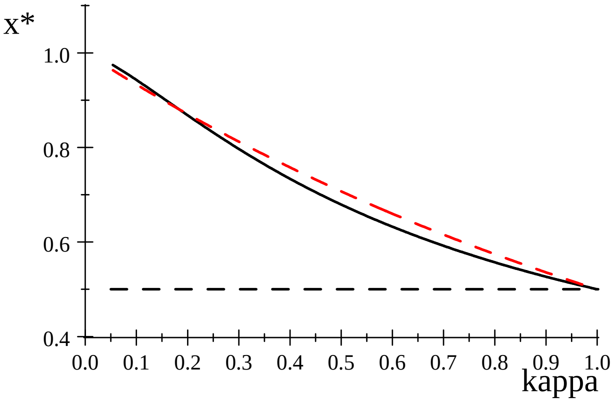

Figure 4: The optimal one-shot deviation for moralists in the repeated game.

Like Bernheim and Stark, we proceed by considering the numerical example when $\mu = 0.01$ and $\gamma = 1/4$, and we assume $\kappa > 0.01$, which guarantees an interior solution, both for x^K and x^{DK}. We use the approximation $x^{DK} = \exp(-\kappa \cdot \ln 2)$, indicated by the dashed curve in the diagram. This gives the approximation

$$\begin{aligned} w^D &= 2^{-\gamma} \cdot (1-\kappa) \cdot \exp(-\gamma\kappa \ln 2) + \kappa \cdot \left[(1-\exp(-\kappa \ln 2))^{\gamma} \exp(-\gamma\kappa \ln 2)\right] \\ &= \left[(1-\kappa) 2^{-\gamma} + \kappa \cdot (1-\exp(-\kappa \ln 2))^{\gamma}\right] \cdot \exp(-\gamma\kappa \ln 2). \end{aligned}$$

The condition (12) for sustainable cooperation can thus be written as

$$\delta \geq \frac{\left[(1-\kappa) 2^{-\gamma} + \kappa \cdot (1-\exp(-\kappa \ln 2))^{\gamma}\right] \cdot \exp(-\gamma\kappa \ln 2) - 4^{-\gamma}}{\left[(1-\kappa) 2^{-\gamma} + \kappa \cdot (1-\exp(-\kappa \ln 2))^{\gamma}\right] \cdot \exp(-\gamma\kappa \ln 2) - \kappa^{\gamma}(1+\kappa)^{-2\gamma}}$$

See diagram below, showing the right-hand side as a function of κ (for $\kappa \geq 0.05$) when $\gamma = 1/4$. The dashed curve is drawn for altruists with $\alpha = \kappa$. We see that, for $\gamma = 1/4$, cooperation is somewhat harder to sustain between moralists than between altruists with $\alpha = \kappa$. In sum, in this numerical example cooperation is easiest to maintain between purely self-interested individual than between altruists, and easier to sustain between altruists than moralists.

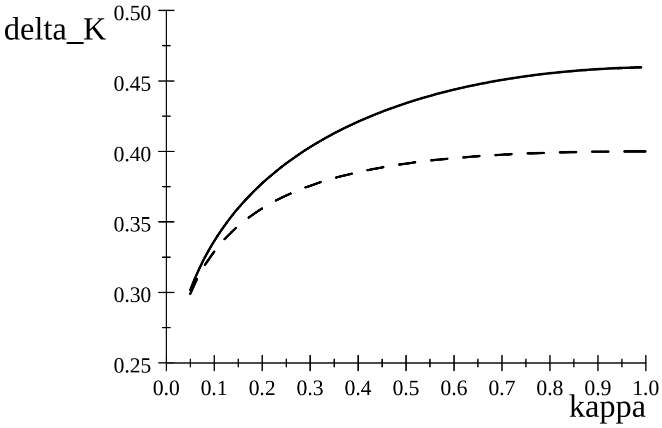

Figure 5: The critical discount factor for cooperation between moralists (solid) and altruists (dashed) in the repeated game.

Does this qualitative result partly depend on the numerical approximation? Does it hold for all γ? In order to investigate these issues, assume that $\kappa = \alpha$, and note that $\delta_A \leq \delta_K$ if and only if $(1+\alpha)\, w^D \geq v^D$, an inequality that can be written as

$$\max_{x \in X} \left[(1+\alpha) \cdot \left[(1-\alpha)^\gamma + \alpha \left[2\,(1-x) \right]^\gamma \right] x^\gamma - \alpha \cdot \left[1 + \alpha^{1/(\gamma-1)} \right]^{1-\gamma} \right] \geq 0 \qquad (13)$$

The left-hand side is positive at $\alpha = 0$, and by continuity this also holds for all $\alpha > 0$ that are small enough. For $\alpha = 1$, (13) holds with equality, since then the inequality boils down to

$$\max_{x \in X} \quad [(1-x)\, x]^\gamma \quad \geq \quad 4^{-\gamma},$$

which holds by equality. See diagram below, showing isoquants for the maximand in (13). The thick curve is the zero isoquant and the thin curves positive isoquants. The diagram suggests that for every $\alpha \in (0,1)$ there exists an $x \in int\,(X)$ such that (13) holds strictly. Hence, the difference between altruism and morality is not due to the approximation of x^{DK}.

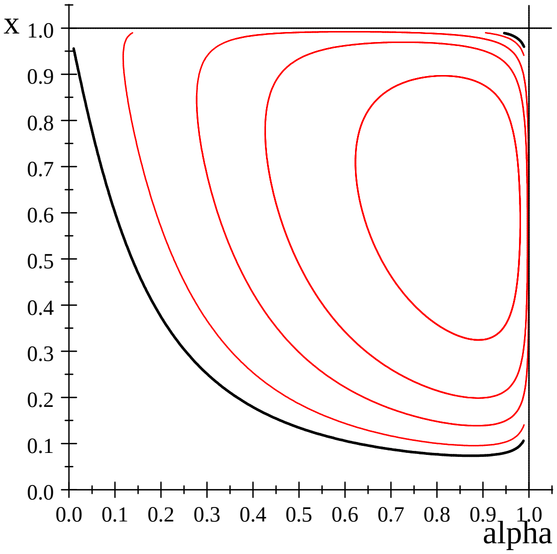

Figure 6: Contour map for the maximand in (13).

4.3. Repeated-game morality. The stage game is now the two-player simultaneous-move game in which each player's strategy set is $X = [0, 1 - \mu]$ for some small $\mu > 0$. If player 1 chooses $x \in X$ and player 2 chooses $y \in X$, material payoffs are

$$\pi_1 = [x(1-y)]^\gamma$$

and

$$\pi_2 = [y(1-x)]^\gamma$$

where $0 < \gamma < 1/2$. We consider the infinitely repeated game with discount factor $\delta \in (0, 1)$ and material payoff functions

$$\Pi_1(\sigma, \tau) = (1-\delta) \sum_{t=0}^{\infty} \delta^t \pi_1(x_t, y_t),$$

where σ and τ are 1's respectively 2's repeated-games strategy, and likewise for 2. Endowed with morality in line with *Homo moralis*, the players' utility functions are thus

$$W_1(\sigma, \tau) = (1-\kappa_1) \cdot \Pi_1(\sigma, \tau) + \kappa_1 \cdot \Pi_1(\sigma, \sigma),$$

and likewise for player 2.

Suppose that $\kappa_1 = \kappa_2 = \kappa$, and consider potential symmetric subgame-perfect equilibria, $\sigma = \tau = \sigma^K$, where σ^K prescribes play of $x_t = 1/2$ initially and after all histories h in which both parties have always taken this action. After any other history, σ^K prescribes play of

$$x^K = \min\left\{\frac{1}{1+\kappa}, 1-\mu\right\}.$$

Does the so defined strategy pair $\left(\sigma^K, \sigma^K\right)$ constitute a subgame-perfect equilibrium for any δ? We investigate this question under the assumption that

$$\kappa \geq \frac{\mu}{1+\mu}.$$

First we consider any history $h \in H^C$ of constant cooperation, then we consider histories $h \notin H^C$.

For histories $h \in H^C$, play according to σ^K results in utility

$$W_1\left(\sigma^K, \sigma^K\right) = \Pi_1\left(\sigma^K, \sigma^K\right) = (1-\delta)\sum_{t=0}^{\infty}\delta^t 4^{-\gamma} = 4^{-\gamma} = w^C.$$

After any history $h \notin H^C$, play according to σ^K results in utility

$$\begin{aligned} W_1\left(\sigma^K, \sigma^K\right) &= \Pi_1\left(\sigma^K, \sigma^K\right) \\ &= (1-\delta)\sum_{t=0}^{\infty}\delta^t\left[\frac{\kappa}{(1+\kappa)^2}\right]^{\gamma} = \kappa^{\gamma}(1+\kappa)^{-2\gamma} = w^N. \end{aligned}$$

It remains to see under what conditions, if any, a unilateral one-shot deviation is profitable, in each case. For histories in H^C, there exists no profitable one-shot deviation iff

$$\delta \geq \frac{w^{DR} - w^C}{w^{DR} - w^N}, \tag{14}$$

where

$$w^{DR} = \max_{\sigma} \quad (1-\kappa)\cdot\Pi_1\left(\sigma, \sigma^K\right) + \kappa\cdot\Pi_1(\sigma, \sigma),$$

and σ is a one-shot deviation from σ^K. Say that σ prescribes play of x in the deviation period. Then the maximand is

$$\begin{aligned} & (1-\kappa)\cdot\Pi_1\left(\sigma, \sigma^K\right) + \kappa\cdot\Pi_1(\sigma, \sigma) \\ = \; & (1-\kappa)\cdot\left[(1-\delta)\cdot(x/2)^{\gamma} + \delta\cdot w^N\right] + \kappa\cdot\left[(1-\delta)\cdot[x(1-x)]^{\gamma} + \delta\cdot w^N\right] \\ = \; & (1-\delta)(1-\kappa)\cdot(x/2)^{\gamma} + (1-\delta)\kappa\cdot[x(1-x)]^{\gamma} + \delta\cdot w^N \end{aligned}$$

Hence,

$$w^{DR} = \max_{x\in X} \quad (1-\kappa)\cdot(x/2)^{\gamma} + \kappa\cdot[(1-x)x]^{\gamma},$$

and thus $w^{DR} = w^D$. In sum, the condition on the critical value of the discount factor for sustainability of cooperation is identical with that when morality is defined in the stage game.

Remark 1. *The same qualitative conclusion should hold also for the altruism model, namely that it does not matter if one defines altruism in the stage game (as do Bernheim and Stark) or in the repeated game. In that setting this is not surprising since preferences are consequentialistic. It is not as evident for morality, since this is partly deontological.*

5. Coordination

Suppose there are n players who simultaneously choose between two actions, A and B. Write $s_i \in S = \{0,1\}$ for the choice of individual i, where $s_i = 1$ means that i chooses A, and $s_i = 0$ that instead B is chosen. The material payoff to an individual from choosing A is when n_A others choose action A is $n_A \cdot a$. Likewise, the individual's material payoff from choosing B when n_B others choose B is $n_B \cdot b$, where $0 < b < a$. Examples abound. We will think of A and B as two distinct "norms", with A being the socially efficient norm. We examine under which conditions the socially inefficient norm B can be sustained in equilibrium. We will also investigate if both norms can be simultaneously and partly sustained in heterogenous populations, in the sense that some individuals take action A while others take action B.

Writing $\boldsymbol{s}_{-i} \in S^{n-1}$ for the strategy profile of i's opponents and $u_i : S^n \to \mathbb{R}$ for the payoff function of a purely self-interested player $i = 1, ..., n$, we have

$$u_i\left(s_i, \boldsymbol{s}_{-i}\right) = as_i \cdot \sum_{j \neq i} s_j + b\left(1 - s_i\right) \cdot \sum_{j \neq i} \left(1 - s_j\right). \tag{15}$$

The utility function of an altruistic player i with degree of altruism $\alpha_i \in [0,1]$ is

$$v_i\left(s_i, \boldsymbol{s}_{-i}\right) = u_i\left(s_i, \boldsymbol{s}_{-i}\right) + \alpha_i \cdot \sum_{j \neq i} u_j\left(s_j, \boldsymbol{s}_{-j}\right). \tag{16}$$

Evidently the efficient norm A, that is all playing strategy 1, can always be sustained as a Nash equilibrium for arbitrarily altruistic players. But also the inefficient norm B is a Nash equilibrium. The reason is simple. If all others choose B, then so will any player i, no matter how altruistic. However, this last conclusion does not hold for moralists. Sufficiently moral persons will deviate to the efficient norm even if many or most others stick to the inefficient norm. This is the issue we will here examine.

Consider *Homo moralis* players, where player i has degree of morality $\kappa_i \in [0,1]$. Such a player's utility function is

$$w_i\left(s_i, \boldsymbol{s}_{-i}\right) = \mathbb{E}_{\kappa_i}\left[u_i\left(s_i, \tilde{\boldsymbol{s}}_{-i}^m\right)\right], \tag{17}$$

where $\tilde{\boldsymbol{s}}_{-i}^m$ is a random vector in S^{n-1} such that with probability $\left(\kappa_i\right)^m$ exactly $m \in \{0, ..., n-1\}$ of the $n-1$ components of $\boldsymbol{s}_{-i}$ are replaced by s_i, with equal probability for each subset of size m, while the remaining components of $\boldsymbol{s}_{-i}$ keep their original values. Thanks to the linearity of the material payoff function (15), the utility function w_i can be written as

$$\begin{aligned} w_i\left(s_i, \boldsymbol{s}_{-i}\right) &= \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_i^m \left(1-\kappa_i\right)^{n-m-1} \cdot \left[a s_i \cdot \left(m s_i + \frac{n-1-m}{n-1} \cdot \sum_{j \neq i} s_j \right) \right. \\ & \left. + b\left(1-s_i\right) \cdot \left(m \cdot \left(1-s_i\right) + \frac{n-1-m}{n-1} \cdot \sum_{j \neq i} \left(1-s_j\right) \right) \right]. \end{aligned}$$

The efficient norm A can clearly be sustained as a Nash equilibrium, since when all the others are playing A, individual i gets utility $(n-1)\,a$ from taking action A and

$$b \cdot \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_i^m \left(1-\kappa_i\right)^{n-m-1} m = b\,(n-1)\,\kappa_i$$

from taking action B. By contrast, the inefficient norm cannot be sustained for all degrees of morality. To see this, first suppose all individuals have the same degree of morality $\kappa \in (0,1)$. If all the others are playing B, any individual gets utility $(n-1)\,b$ from also playing B and would get utility

$$a \cdot \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa^m \left(1-\kappa\right)^{n-m-1} m = a\,(n-1)\,\kappa$$

from deviating to A. Hence, the inefficient norm can be sustained in Nash equilibrium if an only if $\kappa \leq b/a$.

This result shows that morality can have a qualitatively different effect than altruism upon behavior in interactions with strategic complementarities. In the present case of a simple coordination game, morality eliminates the inefficient equilibrium if and only if the common degree of morality κ exceeds b/a. By contrast, the inefficient equilibrium is still an equilibrium under any degree of altruism. No matter how much the parties care for each other, they always want to use the same strategy, even if this results in a socially inefficient outcome. Moralists, if sufficiently fervent, are partly deontologically motivated and evaluate own acts not only in terms of their expected consequences, given others' action, but also in terms of what ought to be done.

We now consider heterogeneous populations. Consider, first, a coordination game, as defined above, played by $n > (a+2b)/b$ individuals, among which all but one are purely self-interested and the remaining individual is a *Homo moralis* with degree of morality $\kappa > b/a$. Under complete information, such a game has a Nash equilibrium in which all the self-interested play B while the unique *Homo moralis* plays A. In this equilibrium, the moral player exerts a negative externality on the others — causes partial mis-coordination. Had the moralist instead been an altruist, he would also play B if the others do, and would thus be behaviorally indistinguishable from the purely self-interested individuals. More generally, altruists as well as self-interested individuals do not care about "the right thing to do" should others do likewise. They only care about the consequences for own and—if altruistic—others' material payoffs, from their unilateral choice of action. By contrast, moralists care also care about what would happen if, hypothetically, others would act like them. In coordination games, this may cause a bandwagon effect reminiscent of that shown in Granovetter's (1978) threshold model of collective action, a topic to which we now turn.

Like Granovetter, we analyze a population in which each individual faces a binary choice and takes a certain action, say A, if and only if sufficiently many do likewise. More precisely, each individual has a population threshold for taking action A. Our model of coordination can be recast in these terms. Indeed, for each individual $i = 1, 2, ..., n$, defined by his personal degree of morality $\kappa_i \in [0,1]$, one can readily determine the minimum number of other individuals who must take action A before he is willing to do so. Consider any player i's choice. If he expects $\tilde{n} \in \{0, ..., n-1\}$ others to take action B, then his utility from taking action B is

$$\begin{aligned} w_i\left(0, \boldsymbol{s}_{-i}\right) &= b \cdot \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_i^m \left(1-\kappa_i\right)^{n-m-1} \left[\frac{n-1-m}{n-1} \cdot (n-\tilde{n}-1) + m\right] \\ &= b \cdot \left[(n-\tilde{n}-1) + \tilde{n}\kappa_i\right] \end{aligned}$$

while from taking action A it is

$$\begin{aligned} w_i\left(1, \boldsymbol{s}_{-i}\right) &= a \cdot \sum_{m=0}^{n-1} \binom{n-1}{m} \kappa_i^m \left(1-\kappa_i\right)^{n-m-1} \left[\frac{n-1-m}{n-1} \cdot \tilde{n} + m\right] \\ &= a \cdot \left[\tilde{n} + (n-\tilde{n}-1)\,\kappa_i\right]. \end{aligned}$$

Hence, individual i will take action A if and only if

$$\frac{a}{b} \geq \frac{n-\tilde{n}-1+\tilde{n}\kappa_i}{\tilde{n}+(n-\tilde{n}-1)\,\kappa_i},$$

or

$$\frac{\tilde{n}}{n-1} \geq \frac{b-\kappa_i a}{\left(1-\kappa_i\right)(a+b)}.$$

In other words, individual i's *threshold* $\theta_i \in \mathbb{R}$ is

$$\theta_i = \frac{b-\kappa_i a}{\left(1-\kappa_i\right)(a+b)}.$$

Whenever individual i expects the population share $x = \tilde{n}/(n-1)$ of others taking action A to exceed (respectively, fall short of) his or her threshold θ_i, he/she takes action A (respectively B). We note that the threshold of an individual is strictly decreasing in the individual's degree of morality, and that individuals with high degrees of morality have negative thresholds. Hence, such individual will take action A even alone. The threshold of an individual with zero degree of morality, that is, *Homo oeconomicus*, is $b/(a+b)$. See diagram below, drawn for different values of $v = a/b$, and with population shares (in percentages) on the vertical axis.

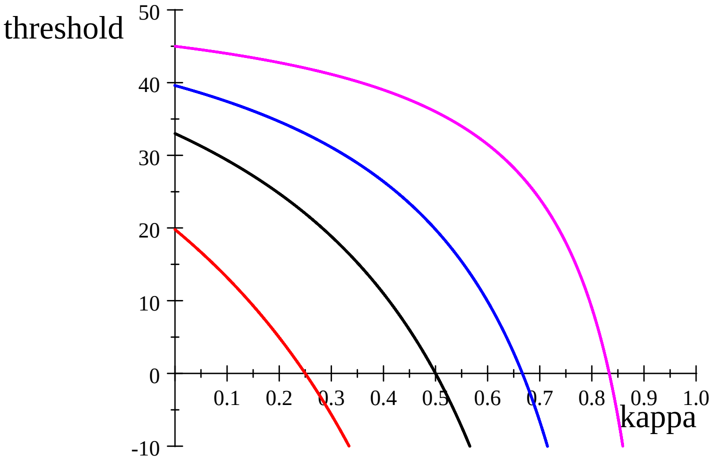

Figure 7: Thresholds for switching to A, as a function of the degree of morality, in a population of size $n = 100$.

Starting from the bottom, the curves are drawn for $v = 4$, $v = 2$, $v = 1.5$, and $v = 1.2$. The bottom curve, the one for $v = 4$, shows that an individual with degree of morality $\kappa = 0.25$ is willing to switch from B to A even if nobody else switches, an individual with degree of morality $\kappa = 0.1$ is willing to make this switch if 14% of the others also switch, etc. This curve also reveals that as long as there is at least 20% who are sufficiently moral, and thus willing to switch even if nobody else does, or only a small number have switched, then a bandwagon effect among myopic individuals will eventually lead the whole population to switch, step by step, even if as many as 80% of the individuals are driven by pure self interest.

Let F be any continuous cumulative distribution function on $\mathbb{R}$ such that for every $\theta_i \in \mathbb{R}$, $F(\theta_i)$ is the population share of individuals with thresholds not above θ_i. Then $F : \mathbb{R} \to [0,1]$ is a continuous representation of the cumulative threshold distribution in the population, with $F(0) \geq 0$ and $F(x) = 1$ for all $x \geq b/(a+b)$. By Bolzano's intermediate-value theorem, $F(x) = x$ for at least one $x \in X = [0,1]$.[7]

[7] To see this, let $\phi(x) = F(x) - x$ for all $x \in [0,1]$, and note that ϕ is continuous with $\phi(0) \geq 0$ and $\phi(1) \leq 0$.

Let $X^* \subseteq [0, 1]$ be the non-empty and compact set of such fixed points. See diagram below, where each of the three curves is the CDF of a potential threshold distribution.

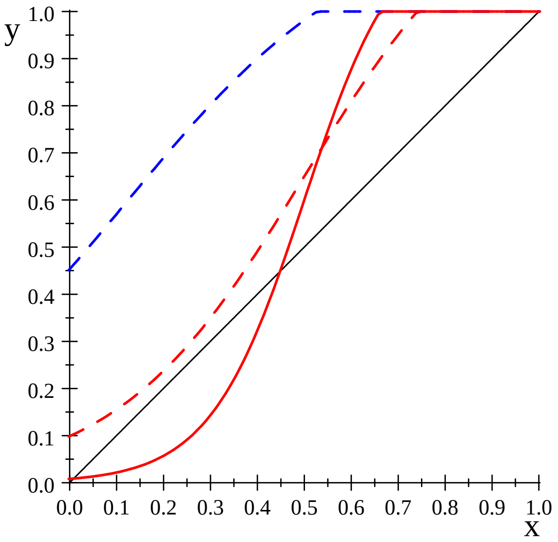

Figure 8: Fixed points for coordination in morally heterogeneous populations.

The two dashed curves represent relatively heterogenous populations, and those curves have one intersection with the diagonal, and hence the unique fixed point then is $x^* = 1$. The solid curve represents a relatively homogeneous population and this distribution function has three intersections with the diagonal, and thus three fixed points; one close to zero, another near 0.45, and the third one being $x^* = 1$. All fixed points are Nash equilibria in a continuum population, and are approximate Nash equilibria in finite but large populations.

In the diagram, all fixed points except the one near 0.45 have index +1. Those equilibria are stable in plausible population dynamics, while the fixed point near 0.45 has index -1 and is dynamically unstable.[8] We briefly consider a few dynamic scenarios. Consider first a relatively heterogeneous population with morality distribution such that there is only one fixed point, which then necessarily is $x^* = 1$. If initially all individuals were to take action B, then all those with non-positive thresholds θ (that is, with relatively high morality) would switch to A. If others see this, then the most moral among them (that is, those with lowest threshold) will follow suit. Depending on population size and its morality distribution, this process may go on until

[8] A fixed point has index +1 if the curve $y = F(x)$ intersects the diagonal, $y = x$, from above. In general, an index of +1 usually implies strong forms of dynamic stability, while an index of -1 usually implies instability, see McLennan (2016), and the references therein, for recent discussions and analyses of index theory in economics and game theory.

the population shares taking action A reaches or surpasses $b/(a+b)$, at which point all remaining individuals will switch to A. Next, consider a relatively homogenous population with smallest fixed point $x^* < 1$, again initially with all individuals taking action B. Again those with non-positive thresholds (if any), will switch to A, which may inspire others to also switch etc. This adjustment process may go on until the population share taking action A reaches or surpasses x^*, at which point the process will either halt or switch back and forth close to x^*. Hence, the population may get stuck there. Had it instead started somewhere above the middle fixed point, it could lead the population gradually towards norm A and finally jump to that norm.

A discrete-time version of this process is as follows. Consider a situation in which initially only strategy B exists, so that initially everybody plays B. Suddenly, strategy A appears, the interpretation being that it is discovered or invented. For each threshold number of individuals $\tilde{n} \in \{0, 1, 2, ...n-1\}$, let $g(\tilde{n})$ be the number of individuals who have that threshold. If $g(0) = 0$, then nobody ever switches to A. But if $g(0) > 0$, the number of individuals $N(t)$ who have switched from B to A at time $t = 1, 2, ...$, where t denotes the number of time periods after strategy A was discovered, we have $N(1) = g(0)$, and

$$N(t) = \sum_{j=0}^{N(t-1)} g(j).$$

for all $t > 1$. The process stops before everybody has switched if there exists some t such that $N(t+1) = N(t)$, i.e., if

$$\sum_{j=N(t-1)}^{N(t)} g(j) = 0.$$

Otherwise, it goes on until the whole population has switched to the efficient norm. In this process *Homo moralis* act as leaders, because they are willing to lead by example. By contrast, altruists as well as self-interested individuals do not care about the right thing to do, should others follow their lead. They care about own material payoff, as well as that of others for altruists, given what the others do. Hence, the cascading effect obtained with moral individuals does not obtain in groups of altruists or self-interested people. We illustrate with two examples, both in which $n = 100$. The following table shows the distribution of the thresholds. In the first example, a total of 21 individuals switch, and this takes four periods. In the second example, all individuals have switched after six periods, in spite of a slower start. Indeed, in the first example, we have $N(1) = 5$, $N(2) = 5 + 7 = 12$, $N(3) = 12 + 6 = 18$, $N(4) = 18 + 3 = 21$, but since the remaining individuals require at least 22 people to have switched before them, they do not switch. In the second example, the process starts with just one individual switching, $N(1) = 1$, but then $N(2) = 5$, $N(3) = 10$, $N(4) = 16$, $N(5) = 32$, $N(6) = 100$.

TABLE 1

$g(0)$	5
$g(4)$	7
$g(9)$	6
$g(14)$	3
$g(22)$	10
$g(23)$	11
$g(24)$	12
$g(25)$	13
$g(26)$	14
$g(27)$	19

$g(0)$	1
$g(1)$	4
$g(4)$	5
$g(8)$	6
$g(12)$	7
$g(16)$	9
$g(18)$	10
$g(20)$	11
$g(22)$	13
$g(23)$	15
$g(26)$	19

6. Concluding remarks

Altruism and morality are considered virtues in almost all societies and religions worldwide. We do not question this here. Instead, we ask whether altruism and morality help improve the material welfare properties of equilibria in strategic interactions. Our analysis reveals a complex picture; altruism is sometimes better than morality, sometimes the reverse is true, sometimes they are equivalent, and sometimes self-interest is best! Our conclusion is thus that economists cannot simply assume that altruism and morality lead to better outcomes. The incentives of altruists and moralists in each interaction need to be carefully analyzed. Nonetheless, our analyses unveil two interesting phenomena, that we believe to be robust and general.

First, in infinitely repeated interactions the Nash-reversion strategy may be less powerful with altruists and moralists than with self-interested individuals. While altruists and moralists are less tempted to deviate from the efficient strategy and less prone to punish each other—an altruist internalizes the pain inflicted on the opponent and a moralist internalizes what would happen if both were to deviate simultaneously—the stage-game Nash equilibrium is more efficient with altruists and moralists than with self-interested players, rendering the punishment following a deviation less painful. In the stage game we consider, the latter effect is always strong enough to outweigh the former, so that both altruism and morality worsen the prospects for long-run social efficiency. More extensive analyses are called for to investigate whether this result also obtains for other game specifications and for other punishment strategies.

Second, our analysis of coordination games unveils a fundamental difference between *Homo moralis* on the one hand, and self-interested and altruistic individuals on the other hand. Indeed, while *Homo moralis* preferences have the potential to eliminate socially inefficient equilibria, neither self interest nor altruism have. The reason is clear: while a *Homo moralis* is partly driven by the right thing to do (in terms of the material payoffs) if others were to follow his behavior, a self-interested or an altruistic individual is solely driven by what others actually do. We also show that individuals with *Homo moralis* preferences may trigger a group of people to switch from an inefficient to an efficient "norm" through a cascading effect, even if individuals are myopic and do not foresee that they may induce others to also switch. Such cascading effects do not arise in groups of self-interested or altruistic individuals.

Advances in behavioral economics provide economists with richer and more realistic views of human motivation. Sound policy recommendations need to be based on such more realistic views. Otherwise, the recommendations are bound to fail, and may even be counter-productive. Our results show how altruism and morality may affect the material welfare properties of equilibrium outcomes in a few, but arguably prototypical, strategic interactions. Our results suggest that these two types of pro-social preferences sometimes have similar, and sometimes sharply distinct effects on equilibrium outcomes. Arguably, much more theoretical as well as empirical work is needed for a fuller understanding to be reached.

7. References

Akerlof, G. and R. Kranton (2000): "Economics and Identity," *Quarterly Journal of Economics,* 115, 715-753.

Alger, I. and R. Renault (2007): "Screening Ethics when Honest Agents Care about Fairness," *International Economic Review*, 47, 59-85.

Alger, I., and J. Weibull (2013): "Homo Moralis – Preference Evolution under Incomplete Information and Assortativity," *Econometrica*, 81, 2269-2302.

Alger, I., and J. Weibull (2016): "Evolution and Kantian Morality," *Games and Economic Behavior*, 98, 56-67.

Andreoni, J. (1988): "Privately Provided Public Goods in a Large Economy: The Limits of Altruism," *Journal of Public Economics*, 35, 57-73.

Andreoni, J. (1990): "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100, 464-477.

Arrow, K. (1973): "Social Responsibility and Economic Efficiency," *Public Policy*, 21, 303-317.

Bacharach, M. (1999): "Interactive Team Reasoning: A Contribution to the Theory of Cooperation," *Research in Economics*, 53, 117-147.

Becker, G. (1974): "A Theory of Social Interaction", *Journal of Political Economy*, 82, 1063-1093.

Becker, G. (1976): "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology", *Journal of Economic Literature*, 14, 817-826.

Bénabou, R. and J. Tirole (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652-1678.

Bergstrom, T. (1995): "On the Evolution of Altruistic Ethical Rules for Siblings," *American Economic Review,* 85, 58-81.

Bergstrom, T. (1989): "A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries," *Journal of Political Economy,* 97, 1138-1159.

Bergstrom, T. (2009): "Ethics, Evolution, and Games among Neighbors," Working Paper UCSB.

Bernheim, B.D. (1994): "A Theory of Conformity," *Journal of Political Economy*, 102:841–877.

Bernheim, B.D. , and O. Stark (1988): "Altruism within the Family Reconsidered: Do Nice Guys Finish Last?" *American Economic Review,* 78, 1034-1045.

Bicchieri, C. (1997): *Rationality and Coordination.* Cambridge: Cambridge University Press.

Binmore, K. (1994): *Game Theory and The Social Contract, Volume 1: Playing Fair.* Cambridge USA: MIT Press.

Bourlès, R., Y. Bramoullé, and E. Perez-Richet (2017): "Altruism in Networks," *Econometrica*, 85, 675-689.

Brekke, K.A., S. Kverndokk, and K. Nyborg (2003): "An Economic Model of Moral Motivation," *Journal of Public Economics*, 87, 1967–1983.

Collard, D. (1975): "Edgeworth's Propositions on Altruism," *Economic Journal*, 85, 355-360.

Dufwenberg, M., P. Heidhues, G. Kirchsteiger, F. Riedel, and J. Sobel (2011): "Other-Regarding Preferences in General Equilibrium," *Review of Economic Studies*, 78, 613-639.

Edgeworth, F.Y. (1881): *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences.* London: Kegan Paul.

Englmaier, F., and A. Wambach (2010): "Optimal Incentive Contracts under Inequity Aversion," *Games and Economic Behavior*, 69, 312-328.

Fehr, E., and K. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817-868.

Gauthier, D. (1986): *Morals by Agreement.* Oxford: Oxford University Press.

Granovetter, M. (1978): "Threshold Model of Collective Behavior," *American Journal of Sociology,* 83, 1420-1443.

Huck, S., D. Kübler, and J.W. Weibull (2012): "Social Norms and Economic Incentives in Firms," *Journal of Economic Behavior & Organization*, 83, 173-185.

Kandori, M., G.T. Mailath, and R. Rob (1993): "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 61, 29-56.

Laffont, J.-J. (1975): "Macroeconomic Constraints, Economic Efficiency and Ethics: an Introduction to Kantian Economics," *Economica*, 42, 430-437.

Levine, D. (1998): "Modelling Altruism and Spite in Experiments," *Review of Economic Dynamics*, 1, 593-622.

Lindbeck, A., S. Nyberg, and J. Weibull (1999): "Social Norms and Economic Incentives in the Welfare State," *Quarterly Journal of Economics*, 114, 1-33.

Lindbeck, A., and J. Weibull (1988): "Altruism and Time Consistency - the Economics of Fait Accompli," *Journal of Political Economy*, 96, 1165-1182.

McLennan, A. (2016): "The index +1 principle", mimeo., University of Queensland.

Myerson, R. and J. Weibull (2015): "Tenable Strategy Blocks and Settled Equilibria", *Econometrica*, 83, 943-976.

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

Sethi, R., and E. Somanathan (1996): "The Evolution of Social Norms in Common Property Resource Use;" *American Economic Review,* 86, 766-788.

Smith, A. (1759): *The Theory of Moral Sentiments.* Reedited (1976), Oxford: Oxford University Press.

Smith, A. (1776): *An Inquiry into the Nature and Causes of the Wealth of Nations.* Reedited (1976), Oxford: Oxford University Press.

Sugden R (2003): "The Logic of Team Reasoning," *Philosophical Explorations*, 6:165–181

Young, P. (1993): "Conventions," *Econometrica*, 61, 57-84.