WORKING PAPERS

N° 17-833

August 2017

“Strategic Behavior of Moralists and Altruists”

Ingela Alger and Jörgen W. Weibull

Toulouse School of Economics

Strategic Behavior of Moralists and Altruists

Ingela Alger[*]and Jörgen W. Weibull[†]

August 27, 2017[‡]

Abstract. Does altruism and morality lead to socially better outcomes in strategic interactions than selfishness? We shed some light on this complex and non-trivial issue by examining a few canonical strategic interactions played by egoists, altruists and moralists. By altruists we mean people who do not only care about their own material payoffs but also about those to others, and by a moralist we mean someone who cares about own material payoff and also about what would be his or her material payoff if others were to act like himself or herself. It turns out that both altruism and morality may improve or worsen equilibrium outcomes, depending on the nature of the game. Not surprisingly, both altruism and morality improve the outcomes in standard public goods games. In infinitely repeated games, however, both altruism and morality may diminish the prospects of cooperation, and to different degrees. In coordination games, morality can eliminate socially 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*

[*] 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

[‡] We thank Ted Bergstrom and two anonymous referees for helpful comments, and Rémi Leménager for research assistance. 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).

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 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

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

[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.

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.

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, results which show that 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 payoff and own material payoff 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 static and repeated interactions. Some of the results may appear surprising and counter-intuitive. 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, the second is partly deontological. In other words, the first motivation is only concerned with resulting material allocations, the second places some weight on "duty" or the moral value of acts, a concern about what is "the right thing to do" in the situation at hand.

Our study complements other theoretical analyses of the effects of pro-social preferences and/or moral values on the qualitative nature of equilibrium outcomes in a variety of strategic interactions. In economics, see Arrow (1973), Becker (1974), Andreoni (1988, 1990), Bernheim (1994), Levine (1998), Fehr and Schmidt (1999), Akerlof and Kranton (2000), Bénabou and Tirole (2006), Alger and Renault (2007), Ellingsen and Johannesson (2008), Englmaier and Wambach (2010), Dufwenberg et al. (2011), and Sarkisian (2017). For related models of social norms, see Young (1993), Kandori, Mailath, and Rob (1993), Sethi and Somanathan (1996), Bicchieri (1997), Lindbeck, Nyberg, and Weibull (1999), Huck, Kübler, and Weibull (2012),

[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).

and Myerson and Weibull (2015).[4]

Our study also complements a large literature on theoretical analyses of the evolution of behaviors in populations. For recent contributions, see Lehmann and Rousset (2012), Van Cleve and Akçay (2013), Allen and Tarnita (2014), Ohtsuki (2014), Peña, Nöldeke, and Lehmann (2015), and Berger and Grüne (2016). For surveys of related work on agent-based simulation models, see Szabó and Borsos (2016) and Perc et al. (2017).

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 (Section 3), and coordination games (Section 4), and finally conclude.

2. Definitions and preliminaries

We consider n-player normal-form games (for any $n > 1$) in which each player has the same set X of (pure or mixed) strategies, and $\pi(x, \boldsymbol{y}) \in \mathbb{R}$ 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 $\pi(x, \boldsymbol{y})$ 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

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

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:

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

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

$$w(x_i, \boldsymbol{x}_{-i}) = \mathbb{E}\left[\pi\left(x_i, \tilde{\boldsymbol{x}}_{-i}^m\right)\right], \tag{2}$$

where $\tilde{\boldsymbol{x}}_{-i}^m$ is a random $(n-1)$-vector such that with probability $\kappa^m (1-\kappa)^{n-m-1}$ exactly $m \in \{0, ..., n-1\}$ of the $n-1$ components of $\boldsymbol{x}_{-i}$ are replaced by x_i, while

[4] For a recent comprehensive textbook treatment of behavioral economics, see Dhami (2016).

the remaining components of $\boldsymbol{x}_{-i}$ keep their original values (for each m, there are $\binom{n-1}{m}$ ways to replace m of the $n-1$ components of $\boldsymbol{x}_{-i}$). For instance, writing x_j and x_k for the strategies of i's two opponents when $n=3$:

$$\begin{aligned} w\left(x_i, x_j, x_k\right) &= (1-\kappa)^2 \pi\left(x_i, x_j, x_k\right)+\kappa(1-\kappa) \pi\left(x_i, x_i, x_k\right) \\ &\quad +\kappa(1-\kappa) \pi\left(x_i, x_j, x_i\right)+\kappa^2 \pi\left(x_i, x_i, x_i\right) . \end{aligned} \tag{3}$$

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 the material 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).

2.1. Necessary first-order conditions. Consider a simple public goods game, with

$$\pi\left(x_i, \boldsymbol{x}_{-i}\right)=\left(x_i+\sum\nolimits_{j \neq i} x_j\right)^{1/2}-x_i^2, \tag{4}$$

where $x_i \geq 0$ is i's contribution to the public good. Assume further that $X=\mathbb{R}$. It turns out that in this interaction equilibria in Γ^κ coincide with those in G^α when $\alpha=\kappa$.

More generally, for interactions in which the strategy set X is an interval and π is continuously differentiable, any interior symmetric Nash equilibrium strategy x^* in game G^α, for any $0 \leq \alpha<1$, satisfies the first-order condition

$$\left.\frac{\partial \pi\left(x_i, \boldsymbol{x}_{-i}\right)}{\partial x_i}\right|_{x_1=\ldots=x_n=x^*} + (n-1) \alpha \cdot \left.\frac{\partial \pi\left(x_i, \boldsymbol{x}_{-i}\right)}{\partial x_n}\right|_{x_1=\ldots=x_n=x^*} = 0. \tag{5}$$

(By permutation invariance of π, all partial derivatives with respect to other players' strategies are identical). Moreover, (5) is also necessary for an interior strategy x^* to be a symmetric Nash equilibrium strategy in the same interaction between moralists, Γ^κ for $\kappa=\alpha$ (Alger and Weibull, 2016). Higher-order conditions may differ, however, so that the set of symmetric equilibria do not necessarily coincide. Nevertheless, in the above public good example they do coincide.[5]

Figure 1 shows the unique symmetric Nash-equilibrium contribution in the public goods game between moralists, Γ^κ, as a function of community size n, for different

[5] See also Bergstrom (1995) for an example for $\kappa=1/2$ and $n=2$.

degrees of morality, with higher curves for higher degrees of morality. This is also the unique symmetric Nash-equilibrium contribution in the public goods 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.

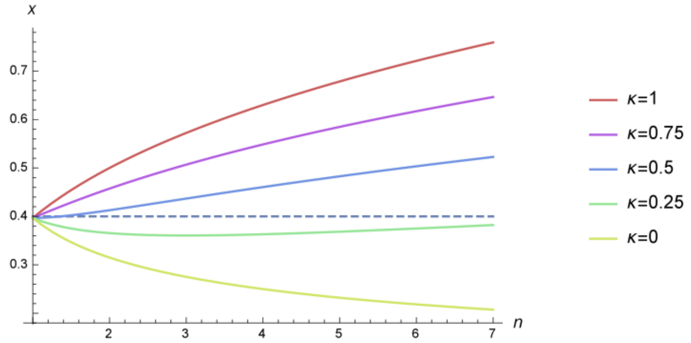

Figure 1: The unique Nash equilibrium contribution in the public-goods game for different degrees of morality.

2.2. Two-by-two games. We now briefly consider symmetric one-shot two-by-two games, with π_{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:

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

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

$$\begin{aligned} v(x,y) \quad &= \quad \pi_{11}xy + \pi_{12}x(1-y) + \pi_{21}(1-x)y + \pi_{22}(1-x)(1-y) \\ &\quad +\alpha \cdot [\pi_{11}xy + \pi_{12}y(1-x) + \pi_{21}(1-y)x + \pi_{22}(1-x)(1-y)], \end{aligned} \tag{6}$$

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

$$\begin{aligned} w(x,y) \quad &= \quad (1-\kappa) \cdot [\pi_{11}xy + \pi_{12}x(1-y) + \pi_{21}(1-x)y + \pi_{22}(1-x)(1-y)] \\ &\quad +\kappa \cdot \left[\pi_{11}x^2 + (\pi_{12} + \pi_{21})x(1-x) + \pi_{22}(1-x)^2\right]. \end{aligned} \tag{7}$$

Depending on whether the sum of the diagonal elements of the payoff matrix, $\pi_{11} + \pi_{22}$, exceeds, equals, or falls short of the sum of the off-diagonal elements, $\pi_{12} + \pi_{21}$, 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 $\alpha = \kappa$.[6]

As an illustration, consider a prisoner's dilemma with the first pure strategy representing "cooperate", that is, payoffs $\pi_{21} > \pi_{11} > \pi_{22} > \pi_{12}$. Using the standard notation $\pi_{11} = R$, $\pi_{12} = S$, $\pi_{21} = T$ and $\pi_{22} = P$, 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 $\kappa \geq \kappa^*$ where

$$\kappa^* = \frac{T-R}{T-P}, \tag{8}$$

and that it is a Nash equilibrium in G^κ if and only if $\alpha \geq \alpha^*$, where

$$\alpha^* = \frac{T-R}{R-S}. \tag{9}$$

We note that

$$\begin{cases} \alpha^* < \kappa^* & \text{if } R-S > T-P \\ \alpha^* = \kappa^* & \text{if } R-S = T-P \\ \alpha^* > \kappa^* & \text{if } R-S < T-P \end{cases}$$

In other words, it takes less altruism to turn cooperation into an equilibrium than it takes morality when the payoff loss $R-S$ inflicted upon an opponent by defecting—which an altruist cares about—exceeds the difference between the own payoff gain T from defecting unilaterally and from defecting together, P, a payoff difference a moralist cares about. The reverse is true when $R-S < T-P$.

We next turn to exploring unchartered territories, by studying repeated interactions (Section 3) and coordination (Section 4).

3. Repetition

We analyze infinite repetition of two distinct classes of interaction: prisoners' dilemmas and sharing games, respectively.

3.1. Repeated prisoners' dilemmas. Consider an infinitely repeated prisoner's dilemma with payoffs as above and with a common discount factor $\delta \in (0,1)$. We will provide necessary and sufficient conditions for grim trigger (that is, cooperate until someone defects, otherwise defect forever), if used by both players, to constitute a subgame-perfect equilibrium that sustains perpetual cooperation.[7] We do this first for a pair of equally altruistic players, then for a pair of equally moral players, and finally compare the ability to sustain cooperation of altruists with that of moralists.

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

[7] An analysis of more general repeated-games strategies falls outside the scope of this paper.

If played by two equally altruistic individuals with degree of altruism α, the stage-game utilities to the row player are the following (see (6)):

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

Grim trigger, if used by both players, constitutes a subgame perfect equilibrium that sustains perpetual cooperation if

$$(1+\alpha)\,R \;\geq\; (1-\delta)\cdot(T+\alpha S) \;+\; \delta\,(1+\alpha)\,P \tag{10}$$

and

$$\alpha \leq \frac{P-S}{T-P}. \tag{11}$$

The first inequality makes one-shot deviations from cooperation unprofitable. The left-hand side is the per-period payoff obtained if both players always cooperate. If one player defects, he gets the "temptation utility" $T+\alpha S$ once, and then the punishment payoff $(1+\alpha)\,P$ forever thereafter. Inequality (10) compares the present value of continued cooperation with the present value from a one-shot deviation. The second inequality, (11), makes a one-shot deviation from non-cooperation (play of (D,D)) unprofitable; this inequality is necessary for the threat to play D following defection to be credible. For further use below, we note that (10) can be written more succinctly as a condition on δ, or the players' patience, namely as, $\delta \geq \delta_A$, where

$$\delta_A = \frac{T-R-\alpha\,(R-S)}{T-P-\alpha\,(P-S)}. \tag{12}$$

Furthermore, denote by α^{**} the threshold value for α defined by (11).

In sum, a pair of equally altruistic players can sustain perpetual cooperation either if altruism is strong enough, $\alpha \geq \alpha^*$ (see (9)), in which case (C,C) is a Nash equilibrium of the stage game and hence needs no threat to be sustained, or if players are selfish enough to credibly punish defection, $\alpha \leq \min\{\alpha^*, \alpha^{**}\}$, and players are patient enough to prefer the long-term benefits from cooperation than the immediate reward from defection, $\delta \geq \delta_A$. In the intermediate case, that is, when $\alpha^{**} < \alpha < \alpha^*$, cooperation is not sustainable for any discount factor $\delta \in [0,1]$.

For example, suppose that $T=10$ and $S=0$. If $R=8$ and $P=4$, then $\alpha^*=1/4$ and $\alpha^{**}=2/3>\alpha^*$. In this case, cooperation is sustainable for any discount factor if $\alpha \geq 1/4$, and for any sufficiently high discount factor ($\delta \geq (1-4\alpha)\,/\,(3-2\alpha)$) if $\alpha < 1/4$. By contrast, if $R=6$ and $P=2$, $\alpha^*=2/3$ and $\alpha^{**}=1/4$. In this case, cooperation is sustainable for any discount factor if altruism is strong ($\alpha \geq 2/3$)

and for any sufficiently high discount factor ($\delta \geq (2-3\alpha)/(4-\alpha)$) if altruism is weak ($\alpha \leq 1/4$), but cooperation is not sustainable at all for intermediate degrees of altruism ($1/4 < \alpha < 2/3$).

Turning now to moralists, the stage-game utilities to a row player with degree of morality κ are given in (7), so we now have

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

Comparison with the utility 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, grim trigger sustains perpetual cooperation between two equally moral individuals as a subgame perfect equilibrium outcome if $\delta \geq \delta_K$, where

$$\delta_K = \frac{T-R-\kappa(T-P)}{T-P-\kappa(T-P)}, \tag{13}$$

and $\kappa \leq \kappa^{**}$, where

$$\kappa^{**} = \frac{P-S}{R-S}. \tag{14}$$

In sum, a pair of equally moral players can sustain perpetual cooperation either if $\kappa \geq \kappa^*$ (see (8)), in which case (C,C) is an equilibrium of the stage game and the threat to punish by playing D is not necessary to sustain cooperation in the repeated interaction, or if $\kappa \leq \min\{\kappa^*, \kappa^{**}\}$ and $\delta \geq \delta_K$.

We now turn to comparing a pair of selfish players to a pair of altruists or a pair of moralists. For selfish players, grim trigger constitutes a subgame perfect equilibrium that sustains perpetual cooperation if $\delta \geq \delta_0$, where

$$\delta_0 = \frac{T-R}{T-P}. \tag{15}$$

Since $\delta_0 \in (0,1)$ for any values of T, R, and P, and since $\delta_0 > \max\{\delta_A, \delta_K\}$ for any $\alpha > 0$ and $\kappa > 0$, we conclude the following. First, conditional on the threat to punish defectors being credible (i.e., $\alpha \leq \alpha^{**}$ and $\kappa \leq \kappa^{**}$, respectively), altruists and moralists are better at sustaining cooperation than selfish individuals. Second, selfish individuals are better at sustaining cooperation than altruists (resp. moralists) if the latter cannot credibly threaten to punish defectors (i.e., $\alpha > \alpha^{**}$ resp. $\kappa > \kappa^{**}$).

Finally, comparing a pair of equally altruistic players with degree of altruism $\alpha \in [0,1]$ to a pair of equally moral players with degree of morality $\kappa = \alpha$, does one pair face a more stringent challenge to sustain cooperation than the other? To answer this question, we distinguish three cases, depending on whether $T-R$ exceeds, falls short of, or equals $P-S$.

Suppose first that $T-R=P-S$. Observe first that this implies $\alpha^{**}=\kappa^{**}=\alpha^{*}=\kappa^{*}$ (where α^{*} was defined in (9) and κ^{*} in (8)). In other words, (C,C) is an equilibrium of the stage game between altruists whenever it is an equilibrium of the stage game between moralists. Moreover, whenever (C,C) is not an equilibrium of the stage game, altruists and moralists are equally capable of credibly threatening to play D following a defection, so that both altruists and moralists can sustain cooperation if sufficiently patient. However, it is easy to verify that $T-R=P-S$ implies $\delta_K>\delta_A$: thus, if $\delta\in[\delta_A,\delta_K)$, grim trigger constitutes a subgame perfect equilibrium that sustains perpetual cooperation for the altruists but not for the moralists.

Second, suppose that $T-R>P-S$. Observe first that this implies $\alpha^{*}>\kappa^{*}$: this means that if $\kappa^{*}\leq\alpha<\alpha^{*}$, then (C,C) is an equilibrium of the stage game between moralists but not of the stage game between altruists. Since $T-R>P-S$ implies $\kappa^{**}>\kappa^{*}$, $\alpha^{**}<\alpha^{*}$ and $\alpha^{**}<\kappa^{**}$, the conclusion is as follows. When $T-R>P-S$ there exist values of α for which altruists are not able to sustain cooperation for any discount factor δ, whereas a pair of moralists with any degree of morality κ can sustain perpetual cooperation; namely, for any $\delta\in[0,1]$ if $\kappa\geq\kappa^{*}$, and for all $\delta\geq\delta_K$ if $\kappa<\kappa^{*}$.

Finally, suppose that $T-R<P-S$. Then it is straightforward to verify that the opposite conclusion obtains, namely that there exist degrees of morality κ for which moralists are not able to sustain cooperation for any δ, whereas a pair of altruists with arbitrary degree of altruism can sustain perpetual cooperation (for any $\delta\in[0,1]$ if $\alpha\geq\alpha^{*}$, and for all $\delta\geq\delta_A$ if $\alpha<\alpha^{*}$).

3.2. Repeated sharing.

The observation that it may be harder for altruists than for egoists to sustain cooperation in an infinitely repeated game was pointed out by Bernheim and Stark (1988, section II.B). We first recapitulate their model. 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.

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

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

for player 1, and

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

for player 2, where $0 < \gamma < 1/2$.[8] A necessary first-order condition for an interior Nash equilbrium 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 and Stark consider the symmetric case when $\alpha_1 = \alpha_2 = \alpha$, in which case the first-order condition is their equation (16).[9] They use this to identify the following unique symmetric Nash equilibrium of the stage game $x = y = x_A$:

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

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 utility evaluated at the stage-game equilibrium is $v^{NE} = (1+\alpha) \cdot [x_A(1-x_A)]^{\gamma}$ and the utility evaluated at the Pareto-optimal strategy pair is $v^C = (1+\alpha) \cdot 4^{-\gamma}$.

Bernheim and Stark consider an infinitely repeated play of this stage game, with discount factor $\delta \in (0,1)$. They note that perpetual play of "cooperation", (x_C, x_C), is sustained in subgame perfect equilibrium by the threat of (perpetual) reversion to (x_A, x_A) iff $\delta \geq \delta_A$, where

$$\delta_A = \frac{v^D - v^C}{v^D - v^{NE}}, \tag{16}$$

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

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

Solving this maximization problem, we find that a player who would optimally deviate from cooperation would play

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

[8] This is the special case when $k = 1$ in Bernheim and Stark (1988).

[9] 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 $\beta/(1-\beta)$.

Noting that for $\alpha = 1$, $x_D = 1/2$ and $v^D = 2 \cdot 4^{-\gamma} = v^C$, we observe that pure altruists do not benefit from deviation. Hence, pure altruists can sustain cooperation irrespective of δ.[10]

Bernheim and Stark proceed by considering a numerical example, $\mu = 0.01$ and $\gamma = 1/4$, and find that the lowest discount factor δ then needed to sustain cooperation is strictly *increasing* with α. In other words, altruism makes cooperation harder. We proceed in parallel with them by setting $\mu = 0.01$, $\gamma = 1/4$ and/ $\alpha > 0.05$. Then $x_A = 1/(1+\alpha)$,

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

and

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

for all α above approximately 0.05. Figure 2 shows that indeed $x_D \leq 1 - \mu = 0.99$ for such values of α.

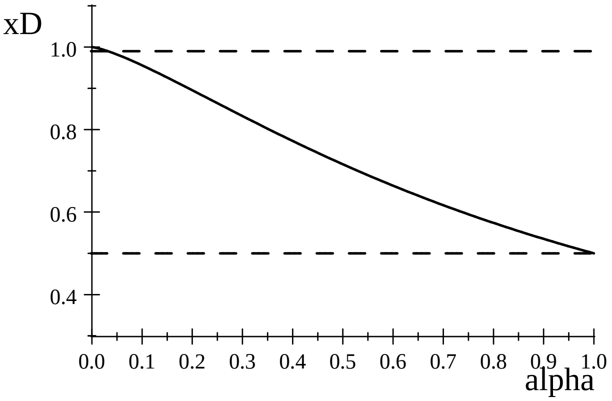

Figure 2: The optimal one-shot deviation for altruists in the repeated game.

For such α,

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

Hence,

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

Figure 3 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 $\alpha \to 1$.

[10] As we will see, a discontinuity will appear in this respect when $\alpha \to 1$.

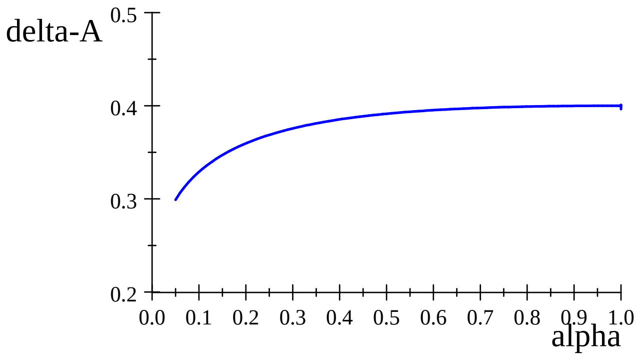

Figure 3: The critical discount factor for cooperation between altruists in the repeated game.

These 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 9 above). In this numerical example, a pair of *Homo oeconomicus* ($\alpha = 0$), can sustain cooperation only if $\delta \gtrsim 0.25$. Altruism thus here 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%.

Morality. We begin by considering the stage-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

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

for player 1, and

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

for player 2, where $0 < \gamma < 1/2$. A necessary first-order condition for an interior Nash equilibrium is thus

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

for player 1, and likewise for player 2. Suppose that $\kappa_1 = \kappa_2 = \kappa$. Then the unique symmetric equilibrium strategy is

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

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 $x_A = x_K$.

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

$$w^{NE} = [x_K (1 - x_K)]^\gamma = \left[\frac{\kappa}{(1+\kappa)^2} \right]^\gamma .$$

The unique symmetric Pareto-optimal strategy is still $x_C = 1/2$, and the utility evaluated at this strategy 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 (x_C, x_C), is sustained in subgame perfect equilibrium by the threat of (perpetual) reversion to (x_K, x_K) if and only if $\delta \geq \delta_K$, where

$$\delta_K = \frac{w^D - w^C}{w^D - w^{NE}}, \tag{17}$$

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

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

Solving this maximization problem, we find that a player who would optimally deviate from cooperation would play $x_{DK} = \min\{x^*, 1 - \mu\}$, where x^* is the unique solution to the fixed-point equation

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

Figure 4 plots the solution as a function of κ, for $\gamma = 1/4$ (and for $\kappa \geq 0.05$).

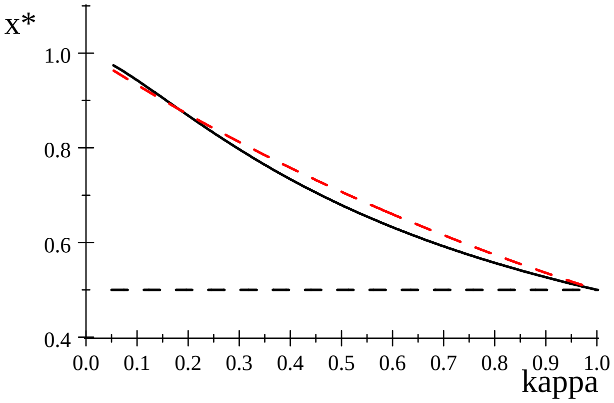

Figure 4: The optimal one-shot deviation for moralists in the repeated game.

We proceed by considering the numerical example that we studied under altruism. Let thus $\mu = 0.01$ and $\gamma = 1/4$, and assume that $\kappa > 0.01$ (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

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

The condition (17) for sustainable cooperation can thus be written as

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

Figure 5 shows 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 between moralists.

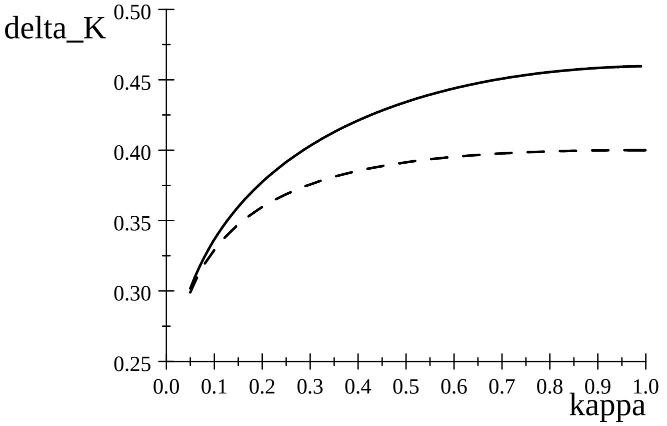

Figure 5: The critical discount factor for cooperation between moralists (solid) and altruists (dashed) in the repeated game.

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

$$\alpha \cdot \left[1 + \alpha^{1/(\gamma-1)}\right]^{1-\gamma} \leq \max_{x \in X} \; (1+\alpha) \cdot [(1-\alpha)^\gamma + \alpha\, (2\,(1-x))^\gamma] \cdot x^\gamma. \tag{18}$$

This inequality clearly holds strictly at $\alpha = 0$, and by continuity also for all $\alpha > 0$ that are small enough. For $\alpha = 1$, (18) holds with equality, since then it boils down to

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

which clearly holds by equality. See Figure 6, which shows isoquants for the difference between the right-hand and left-hand sides in (18). The thick curve is the zero isoquant (where the inequality is an equality) and the thin curves positive isoquants (where the inequality is slack). The diagram suggests that for every $\alpha \in (0,1)$ there exists an $x \in int\,(X)$ such that (18) holds strictly. Hence, the difference between altruism and morality is not due to the approximation of x_{DK}.

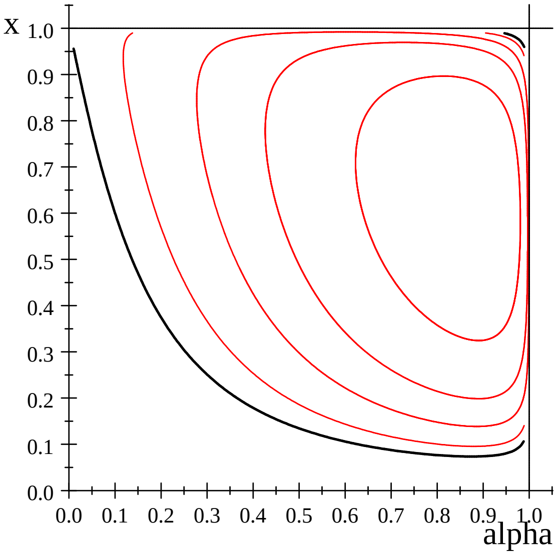

Figure 6: Contour map for the maximand in (18).

3.3. Preference representations. Both in the repeated prisoners' dilemma and in the repeated sharing game, we represented the players' (selfish, altruistic, moral) utility functions over behavior strategies in the repeated game as the normalized present values of their per-period (selfish, altruistic, moral) utilities as defined over their actions in the stage game. Is this consistent with defining their utility functions directly in the repeated game, the game they actually play?

Consider the infinitely repeated play of any symmetric two-player game in material payoffs with common strategy set X and material payoff function $\pi : X^2 \rightarrow \mathbb{R}$, and with common discount factor $\delta \in (0,1)$. In terms of normalized present values, the material payoff function of a player using behavior strategy σ in the repeated game, when the opponent uses behavior strategy τ, is then

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

where x_t is the player's own action in period t and y_t the action of the opponent. The function Π is thus a selfish players' utility function in the repeated game.

First consider altruistic players. By definition, the utility function, in the repeated game, of such a player with degree of altruism $\alpha \in [0,1]$ is

$$\begin{aligned} V_\alpha(\sigma,\tau) &= \Pi(\sigma,\tau) + \alpha \cdot \Pi(\tau,\sigma) \\ &= (1-\delta)\cdot\left(\sum_{t=0}^{\infty}\delta^t \cdot [\pi(x_t,y_t) + \alpha\pi(y_t,x_t)]\right). \end{aligned}$$

Hence, the utility function coincides with the normalized present value representation that we used in our analysis of the prisoners' dilemma and sharing game.

Secondly, for a *Homo moralis* player with degree of morality $\kappa \in [0,1]$, the utility function in the repeated game is, by definition,

$$\begin{aligned} W_\kappa(\sigma,\tau) &= (1-\kappa)\cdot\Pi(\sigma,\tau) + \kappa\cdot\Pi(\sigma,\sigma) \\ &= (1-\delta)\cdot\left(\sum_{t=0}^{\infty}\delta^t\cdot[(1-\kappa)\cdot\pi(x_t,y_t) + \kappa\cdot\pi(x_t,x_t)]\right), \end{aligned}$$

so also the repeated-games utility function of a moralist coincides with the normalized present value representation that we used in the two games.

In sum: the additive separability over time, inherent in the very definition of payoff functions in repeated games, makes the difference between "stage-game preferences" and "repeated-games preferences" immaterial, both in the case of altruism and in the case of morality.

4. 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. Let the material payoff to an individual from choosing A when n_A others choose action A be $n_A \cdot a$. Likewise, let the individual's material payoff from choosing B when n_B others choose B be $n_B \cdot b$, where $0 < b < a$. Examples abound. 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

$$u_i\left(s_i, \boldsymbol{s}_{-i}\right) = as_i \cdot \sum_{j \neq i} s_j + b\left(1 - s_i\right) \cdot \sum_{j \neq i}\left(1 - s_j\right). \tag{19}$$

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

$$v_i\left(s_i, \boldsymbol{s}_{-i}\right) = u_i\left(s_i, \boldsymbol{s}_{-i}\right) + \alpha_i \cdot \sum_{j \neq i} u_j\left(s_j, \boldsymbol{s}_{-j}\right). \tag{20}$$

Evidently the efficient norm A, that is all playing A, can always be sustained as a Nash equilibrium for arbitrarily altruistic players. But also the inefficient norm B is a Nash equilibrium. For if all others choose B, then so will any player i, no matter how altruistic. We will now see that this last conclusion does not hold for moralists.

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

$$w_i\left(s_i, \boldsymbol{s}_{-i}\right) = \mathbb{E}_{\kappa_i}\left[u_i\left(s_i, \tilde{\boldsymbol{s}}_{-i}^m\right)\right], \tag{21}$$

where $\tilde{\boldsymbol{s}}_{-i}^m$ is a random vector in S^{n-1} such that with probability $\kappa_i^m\left(1 - \kappa_i\right)^{n-m-1}$ exactly $m \in \{0, ..., n-1\}$ of the $n - 1$ components of $\boldsymbol{s}_{-i}$ are replaced by s_i, while the remaining components of $\boldsymbol{s}_{-i}$ keep their original values. Thanks to the linearity of the material payoff function (19), the utility function w_i can be written as

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

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

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

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

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

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 examine heterogeneous populations. First, suppose that the coordination game defined above is 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

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

while from taking action A it is

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

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

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

or

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

In other words, 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* $\theta_i\in\mathbb{R}$, where

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

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. Moreover, individuals with high enough degrees of morality have negative thresholds, and will thus take action A even alone. The threshold of an individual with zero degree of morality, that is, *Homo oeconomicus*, is $b/(a+b)$.

Figure 7 below shows the threshold as a function of κ_i for different values of $v=a/b$, and with population shares (in percentages) on the vertical axis. 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.

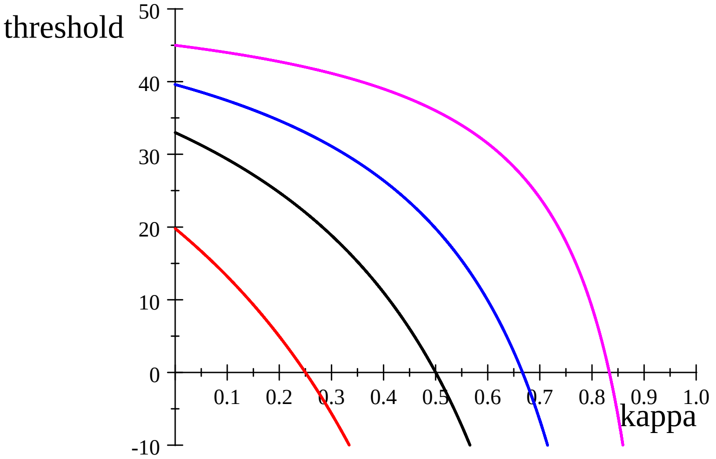

Figure 7: Thresholds for switching to A, as a function of the degree of morality, in a population of size $n = 100$.

Let F be any continuous cumulative distribution function (CDF) 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 $x \in X = [0,1]$.[11] Let $X^* \subseteq [0,1]$ be the non-empty and compact set of such fixed points.

Figure 8 below shows three different CDFs. 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.[12]

[11] 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 $\phi(1) \leq 0$.

[12] 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.

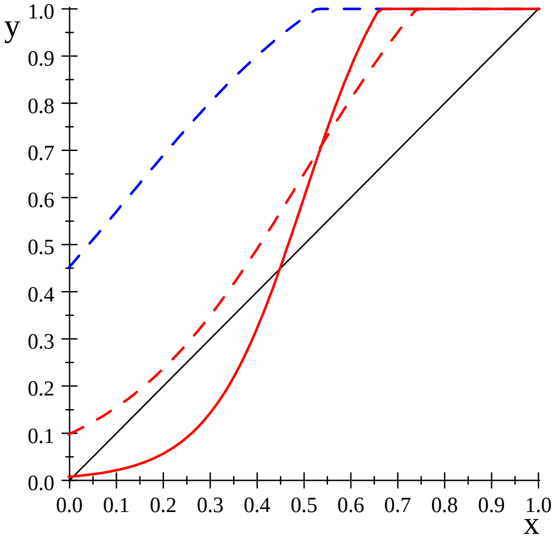

Figure 8: Fixed points for coordination in morally heterogeneous populations.

Figure 8 can be used for discussion of dynamic scenarios. Suppose that initially all individuals were to take action B. All those with non-positive thresholds θ (that is, with relatively high morality) would immediately 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 the population shares taking action A reaches or surpasses $b/\left(a+b\right)$, at which point all remaining individuals will switch to A. This is what may happen in a relatively heterogeneous population with morality distribution such that there is only one fixed point, which then necessarily is $x^{*}=1$. By contrast, in a relatively homogenous population with smallest fixed point $x^{*}<1$, once the adjustment process reaches the point where the population share taking action A is x^{*}, 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 $g\left(\tilde{n}\right)$ be the number of individuals who have that threshold. If $g\left(0\right)=0$, then nobody ever switches to A. But if $g\left(0\right)>0$, the number of individuals $N\left(t\right)$ 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 $N\left(1\right) = g\left(0\right)$, and

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

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

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

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 $N\left(1\right) = 5$, $N\left(2\right) = 5 + 7 = 12$, $N\left(3\right) = 12 + 6 = 18$, $N\left(4\right) = 18 + 3 = 21$, 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, $N\left(1\right) = 1$, but then $N\left(2\right) = 5$, $N\left(3\right) = 10$, $N\left(4\right) = 16$, $N\left(5\right) = 32$, $N\left(6\right) = 100$.

TABLE 1

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

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

5. 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; sometimes altruism and morality have beneficial effects, sometimes altruism is better than morality, sometimes the reverse is true, sometimes they are equivalent, and sometimes self-interest is best! The commonly held presumption that altruism and morality always lead to better outcomes is thus not generally valid. Our analysis unveiled two non-trivial and potentially important phenomena that we believe are robust and general. However, before attacking these two phenonema, we showed that in canonical and one-shot public-goods games with arbitrary many participants, altruism and morality are behaviorally undistinguishable and lead to unambiguously increase material welfare in equilibrium. We also showed that altruism and morality induce different behaviors and outcomes in simple 2×2 games. With these observations as a back-drop, we turned to the above-mentioned two phenomena.

The first phenomenon is that it may be more difficult to sustain long-run cooperation in infinitely repeated interactions between altruists and moralists than between egoists. More specifically, we showed this for infinitely repeated prisoners' dilemmas and infinitely repteated sharing games, in both cases focussing on repeated-games strategies based on the threat of perpetual play of the state-game Nash equilibrium. While altruists and moralists are less tempted to deviate from cooperation and less prone to punish each other—an altruist internalizes the pain inflicted upon the opponent and a moralist internalizes what would happen if both were to deviate simultaneously—the stage-game Nash equilibrium between altruists and between moralists results in higher material payoffs than between self-interested players. This renders the punishment following a deviation less painful, both for the deviator and for the punisher. In the stage games considered here, 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 in order to investigate whether this result obtains for other stage-games and punishment strategies (see e.g. Mailath and Samuelson, 2006).

The second phenomenon is that morality, but not altruism, can eliminate socially inefficient equilibria in coordination games. More precisely, while *Homo moralis* preferences have the potential to eliminate socially inefficient equilibria, neither self interest nor altruism can. The reason is that 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 altruistic individual is solely driven by what others actually do, and hence has no incentive to unilaterally deviate from an inefficient equilibrium. We also showed that when coordination games are played in heterogeneous populations, individuals with a high degree of morality, even if acting myopically, may initiate population cascades away from inefficient equilibria towards a more efficient social "norm". In such cascades, the most morally motivated take the lead and are followed by less morally motivated individuals and may finally be followed even by purely self-interested individuals (when sufficiently many others have switched).

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 behavior and welfare in a few, but arguably canonical, strategic interactions. Clearly, much more theoretical and empirical work is needed for a fuller understanding to be reached, and we hope that this paper can serve as an inspiration.

6. 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.

Allen, B., and C. Tarnita (2014): "Measures of Success in a Class of Evolutionary Models with Fixed Population Size and Structure," *Journal of Mathematical Biology*, 68, 109–143.

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.

Berger, U. and A. Grüne (2016): "On the Stability of Cooperation under Indirect Reciprocity with First-Order Information," *Games and Economic Behavior*, 98, 19-33.

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.

Dhami, S. (2016) *The Foundations of Behavioral Economic Analysis*, Oxford: Oxford University Press.

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.

Ellingsen, T., and M. Johannesson (2008): "Pride and Prejudice: The Human Side of Incentive Theory," *American Economic Review*, 98, 990-1008.

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.

Lehmann L., and F. Rousset (2012): "The Evolution of Social Discounting in Hierarchically Clustered Populations," *Molecular Ecology*, 21, 447-471.

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.

Mailath, G., and L. Samuelson (2006): *Repeated Games and Reputations.* Oxford University Press: Oxford.

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.

Ohtsuki, H. (2010): "Evolutionary Games in Wright's Island Model: Kin Selection Meets Evolutionary Game Theory," *Evolution*, 64, 3344–3353.

Peña J., G. Nöldeke, and L. Lehmann (2015): "Evolutionary Dynamics of Collective Action in Spatially Structured Populations," *Journal of Theoretical Biology*, 382, 122-136.

Perc, M., J.J. Jordan, D.G. Rand, Z. Wangf, S. Boccaletti, and A. Szolnoki (2016): "Statistical Physics of Human Cooperation," *Physics Reports*, 687, 1-51.

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

Sarkisian, R. (2017): "Team Incentives under Moral and Altruistic Preferences: Which Team to Choose?" Working Paper Toulouse School of Economics.

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

Szabó, G., and I. Borsos (2016): "Evolutionary Potential Games on Lattices," *Physics Reports*, 624, 1-60.

Van Cleve, J., and E. Akçay (2014): Pathways to Social Evolution: Reciprocity, Relatedness, and Synergy," *Evolution*, 68, 2245–2258.

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