Half-Hearted Cooperation: A Theory of the Evolution of Altruistic Preferences*

Alexander White†

March 16, 2013

Abstract

This paper presents an evolutionary model in which altruists and egoists simultaneously survive natural selection. Successive generations of randomly paired agents play a two-stage game consisting first of a choice of technology and second a choice of effort level. This setting induces a form of cooperation if *at least one* of the pair is an altruist but not when both are egoists. As a result, in the population steady state there is a positive fraction of both types of agent that is a function of the technology and of the altruists' degree of regard for their opponents.

PRELIMINARY DRAFT

Keywords: Altruism, Cooperation, Economics and Biology, Evolution of Preferences, Contribution and Free-Riding, Prisoners' Dilemma, Technology Adoption

JEL Codes: D64, D74, H41, O33, L23

*I thank Gilles Saint-Paul and Eyal Winter for helpful comments as well as Tesary Lin and Hang Zhou for terrific research assistance. This article is based on a paper written during my doctoral studies at the Toulouse School of Economics, and I am especially grateful to Paul Seabright for his advising and support of this project.

†Tsinghua University School of Economics and Management; awhite@sem.tsinghua.edu.cn

1 Introduction

People often cooperate with one another. The fact that they do has been the subject of a great deal of study revolving around a particular puzzle: in individual instances when people cooperate each one would seem to be better off if he didn't. Given this, how is cooperation sustained? Or, from an evolutionary perspective, what forces allow cooperative behavior to survive natural selection?

How might an outside observer witnessing cooperation explain why it took place? He might say that, despite the appearance given by the individual occurrence, each individual was really pursuing his longer-term own self-interest. Or, he might say that each cooperator values not only his own well-being, but also that of the others.

These two answers represent a dichotomy between two general types of explanation for cooperative behavior, which we might call 'circumstantial' and 'motivational'. The first appeals to the presence of relevant circumstances beyond the specific instance which dictate that a self-interested person cooperate on the particular occasion in question. The second appeals to the idea that the cooperating individuals are motivated by something other than their own self-interest.

In this paper, I suggest that the *both* of these answers could be correct, but that the relation between the two is interdependent in a subtler way than one might imagine. While sometimes, circumstances lead a purely self-interested person to behave cooperatively, this set of circumstances depends on whom he is dealing with. The presence of an altruistic partner can lead an egoist to cooperate in situations where he wouldn't have if his partner were also an egoist. In such situations, the outside observer would have to look closely in order to decipher each individual's true motivations. Upon seeing mutual cooperation, he could be sure that at least one of the two parties was altruistic. Without a finer lens through which to view the interaction, however, he might remain unsure whether both were altruists.

Such situations where mutual cooperation depends on the presence of at least one altruist provide a possible explanation for how altruists and egoists could both be evolutionarily successful. If altruists can engage in successful cooperation with whomever they encounter, whereas egoists require an altruistic partner, the greater *frequency* of cooperation favors the

altruists. However, if when egoists interact with altruists, the former does somewhat better, then the *higher payoff in particular instances* of cooperation favors the egoists. This type of situation could thus support the belief that in the long run, the population can arrive at a mixed equilibrium consisting of both altruists and egoists.

What type of situation might lead egoists to cooperate only with altruists? The kind that I consider in the paper is one in which two parties must first commit to a means of survival and then rely on this particular means as a source of sustenance. So, once the survival technique has been chosen, there is further work left to be done.

Suppose that there are two survival techniques, C and D, available for each individual to pick. On the on hand, C is more efficient than D. On the other hand, the D allows the individual who selects it to capture all of the output he produces, whereas C confers some benefit on the other. Furthermore, suppose that when both choose C, they must to some extent work collectively, so that if they both choose this technique, each retains a greater proportion of his own output than he would if the other had chosen D.

In what follows, I show that in situations of this sort, an egoist paired with an altruist will choose the more efficient technique, C. In contrast, when two egoists are paired together, both will choose D. As a result of this partner-dependent choice on the part of the egoist, both altruists and egoists are evolutionarily viable.

To show this, I offer a game-theoretic model that borrows elements from different areas of the literature that examines why cooperative behavior might emerge in an evolutionary setting.[1] As in many models of cooperation, dynamics play a central role. In my model, the game takes place over two stages. In the first, each player must select a technology, and in the second, each must choose an amount of effort. Because of this structure, the model contains the typical feature whereby egoists' decision to cooperate *now* is driven by the fact that *in the future* there will be common knowledge of the decision they have taken.

Unlike many models, mine contains no repeated interaction. This separates it from a vast body of research that relies on the notion of indefinite repetition as the basis for cooperation. As a result, the perspective I offer differs from that taken by notable authors such as Ken Binmore, who, in a series of books (Binmore, 1994, 1998, 2005), argues that the relevant type of

[1] A seminal work that is a catalyst for much of this literature is Trivers (1971). See Bergstrom (2002) for an excellent survey of subsequent developments.

situation to study in order to understand cooperation is best represented by the indefinitely-repeated prisoner's dilemma. Using this framework, he develops a theory in which morality is a product of self-interested players' involvement in games with multiple equilibria.

The approach used in this model is similar to Binmore's in that the players are versions of what he terms *homo economicus*. In other words, their behavior is characterized by the maximization of rational preferences, which depend on their own, and, in the case of altruists, their partner's 'genetic fitness'. This latter quantity is then used as the replication dynamic in determining the population equilibrium. Thus, this paper follows what is known in the literature as the 'indirect evolutionary approach', pioneered by Güth and Yaari (1992) and Güth (1995) and adopted, for example, by Samuelson and Swinkels (2006) and Dekel, Ely, and Yilankaya (2007). One recent work following this approach, Alger and Weibull (2010), also considers the interaction between altruism and effort choice. However, that paper's focus is quite different in that it considers the effect that the prospect of *ex post* transfers among altruistic kin has on an earlier choice of effort level.

The approach I (as well as the above authors) take contrasts with many evolutionary models whose players are what Binmore author calls *homo behavioralis* (Binmore, 1994, pp. 187). In models of this type, the individual agents' strategies are predetermined by 'nature' and thus the true players might be thought of as genes, which 'program' their hosts' behavior and reproduce in proportion to their hosts' payoffs.

Two models that use versions of homo behavioralis but which resemble mine in that they use cooperation as a device leading to mixed populations are those of Frank (1987)[2] and, more recently, Choi and Bowles (2007). In the first, both 'honest' and 'dishonest' agents give off noisy signals of their type and can thus coexist in a one-shot prisoner's dilemma setting with payoff-based replication. In the second, the population consists of groups of two-dimensional agents that can be both altruistic or not and 'parochial' or not. After playing an intra-group public goods game, each engages in either inter-group war or inter-group peace with another. The authors find that both groups with many parochial altruists and groups with many tolerant egoists tend to be prolific at steady state. A potentially appealing feature of the mechanism presented in the current model, that differentiates it from these two is that it does not depend

[2]In Frank's model, agents are not totally void of rationality, since, despite being able to play only one strategy in the game, they exercise preferences in their choice of partners.

on 'group selection', whereby altruists' success is due to their ability to locate and pair up with other altruists. Instead, I assume that players are matched randomly, and the paper's argument is robust to variations along this dimension.

The rest of the paper is organized as follows: Section 2 presents the basic model. Section 3 studies the game played by each generation of agents, and Section 4 analyzes the population steady state. Section 4 concludes.

2 The Model

The model takes the following structure. First, a given generation of agents plays the game. Second, they reproduce according to their equilibrium fitness. The equilibrium fitness from the game is used to calculate the steady state proportion of each type of player, and this is considered to be the population's long-run equilibrium. The main result is that due to their ability to induce egoists to cooperate, altruists – with seemingly disadvantageous utility functions – survive alongside egoists in the game's evolutionary process.

In a given generation, players from the population are randomly matched and then play a two-stage game, the equilibrium outcome of which depends on the their utility functions. These utility functions, however, need not correspond directly to the players' reproductive fitness. For egoists, or 'self-regarding' players, fitness and utility are represented by the same function, while altruists, or 'other-regarding' players' utility functions depend on their own fitness and on that of the player with whom they are paired, which we refer to as their 'opponent'.

In the first stage of the game, the players' moves determine the technology for producing fitness that will be used in the second stage. One of these technologies, C, is more efficient than the other; however, if one player cooperates, i.e. chooses C, and the other player chooses D, then the latter, defecting player, captures, at the margin, all of the addition fitness afforded by the more efficient technology. If both cooperate, then the total output is split equally. In the second stage, the technology having been determined, players choose a level of effort.

In Section 3, we show that, while two self-regarding players *never* cooperate with one another, each *does* cooperate when paired with an other-regarding player, provided that

the latter is altruistic enough and provided that the additional surplus generated by the 'cooperative technology', C, is great enough. In Section 4, we show that when such an equilibrium exists, the population steady state may entail a relatively even mix of self and other-regarding players.

2.1 The Players

We assume all players' utility functions to take the form $u_i = f_i + \theta_i f_j$, where the parameter θ_i reflects player i's degree of regard for his opponent, j's, fitness. We assume that θ_i is a genetically inherited trait and is therefore, from player i's perspective, an exogenously given feature of his preference ordering.

We restrict attention to the case of a population that consists of two types of players. One type, the self-regarding players, do not care about the fitness of their opponent and thus are of the type $\theta_i = 0$. The second type is other-regarding, and, in the case that we consider, (positively) altruistic, and thus, for such players, $\theta_i = \overline{\theta} > 0$.

2.2 The Game

The players of each generation play a two-stage game, taking one action at each stage. First, they simultaneously select a production technology, and second, they simultaneously select a level of effort. We assume complete information, so each player knows both his opponent's type when deciding which technology to use[3] and the choice of technology that the other player has made when he selects his effort level. Our solution concept is subgame perfect equilibrium.

First, each individual selects a production technology-cooperation (C) or defection (D). We denote this action by σ, where $\sigma = \sigma_i \times \sigma_j \in \{C, D\} \times \{C, D\}$. Second, each player selects a level of effort, e_i. The fitness payoffs are shown in the following matrix,

[3]In future research, we plan to relax this assumption that, when players choose their technologies, they have complete information of their oponent's type. Note, however, that, unlike in various prior works, such as Frank (1987), the argument in this paper for the adaptability of altruism does not depend on such knowledge of other players' types as a determinant of which players are matched with each other. Moreover, we do not assume that altruists have a greater ability than egoists to recognize other altruists, e.g., through a 'secret handshake' (see Robson (1990)).

	$\sigma_j = C$	$\sigma_j = D$
$\sigma_i = C$	$f_i(e_i, e_j, \sigma) = \frac{1+\alpha}{2}(\sqrt{e_i} + \sqrt{e_j}) - e_i,$ $f_j(e_j, e_i, \sigma) = \frac{1+\alpha}{2}(\sqrt{e_j} + \sqrt{e_i}) - e_j$	$f_i(e_i, \sigma) = \sqrt{e_i} - e_i$ $f_j(e_j, e_i, \sigma) = \sqrt{e_j} + \alpha\sqrt{e_i} - e_j$
$\sigma_i = D$	$f_i(e_i, e_j, \sigma) = \sqrt{e_i} + \alpha\sqrt{e_j} - e_i$ $f_j(e_j, \sigma) = \sqrt{e_j} - e_j$	$f_i(e_i, \sigma) = \sqrt{e_i} - e_i$ $f_j(e_i, \sigma) = \sqrt{e_i} - e_i$

where $\alpha \in (0, 1)$ is a technological parameter, which remains fixed throughout the evolutionary process.

We note a few basic features of this fitness matrix. When player i defects, he receives all of the output from his effort, e_i, whereas when player i cooperates, his effort exerts a positive externality on the opponent, j. When both players defect, the chance to use the more efficient technology is sacrificed completely, giving

$$f_i(e_i, \{D, D\}) + f_j(e_j, \{D, D\}) = \sqrt{e_i} + \sqrt{e_j} - e_i - e_j.$$

From the bottom right-hand corner of the matrix, we see that in the case of DD, the distribution is such that only a player's own effort affects his fitness.

On the other hand, when both players cooperate, both produce using the more efficient technology, which is, collectively,

$$f_i(e_i, e_j, \{C, C\}) + f_j(e_j, e_i, \{C, C\}) = (1 + \alpha)(\sqrt{e_i} + \sqrt{e_j}) - e_i - e_j.$$

As shown in the top left-hand corner of the above matrix, when both players cooperate, the output is split evenly between the two players.

Finally, when i cooperates and j defects, only players i's production uses the more efficient technology, while player j's uses the less efficient one, yielding

$$f_i(e_i, \{C, D\}) + f_j(e_j, \{D, C\}) = (1 + \alpha)\sqrt{e_i} + \sqrt{e_j} - e_i - e_j.$$

As we see in the matrix, the *distribution* of additional fitness depends crucially on the combined stage 1 decision, σ. Supposing that the opponent, j, cooperates by choosing C, then player

i faces a tradeoff between the use of the more efficient technology, which he can use if also chooses C, and the ability to capture all of the technological benefits, associated with j's effort, that come about from j's choice of C. We now solve for the outcome of the game.

3 Analysis of the Game

We begin by summarizing our results, which we prove in the following two sections. If cooperation is sufficiently more efficient than defection and if other-regarding players are sufficiently altruistic, then for both combinations of players that include at least one altruist, the unique subgame perfect equilibrium is such that $\sigma = \{C, C\}$. In particular, we prove a sufficient criterion for this to be the case:

$$\theta_i > \frac{1}{2} \quad \text{and} \quad \alpha > \frac{2\sqrt{26} - 3}{19} \approx 0.38.$$

When egoists face each other, the unique equilibrium involves $\sigma = \{D, D\}$. For a large subset of parameter values satisfying these conditions, altruists survive alongside egoists in the population steady state.

3.1 Self-Regarding versus Self-Regarding

As a benchmark, consider, first, the case where two self-regarding players face one another. Since for such agents, we have $\theta_i = 0$, utility and fitness are equivalent, $u_i = f_i$. Using backward induction from the above fitness matrix gives Lemma 1.

Lemma 1. *When two self-interested players face each other, the game reduces to a one-stage game with the following payoff matrix:*

	$\sigma_j = C$	$\sigma_j = D$
$\sigma_i = C$	$\frac{3(1+\alpha)^2}{16}, \frac{3(1+\alpha)^2}{16}$	$\frac{1}{4}, \frac{1+2\alpha}{4}$
$\sigma_i = D$	$\frac{1+2\alpha}{4}, \frac{1}{4}$	$\frac{1}{4}, \frac{1}{4}$

In this reduced-form one-stage game, the strategy D weakly dominates C. In the unique equilibrium in weakly dominant strategies, $\{D, D\}$, both players obtain a fitness level of $\frac{1}{4}$.[4]

Proof. See Appendix A.1. □

Let us briefly analyze the above one-stage game. Clearly it is part of a class of one-shot, complete information games of the form

	Cooperate	Defect
Cooperate	x_2, x_2	x_1 , x_3
Defect	x_3 , x_1	x_1 , x_1

where $x_3 > x_2 > x_1$. However, if, in this game, we were to straightforwardly give the row player the same form of other-regarding preferences as those that we consider in our model, we would have a game of the form

	Cooperate	Defect
Cooperate	$(1+\theta_i)x_2, x_2$	$x_1 + \theta_i x_3$, x_3
Defect	$x_3 + \theta_i x_1$, x_1	$(1+\theta_i)x_1$, x_1

As is immediately apparent, for the column player *Defect* remains a weakly dominant strategy. Since, $x_3 > x_2$, the column players' best reply to *Cooperate* is *Defect* and thus there is no possibility for a Nash equilibrium in which both players cooperate. In our two-stage

[4] I shall assume away the unlikely Nash equilibrium in which, when two self-regarding players face each other, one cooperates and the other defects, as this requires the cooperating agent to play a weakly dominated strategy. Moreover, the model can be easily modified, at some notational expense, to make $\{D, D\}$ a strictly dominant strategy in the reduced-form one-stage game played by egoists (and thus for it to take the form of a classic prisonners' dilemma). This can be done, for example, by assuming that, if a player chooses C, he incurs a small fixed cost. All of the qualitative results of the paper carry through under such an assumption.

game, on the other hand, as the proceeding discussion shows, when self-regarding players face other-regarding players, equilibrium *can* entail $\sigma = \{C, C\}$.

3.2 General Pairings

We now look at the cases where self-regarding players face other-regarding players and where other-regarding players face one another. The general stage-one utility payoff matrix, for players of any combination is given in Lemma 2. Having proved this, we will then analyze the two remaining cases by plugging in the appropriate values for θ_i and θ_j.

Lemma 2. *For a given pair of players, i and j, with any combination of preferences, the game can be rewritten as a one-stage game with the following utility payoff matrix for player i:*

	$\sigma_j = C$	$\sigma_j = D$
$\sigma_i = C$	$\frac{(1+\alpha)^2\left(3+\theta_i(3+\theta_i-\theta_j^2)+2\theta_j\right)}{16}$	$\frac{1+\theta_i(1+2\alpha+\alpha^2\theta_i)}{4}$
$\sigma_i = D$	$\frac{1+2\alpha+\theta_i(1-\alpha^2\theta_j^2)+2\alpha^2\theta_j}{4}$	$\frac{1+\theta_i}{4}$

Proof. See Appendix A.2. □

Other-Regarding versus Self-Regarding

We now analyze the case in which an other-regarding player is matched with a self-regarding one. In Proposition 1, we will state a sufficient condition for a unique, cooperative equilibrium. Before doing this, we write the payoff matrix and then state and prove two preliminary lemmas. We let player i, the column player, be other-regarding, and player j, the row player, be self-regarding. To find the payoff matrix, we set $\theta_i = \overline{\theta}$ and $\theta_j = 0$. Plugging these into the general pairings matrix from Lemma 2, we get:

	$\underline{\theta_j = 0}$	
	$\sigma_j = C$	$\sigma_j = D$
$\sigma_i = C$	$\frac{(1+\alpha)^2(3+3\overline{\theta}+\overline{\theta}^2)}{16}, \frac{(1+\alpha)^2(3+2\overline{\theta})}{16}$	$\frac{1+\overline{\theta}+\alpha(2\overline{\theta}+\alpha\overline{\theta}^2)}{4}, \frac{1+2\alpha(1+\alpha\overline{\theta})}{4}$
$\underline{\theta_i = \overline{\theta}}$		
$\sigma_i = D$	$\frac{1+2\alpha+\overline{\theta}}{4}, \frac{1}{4}$	$\frac{1+\overline{\theta}}{4}, \frac{1}{4}$

To interpret this payoff matrix, let us first note that, for the other-regarding player i, if the self-regarding player j defects, i's best response is to cooperate, since $\alpha\left(2\overline{\theta}+\alpha\overline{\theta}^2\right) > 0$. The certainty of this result is most easily understood in connection with the fact that, when self-regarding players face one another, D *weakly* dominates C, and that in the purely self-interested case, C is in fact one of two best responses to D. Here, when we add to player i's preferences, an arbitrarily small level of regard for the opponent, for him, D ceases to be a best response to j's playing D.

We have established that C is i's best response to j's playing D. We now look at j's best response to i's playing C. Lemma 3 gives the relevant condition.

Lemma 3. *For a self-regarding player j facing an other-regarding player i, j's unique best response to $\sigma_i = C$ is the cooperative action, $\sigma_j = C$, if and only if $\theta_i > \frac{1}{2}$. (In the case where $\theta_i = \frac{1}{2}$, both C and D are best responses.)*

Proof. See Appendix A.3. □

This condition, determining whether or not C is a best response for the self-regarding player j, is *independent* of the technology parameter, α. The explanation for this is that α plays what is essentially an analogous role in j's fitness, whether he chooses D or C, since in either case he receives fitness from the effort of his other-regarding opponent. In i's choice of how much effort to make, it is, thus, exclusively the value of $\overline{\theta}$ that influences whether the collective

incentive to expend effort that goes towards j's fitness is sufficient to warrant j's choosing the cooperative technology. [5]

While the opponent must be substantially other-regarding in order for the self-regarding player to cooperate, it is nevertheless, the opponent's marginal preference for his own fitness that plays the key role in making C the self-regarding player's best response. *By cooperating, the self-regarding player allows the other-regarding player to reap more of the benefits of his own effort.* This induces the other-regarding player to exert more effort; and this, then, benefits the self-regarding player.

Next we look at the other-regarding player's decision, in order to verify the conditions under which CC is, in fact, an equilibrium. We state first the (less tidy) condition for C to be a best response for the other-regarding player to the self-regarding player's C.

Lemma 4. *For an other-regarding player, i, facing a self-regarding player, j, i's unique best response to $\sigma_j = C$ is the cooperative action, $\sigma_i = C$, if and only if $\alpha > \frac{A+2\sqrt{B}}{C}$, where $A = 1 - \theta_i(3 + \theta_i)$, $B = 1 + \theta_i^2(2 + \theta_i)$ and $C = 3 + \theta_i(3 + \theta_i)$. (In the case of equality, both C and D are best responses.)*

Proof. See Appendix A.4. □

This condition shows that for the other-regarding player, the technological superiority of the cooperative production method must be sufficiently great, in order for *him* not to want to defect given that the self-regarding player cooperates. In the event that this condition does not hold but the other-regarding player has a value of $\overline{\theta}$ greater than $\frac{1}{2}$, then there is no pure-strategy equilibrium in the game. This is because the other-regarding player's preference for defection against cooperation would, as it were, take the game to DD, but then, his preference for cooperation against defection would bring us back to CC, so there would be no compatible best responses. We can now prove Proposition 1.

Proposition 1. ***Sufficient condition for a unique cooperative equilibrium***
If (a) $\theta_i > \frac{1}{2}$ and (b) $\alpha > \frac{2\sqrt{26}-3}{19} \approx 0.38$, then when an other-regarding player, i, faces a self-regarding player, j, the unique equilibrium is the cooperative one where $\sigma = \{C, C\}$.

[5] We should note that this result is a product of the imposed 50-50 distribution of output when CC is chosen. It would be interesting to examine the results in cases where the CC distribution was allowed to vary or was the product of some form of bargaining.

Proof. We have shown in Lemma 1 that (a) is a necessary and sufficient condition for $\sigma_j = C$ to be a best response to $\sigma_i = C$. From Lemma 1, we have that $\alpha > \frac{A+2\sqrt{B}}{C}$ is a necessary and sufficient condition for $\sigma_i = C$ to be a best response to $\sigma_j = C$. Plugging $\theta_i = \frac{1}{2}$ into we have $\alpha(\theta_i = \frac{1}{2}) = \frac{2\sqrt{26}-3}{19}$. Since the inequality that must hold is

$$\frac{(1+\alpha)^2(3+3\theta_i+\theta_i^2)}{16} > \frac{1+2\alpha+\theta_i}{4},$$

we have that $\alpha > \frac{A+2\sqrt{B}}{C}$ holds for all $\theta_i > \frac{1}{2}$. □

This says that if the cooperative technology is good enough for an other-regarding player i with a value of θ_i equal to $\frac{1}{2}$ to be willing to cooperate against a self-regarding player, then, *a fortiori*, it is good enough for any player with a higher θ_i also to be willing to cooperate. Note that it is not a necessary condition since for an other-regarding player with a value of θ_i greater than $\frac{1}{2}$, the threshold value of α is lower than 0.38. So, provided α is at least this large, any other-regarding player i with θ_i at least $\frac{1}{2}$ can induce a self-regarding player to cooperate.

Other-regarding versus other-regarding

Finally, we consider the case in which the game is played by a pair of other-regarding players. Here, the utility functions are such that $\theta_i = \theta_j = \overline{\theta}$. If we plug these in to the general utility payoff matrix given in Proposition 2, we get the following matrix (for player i):

	$\underline{\theta_j = 0}$	
	$\sigma_j = C$	$\sigma_j = D$
$\sigma_i = C$	$\frac{(1+\alpha)^2[4(1+\overline{\theta})^2-(1+\overline{\theta})^3]}{16}$	$\frac{1+\overline{\theta}(1+2\alpha+\alpha^2\overline{\theta})}{4}$
$\underline{\theta_i = \overline{\theta}}$ $\sigma_i = D$	$\frac{1+2\alpha+\overline{\theta}(1+2\alpha^2-\alpha^2\overline{\theta}^2)}{4}$	$\frac{1+\overline{\theta}}{4}$

As in the case where an other-regarding player faces a self-regarding player, when other-regarding players face each other, C is a best response to D, and the intuition for this is precisely the same as before. Given that the opponent defects, any degree of regard for the opponent's fitness makes C the best response. So, to find the conditions for a unique cooperative equilibrium, it suffices to compare the CC payoff with the DC payoff. Since our interest lies primarily in the case where other-regarding players face self-regarding players, and since it is computationally messy, I will not detail formally the conditions for CC to be an equilibrium in the case of joint other-regarding players.

From the evolutionary standpoint of the model, parameters that yield a CC equilibrium among mutually other-regarding players are of little interest unless they also yield a CC equilibrium when other-regarding agents go up against self-regarding ones. Since, in our model, we imagine that in each generation, players are randomly paired, unless other-regarding players can induce cooperation against selfish ones, they will not be able to survive the evolutionary process. So we simply need to verify that for the crucial parameter values already identified, the CC payoff is greater than the DC payoff. This reduces to the inequality

$$\overline{\theta}^3(\alpha^2 - 3\alpha - 1) + \overline{\theta}^2(\alpha^2 + 2\alpha + 1) + \overline{\theta}(\alpha^2 + 10\alpha + 1) + 3\alpha^2 - 2\alpha - 1 > 0,$$

which can be readily checked to hold for all elements of the set $\{\alpha, \overline{\theta}\} \in [0.38, 1] \times [\frac{1}{2}, 1]$. We note, however, that, for $\overline{\theta}$ sufficiently large, the CC equilibrium among other-regarding agents breaks down. This, however, represents the unlikely scenario in which the agents are so other-regarding, that each 'cannot bear' to see his opponent sacrifice fitness by overexerting himself on behalf of the first, and thus he prefers to deviate to D, in order to induce a change in technology that induces the latter to provide less effort. For our purposes, it thus seems reasonable to choose 1 as a maximum value to consider for $\overline{\theta}$.

To recap what we have shown in Section 3, the two-stage game in which players first choose a technology and then an effort level, if played only by self-interested players, can be rewritten as a one-stage game with a weakly dominant strategy to defect. When we introduce other-regarding preferences to the two-stage game, however, there is significant scope for both players to choose the cooperative technology even when only one of them is other-regarding.

The important factor in the two-stage game that allows for a cooperative equilibrium is the *difference in marginal return to effort* for the other-regarding player when the game is at *CC* versus when the game is at *CD*. Being in the former situation elicits enough additional effort on the part of the player with $\theta_i > \frac{1}{2}$ that the self-regarding player eschews the opportunity to defect. Provided that $\alpha > 0.38$, any other-regarding player who can induce a self-regarding player to cooperate will also, himself, cooperate with the self-regarding player.

4 Evolutionary Analysis

In this section, we examine the population steady state as a function of $\overline{\theta}$ and α, assuming the sufficient criterion stated in Proposition 1 is satisfied, for cooperative equilibria when at least one of the players is an altruist. In particular, we will show that, at steady state, the proportion of other-regarding agents is strictly positive if and only if the worst fitness payoff for other regarding agents is greater than that of self-regarding agents.

In order to prove Proposition 2, we note that we can compute the steady state by setting the average fitness payoffs of the two types equal to one another, since we assume that agents are randomly paired. We assume the population to be a continuous mass of agents, each of whom embodies one of two genes, represented by $\theta = \overline{\theta}$ and $\theta = 0$. After a given generation of agents plays the game, the next generation is populated by a proportion of each type of agent, determined by the average fitness payoff of their 'parents'.

We let $p \in [0, 1]$ denote the fraction of the population with the gene $\theta = 0$ and $1 - p$ the fraction with the other-regarding gene, $\theta = \overline{\theta}$. The steady state value, p^{SS}, thus represents, as a function of θ and α, the point at which neither gene has a tendency to displace the other in the population as a whole. Lemma 5 states the average fitness payoff of each type of agent as a function of p and then gives the steady state proportion of self-regarding agents.

Lemma 5. *(a) Denote the average fitness payoff for self-regarding agents by f_0 and the same for other-regarding agents by $f_{\overline{\theta}}$. Provided that $\{\alpha, \overline{\theta}\} \in (0.38, 1) \times (\frac{1}{2}, 1)$, these are given by the following expressions:*

$$f_0 = p \cdot \frac{1}{4} + [1 - p] \cdot \frac{(1 + \alpha)^2(3 + 2\overline{\theta})}{16}$$

$$f_{\overline{\theta}} = p \cdot \frac{(1+\alpha)^2(3-\overline{\theta}^2)}{16} + [1-p] \cdot \frac{(1+\alpha)^2(3+2\overline{\theta}-\overline{\theta}^2)}{16}$$

(b) Given these average fitness payoffs for the two types of agents the steady state proportion of self-regarding players, p^{SS}, is given by

$$p^{SS} = \frac{\overline{\theta}^2(1+\alpha)^2}{3(1+\alpha)^2 - 4}.$$

Proof. See Appendix A.5. □

We are now in a position to explain the intuition leading up to Proposition 2. From the expressions for f_0 and $f_{\overline{\theta}}$, given in Lemma 5, we see that, when there is cooperation, the highest fitness payoff goes to a self-regarding agent, paired against an other-regarding agent. The second highest cooperative payoff goes to other-regarding agents playing one another, and the third highest to an other-regarding agent against a self-regarding one. Since self-regarding agents, playing intramurally, never use the collective technology, they get such a pairing yields $\frac{1}{4}$ to each agent. The relative size of the cooperative payoffs compared to $\frac{1}{4}$, however, is ambiguous.

If, for example, we assume the value of α to be $\frac{2}{5}$ (near the low end of our assumed range), and take the limit, as $\overline{\theta}$ approaches 1, of the other-regarding agent's fitness payoff when facing a self-regarding one, we find that it equals $\frac{49}{200} < \frac{1}{4}$. As α increases, or $\overline{\theta}$ decreases, the rank of this payoff for the other-regarding agent moves from last place to second-to-last place. In such a case, where this other-regarding payoff occupies last place, the gene $\theta = \overline{\theta}$ will be driven to extinction. Since it is always the case that self-regarding agents receive more fitness than do other-regarding ones in cooperative situations, the indispensable advantage that the $\theta = \overline{\theta}$ gene needs is for its worst fitness payoff be greater than the worst for the self-regarding agents. We now state Proposition 2.

Proposition 2. *The steady state proportion of other-regarding agents is strictly positive if and only if the worst fitness payoff for other-regarding agents is strictly greater than the worst fitness payoff for self-regarding agents.*

Proof. The worst fitness payoff for each type of agent comes when playing against the self-

regarding type. The self-regarding fitness payoff against themselves, $f_{(0,0)}$, is equal to $\frac{1}{4}$. The other-regarding fitness payoff against self-regarding agents,$f_{(\overline{\theta},0)}$, is given by

$$f_{(\overline{\theta},0)} = \frac{(1+\alpha)^2(3-\overline{\theta}^2)}{16}.$$

We show that $f_{(\overline{\theta},0)} > f_{(0,0)} \Leftrightarrow p^{SS} < 1$. Remark that $f_{(\overline{\theta},0)} > f_{(0,0)}$ is equivalent to

$$\frac{(1+\alpha)^2(3-\overline{\theta}^2)}{16} > \frac{1}{4} \quad \Leftrightarrow \quad 1 > \frac{\overline{\theta}^2(1+\alpha)^2}{3(1+\alpha)^2-4} = p^{SS},$$

where the final equality holds due to part (b) of Lemma 5. □

Interpretation

As the proof of Proposition 2 suggests, the proportion of other-regarding genes in the population is in some sense, driven by the relative value of the other-regarding agents' fitness payoff against self-regarding agents versus the self-regarding agents' fitness payoff against themselves. This can be seen more clearly by the following rearrangement of the steady state equation:

$$\frac{(1+\alpha)^2(3-\overline{\theta}^2/p^{SS})}{16} = \frac{1}{4}.$$

This shows that, at steady state, other-regarding agents' fitness payoff against self-regarding agents, with the negative term, $\overline{\theta}^2$, weighted by the proportion of self-regarding players in the population, is equal to the self-regarding agents payoff against their own type. From this equation, we can easily see that as the value of $f_{(\overline{\theta},0)}$ approaches, from above, the value of $f_{(0,0)}$, the proportion of self-regarding genes in the population goes to one.

Using this rearrangement, we are in a better position to make comparisons between each type's best fitness payoffs and between each type's worst fitness payoff. Note that the best payoffs for the two types are, respectively,

$$f_{(0,\overline{\theta})} = \frac{(1+\alpha)^2(3+2\overline{\theta})}{16}$$

for the self-regarding agents, and

$$f_{(\overline{\theta},\overline{\theta})} = \frac{(1+\alpha)^2(3+2\overline{\theta}-\overline{\theta}^2)}{16}$$

for the other-regarding agents. Further note that, when cooperating, self-regarding agents give the effort level $e^{CC}_{(0,\overline{\theta})} = \frac{(1+\alpha)^2}{16}$. We can thus write the difference between the respective best payoffs as

$$f_{(0,\overline{\theta})} - f_{(\overline{\theta},\overline{\theta})} = e^{CC}_{(0,\overline{\theta})} \cdot \overline{\theta}^2. \tag{1}$$

Equation (1) shows that, when self- and other-regarding agents cooperate with one another, the former come away with a fitness differential equal to the amount of effort they give times the square of the other-regarding agent's altruism coefficient.

At *steady state*, there is a similar expression relating the two worst payoffs. Stating the respective worst payoffs, we have, for other-regarding agents,

$$f_{(\overline{\theta},0)} = \frac{(1+\alpha)^2(3-\overline{\theta}^2)}{16},$$

and, for self-regarding agents,

$$f_{(0,0)} = \frac{1}{4} = \frac{(1+\alpha)^2(3-\overline{\theta}^2/p^{SS})}{16}.$$

Taking the difference between these two fitness payoffs gives the expression

$$f_{(\overline{\theta},0)} - f_{(0,0)} = e^{CC}_{(0,\overline{\theta})} \cdot \overline{\theta}^2 \cdot \left[\frac{1-p^{SS}}{p^{SS}}\right]. \tag{2}$$

From equation (2), we see that at steady state, the fitness differential between the worst payoff for the other-regarding agents and the worst payoff for the self-regarding agents is equal to the cooperative effort level of the latter, multiplied by the square of other-regarding agents' altruism coefficient, multiplied by the ratio of other-regarding agents to self-regarding agents in the population.

Interpreted jointly, equations (1) and (2) lead to the following conclusion. On the one

hand, equation (1) shows that the 'selfish gene', $\theta = 0$, is rewarded by natural selection for programming its agents to choose the optimal level of effort for the sake of their own fitness, given the technology that has been chosen. Also, this gene is rewarded for accepting the benefits bestowed on it by the altruistic gene, $\theta = \overline{\theta}$. On the other hand, equation (2) shows the greater frequency with which the altruistic gene is able prompt cooperation. Despite the individual overexertion by agents with this gene, matched together, the other-regarding agents come closer to the optimal collective level of effort. If the former is a more significant boon, then the population tends towards a greater number of self-regarding agents at steady state, whereas if the latter is a greater asset, it tends towards a greater number of other-regarding agents.

5 Conclusion

At the outset of this paper I suggest that some forms of cooperative behavior may be driven by a combination of altruism and self-interested concern for the future. In the model, I have shown a mechanism in which, when pairs of agents interact, one altruist can serve as a catalyst for mutual cooperation, even when his opponent is an egoist. In identical situations, when egoists face each other, they do not cooperate. Since, cooperation comes at some expense to the altruistic agents, after individual instances of cooperation, egoists come away with a higher level of fitness than altruists. However, since altruists elicit cooperation from both types of agent, whereas egoists only from altruists, the altruists achieve a higher worst payoff that the egoists. Calculating the population steady state, I find that this dynamic can lead to stable populations in which both types of agent survive evolution.

In the study of how humans came to be 'the cooperative species', it is a central issue whether or not altruistic traits played an important role. Altruism makes cooperation easier to explain, but, a priori, egoism seems to have a selective advantage. By appealing to situations where commitment implies a distributive indivisibility, this paper shows one mechanism that could have helped altruistic traits overcome their apparent evolutionary disadvantage while at the same time encouraging cooperative behavior by non-altruists.

Appendices

A Omitted Proofs

A.1 Lemma 1

Proof. In each possible state of the second stage, each player i maximizes the function $u_i = f_i$ with respect to e_i. Thus we have the following maximization programs:

(i) If $\sigma = \{C, C\}$, then, in the second stage, player i maximizes $\frac{1+\alpha}{2}\sqrt{e_i} - e_i$, which is solved by $e_i = \frac{(1+\alpha)^2}{16}$. Plugging $e_i = e_j = \frac{(1+\alpha)^2}{16}$ into $\frac{1+\alpha}{2}\left(\sqrt{e_i} + \sqrt{e_j}\right) - e_i$ yields $u_i^{CC} = \frac{3(1+\alpha)^2}{16}$.

(ii) Given any other pair of actions chosen in the first stage, player i maximizes $\sqrt{e_i} - e_i$, which is solved by $e_i = \frac{1}{4}$. If $\sigma = \{C, D\}$ or $\sigma = \{D, D\}$, then plugging this solution into $\sqrt{e_i} - e_i$ gives $u_i^{CD} = u_i^{DD} = \frac{1}{4}$. If $\sigma = \{D, C\}$, then plugging $e_i = e_j = \frac{1}{4}$ into $\sqrt{e_i} + \alpha\sqrt{e_j} - e_i$ gives $u_i^{DC} = \frac{1+2\alpha}{4}$.

□

A.2 Lemma 2

Proof. In each possible state of the second stage, each player, i, maximizes the function $u_i = f_i + \theta_i f_j$ with respect to e_i. Thus, we have the following cases:

(i) If $\sigma = \{C, C\}$, then player i maximizes $\frac{(1+\alpha)(1+\theta_i)}{2}\sqrt{e_i} - e_i$, which is solved by $e_i = \frac{(1+\alpha)^2(1+\theta_i)^2}{16}$. Plugging this and the analogous value of e_j into $\frac{1+\alpha}{2}(\sqrt{e_i} + \sqrt{e_j}) - e_i + \theta_i\left(\frac{1+\alpha}{2}(\sqrt{e_i} + \sqrt{e_j}) - e_j\right)$ gives $u_i^{CC} = \frac{(1+\alpha)^2\left(3+\theta_i(3+\theta_i-\theta_j^2)+2\theta_j\right)}{16}$.

(ii) If $\sigma = \{C, D\}$, then player i maximizes $(1 + \alpha\theta_i)\sqrt{e_i} - e_i$, which is solved by $e_i = \frac{(1+\alpha\theta_i)^2}{4}$. Since the defecting player, j, exerts no fitness externality, we have $e_j = \frac{1}{4}$. Plugging these values into $\sqrt{e_i} - e_i + \theta_i(\sqrt{e_j} + \alpha\sqrt{e_i} - e_j)$ gives $u_i^{CD} = \frac{1+\theta_i\left(1+2\alpha+\alpha^2\theta_i\right)}{4}$.

(iii) If $\sigma = \{D, C\}$, reversing the indices in the equilibrium effort levels found in case (ii) and plugging into $\sqrt{e_i} + \alpha\sqrt{e_j} - e_i + \theta_i(\sqrt{e_j} - e_j)$ gives $u_i^{DC} = \frac{1+2\alpha+\theta_i\left(1-\alpha^2\theta_j^2\right)+2\alpha^2\theta_j}{4}$

(iv) If $\sigma = \{D, D\}$, then, as established in case (ii), we have $e_i = e_j = \frac{1}{4}$. Plugging these values into $\sqrt{e_i} - e_i + \theta_i(\sqrt{e_j} - e_j)$ gives $u_i^{DD} = \frac{1+\theta_i}{4}$.

□

A.3 Lemma 3

Proof. The cooperative action, $\sigma_j = C$ is the unique best response to $\sigma_i = C$ if and only if the inequality $\frac{(1+\alpha)^2(3+2\theta_i)}{16} > \frac{1+2\alpha(1+\alpha\theta_i)}{4}$ holds. This is equivalent to $\theta_i > \frac{1}{2}$. □

A.4 Lemma 4

Proof. The cooperative action, $\sigma_i = C$ is the unique best response to $\sigma_j = C$ if and only if the inequality $\frac{(1+\alpha)^2(3+3\theta_i+\theta_i^2)}{16} > \frac{1+2\alpha+\theta_i}{4}$ holds. We show that this holds iff $\alpha > \frac{A+2\sqrt{B}}{4}$. Equivalent to the above inequality is $(\theta_i^2 + 3\theta_i + 3)\alpha^2 + (2\theta_i^2 + 6\theta_i - 2)\alpha + \theta_i^2 - \theta_i - 1 > 0$. Using the quadratic formula the find the roots of the left hand side gives

$$\alpha = \frac{1 - \theta_i(3 + \theta_i) \pm 2\sqrt{1 + \theta_i^2(2 + \theta_i)}}{3 + \theta_i(3 + \theta_i)} \equiv \frac{A \pm 2\sqrt{B}}{C}.$$

Both 'plus' and 'minus' roots are real for $\theta_i > 0$. We are interested only in positive values of both α and θ_i . But the 'minus' root is positive only when θ_i is negative. Thus we have

$$\alpha > \frac{A + 2\sqrt{B}}{C} \quad \Leftrightarrow \quad \frac{(1 + \alpha)^2(3 + 3\theta_i + \theta_i^2)}{16} > \frac{1 + 2\alpha + \theta_i}{4}.$$

□

A.5 Lemma 5

Proof. (a) In the proof of Lemma 2, we have calculated the level of effort each agent exerts for each state possible stage-two state of the game. Since $\{\alpha, \overline{\theta}\} \in (0.38, 1) \times (\frac{1}{2}, 1)$, we have

(i) when an other-regarding player, i, faces a self-regarding player, j, $\sigma_i = \sigma_j = C$, and thus $f_i = \frac{1+\alpha}{2}(\sqrt{e_i} + \sqrt{e_j}) - e_i = \frac{(1+\alpha)^2(3-\overline{\theta}^2)}{16}$ and $f_j = \frac{1+\alpha}{2}(\sqrt{e_j} + \sqrt{e_i}) - e_j = \frac{(1+\alpha)^2(3+2\overline{\theta})}{16}$,

(ii) when other-regarding agents face each other, we also have $\sigma_i = \sigma_j = C$, hence $f_i = f_j = \frac{1+\alpha}{2}(\sqrt{e_i} + \sqrt{e_j}) - e_i = \frac{(1+\alpha)^2(3+2\overline{\theta}-\overline{\theta}^2)}{16}$,

(iii) when self-regarding agents face one another, $\sigma_i = \sigma_j = D$, so $f_i = f_j = \sqrt{e_i} - e_i = \frac{1}{4}$.

With probability p, a given agent will be paired with a self-regarding opponent and, with probability $1 - p$, he will be paired with an other-regarding opponent. This yields the respective expressions for f_0 and $f_{\overline{\theta}}$.

(b) The steady state proportion of self-regarding players in the population, p^{SS}, is found by equating the average fitness payoffs for self-regarding and other-regarding players:

$$p^{SS} \cdot \frac{1}{4} + [1 - p^{SS}] \cdot \frac{(1 + \alpha)^2(3 + 2\overline{\theta})}{16} = p^{SS} \cdot \frac{(1 + \alpha)^2(3 - \overline{\theta}^2)}{16} + [1 - p^{SS}] \cdot \frac{(1 + \alpha)^2(3 + 2\overline{\theta} - \overline{\theta}^2)}{16}.$$

Solving for p^{SS} yields $p^{SS} = \frac{\overline{\theta}^2(1+\alpha)^2}{3(1+\alpha)^2-4}$.

□

References

Alger, Ingela and Jörgen Weibull (2010), "Kinship, Incentives, and Evolution." *American Economic Review*, 100, 1727–1760.

Bergstrom, Theodore C. (2002), "Evolution of Social Behavior: Individual and Group Selection." *Journal of Economic Perspectives*, 16, 67–88.

Binmore, Ken (1994), *Game Theory and the Social Contract, Volume 1: Playing Fair*. MIT Press, Cambridge, Massachusetts.

Binmore, Ken (1998), *Game Theory and the Social Contract, Vol. 2: Just Playing*. MIT Press, Cambridge, Massachusetts.

Binmore, Ken (2005), *Natural Justice*. Oxford University Press, New York.

Choi, Jung-Kyoo and Samuel Bowles (2007), "The Coevolution of Parochial Altruism and War." *Science*, 636–640.

Dekel, Eddie, Jeffrey C. Ely, and Okan Yilankaya (2007), "Evolution of Preferences." *Review of Economic Studies*, 74, 685–704.

Frank, Robert H. (1987), "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?" *The American Economic Review*, 77, 593–604.

Güth, Werner (1995), "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives." *International Journal of Game Theory*, 24, 323–344.

Güth, Werner and Menahem Yaari (1992), "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game." In *Explaining Process and Change: Approaches to Evolutionary Economics* (Ulrich Witt, ed.), 23–24, University of Michigan Press, Ann Arbor, MI.

Robson, Arthur J. (1990), "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake." *Journal of Theoretical Biology*, 144, 379–396.

Samuelson, L. and J. M. Swinkels (2006), "Information, Evolution and Utility." *Theoretical Economics*, 1, 119–142.

Trivers, Robert L. (1971), "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology*, 46, 35–57.