Preference for mates and the evolution of social norms

Sofia Moroni

Extended Abstract

1 Introduction

A puzzling feature of our species is that we are able to cooperate with and trust fellow humans that we don't expect to encounter in the future and whose true intentions we cannot observe. In this paper we develop an evolutionary theory for the emergence of cooperation. Individuals' fitness depends not only on their ability to produce sustenance for survival but also on their ability to attract a mate to produce offspring. We show that in this environment a preference for cooperation to obtain resources and a taste for mates who cooperate can *evolve simultaneously*. Furthermore, a society with these features cannot be taken over by other preferences. An individual who deviates from cooperative behavior obtains more resource wealth but faces the judgment of potential mates and, therefore, obtains a low fitness overall. Individuals who do not carry the preference for cooperating mates do not have an evolutionary advantage as some of their children will fail to cooperate with others and, hence, struggle to find a mate. Thus, a social norm that requires cooperation can be sustained if members of society have a preference for mates who don't deviate from the social norm.

We develop this argument in a model in which members of society first gather resources and then look for potential mates. Individuals interact with strangers in a resource acquisition game which is modeled as a prisoner's dilemma. Cooperation generates more resources overall, but a player has the option to defect and steal all the resources garnered leaving his fellow player destitute. Because the interaction is one-off, a defector has no chance of being punished by his opponent.

After the resource acquisition game has taken place each player finds a potential mate who observes, perhaps imperfectly, the player's outcome in the resource acquisition game and decides whether to accept or reject the match. A player produce offspring only if he or she successfully forms a match. The number of offspring is proportional to their parents' outcomes in the resource acquisition game. Players' decisions over actions in the game are dictated by their preferences which may not coincide with their actual fitness. Offspring inherit the preference of each parent with probability 1/2.

We ask the following questions. First, what preferences are evolutionarily stable, meaning that they cannot be taken over by a small group of individuals with different preferences? Second, from a given distribution of preferences, what invading groups with different preferences may be favored by evolution, and what is the long run distribution after a successful invasion?

Our first result establishes that a society in which players cooperate and a sufficiently great proportion of the population only accept mates who cooperate is evolutionarily stable. A small invasion by defectors and players who accept defectors does not have an evolutionary advantage. Defectors have little probability of finding a mate and players who accept defectors do either as well as the existing population or worse if their offspring have a penchant for defection.

Our results show that social norms and preferences for mates can give rise to societies in which people cooperate. We obtain cooperation without repeated interactions that allow for punishment, without kin selection, and without "group selection" arguments.[1]

1.1 Related Literature

We build on a literature on the evolutionary origins of cooperation. Classic theories of the origins of cooperation include kin selection (Maynard Smith (1964)) – the idea that evolution will select for cooperation among genetically related in-

[1] In the final section of the paper, we develop an extension in which individuals are able to condition on membership in their own group.

dividuals – “group selection” (Wilson (1975) and a large literature), which posits that groups are able to cooperate among themselves will outcompete groups that fail to cooperate, and reciprocal altruism (Trivers (1971)), in evolution may support cooperative behavior in interactions between individuals who expect to meet again in the future. Models of group selection face the challenge that groups are vulnerable to invasion by defectors. This paper develops a theory of cooperation among unrelated strangers.

In addition, we build on a literature on evolutionary games in economics (Güth and Yaari (1992), Dekel, Ely and Yilankaya (2007), Heifetz, Shannon, Spiegel (2007), Sethi and Somanathan (2002), Kockesen, Ok, Sethi (2000), Robson and Samuelson (2011), Cole, Mailath and Postlewaite (1992)), and on the theory of sexual selection (Fisher (1915), Grafen (1990)).

2 Model

We consider a population that evolves according to a game that determines each individual's evolutionary success. The game consists of two stages. The first stage is the resource acquisition game in which players are matched with other players to play a prisoner's dilemma with the following payoff matrix:

$$\begin{array}{ccc} & C & D \\ C & c, c & 0, a \\ D & a, 0 & d, d \end{array}$$

where $a > c > d > 0$ are even numbers. We assume that $2c > a$ so that cooperation is the most efficient outcome. In the second stage players have to find a partner to produce offspring. The number of offspring of the players depends on their outcomes in the two stages. The players are gendered and are either male or female. Half of the population is male and half is female. For simplicity we assume that players play the resource acquisition game with members of their own gender.

After playing the resource acquisition game, in the second stage males and females are randomly matched to a potential partner. Players observe the outcomes in the resource acquisition game of their match. Players then decide whether to accept or reject their potential partner. Their decision may depend on the outcomes of the resource acquisition game. If any of the two players reject the match then both players die without producing offspring.[2] If the match is accepted by both partners the number of offspring of the couple is the sum of the outcomes in the resource acquisition game of the players. Thus, the fitness of a player (male or female) equals the payoff they obtained in the resource acquisition game in addition to their partner's if they produce offspring and zero otherwise. Players who match either in the resource acquisition game or the offspring game do not observe each other's preferences.

Players may have preferences over the outcomes of the game that do not match their actual evolutionary fitness given by their payoffs in the two-stage fitness game. For example, players may have idiosyncratic preferences over their or potential partner's outcome that influence their decision of whether to accept or reject a match.

Formally, players will be characterized by types that correspond to their preferences over outcomes of the game. Let $\Phi^R = \{(C,C),(D,D),(C,D),(D,C)\}$ be the outcome of the resource acquisition game, where the first coordinate corresponds to a player's action or a potential partner's action. Let $\Phi^O = \{a,r\} \times \{a,r\}$ denote the set of outcomes in the second stage game. The first coordinate corresponds to a player's action (accept or reject) and the second coordinate to the player's second stage match. Thus, each player has a gender $g \in G \equiv \{f,m\}$ and a type θ which is a function $\theta : \Phi^R \times \Phi^R \times \Phi^O \times G \to \mathbb{R}$, where $\theta(o_1^R, o_2^R, o, g)$ is the payoff type θ obtains if the outcome in his or her own resource acquisition game is o_1^R, his/her potential partner's outcome is o_2^R and the outcome in the second stage game is o. Figure 1 illustrates the game. The type of a player is determined by the values at the terminal nodes of the tree. The set of types is denoted Θ.

Half of a couple's offspring is female and half is male. Players inherit the preference of each parent with a probability that depends on the parents' preference types. If a child's parents have types θ_1 and θ_2 the child's preferences are θ_1 with probability $1/2$ and to type θ_2 with probability $1/2$.

[2] Alternatively, we could assume that there are N stages in which players are randomly matched and females decide whether or not to accept the match. Players accept their N'th match but may reject unsuitable matches at a small cost.

Let A denote the set of preference types in the population. A distribution $\alpha \in A$ is of the form $\alpha = (\alpha^f, \alpha^m)$ where α^f and α^m are the distributions over the preferences types in the female population and male populations respectively.

Equilibrium play A behavioral strategy of a player of preference type θ is a mixture over actions in each information set. The resource game strategy of a type θ player of gender g, denoted $b^R_{\theta,g}$, is a mixed strategy over $\{C, D\}$. The second stage strategy, $b^O_{\theta,g} : \Phi^R \times \Phi^R \to \Delta\{a, r\}$ is a function from the resource game outcomes of a player's and the outcome of the player's second-stage match to $\Delta\{a, r\}$. A strategy of a player of type θ has the form $b_{\theta,g} = (b^R_{\theta,g}, b^O_{\theta,g})$. Let B denote the set of strategies.

We assume that play in the population is a Perfect Bayesian Equilibrium. That is, we assume that players have correct beliefs about their opponent's play and choose their actions so as to maximize their expected payoff. In the second stage a player of type θ and gender g, after observing the outcomes $(o^R_1, o^R_2) \in \Phi^R \times \Phi^R$ in the resource acquisition game, chooses

$$b^O_{\theta,g}(o^R_1, o^R_2) \in \text{argmax}_{\sigma\in\Delta\{a,r\}} \int \theta\left(o^R_1, o^R_2, \left(\sigma, b^O_{\tilde{\theta},-g}(o^R_1, o^R_2)\right), g\right) d\alpha^{-g}(\tilde{\theta}). \qquad (1)$$

where $-g$ denotes g's opposite gender and the expectation is taken with respect to the distribution of the type that g encounters in the offspring game. In the resource acquisition game a player of type θ and gender g chooses $b^R_{\theta,g}$ such that

$$b^R_{\theta,g} \in \text{argmax}_{\sigma\in\Delta\{C,D\}} \int \theta\left((\sigma, b^R_{\tilde{\theta},g}), (b^R_{\tilde{\theta}_1,-g}, b^R_{\tilde{\theta}_2,-g}), (b^O_{\theta,g}, b^O_{\tilde{\theta}_1,g}), g\right) d\alpha^{g}(\tilde{\theta}) d\alpha^{-g}(\tilde{\theta}_1) d\alpha^{-g}(\tilde{\theta}_2), \qquad (2)$$

where the expectation is taken over the type of the matches that type θ encounters at each stage of the game.[3] A strategy profile $b = (b_{\theta^f,f}, b_{\theta^m,m})_{(\theta^f,\theta^m)\in\text{supp}(\alpha)}$ that satisfies conditions (1) and (2) is a Perfect Bayesian Equilibrium when the distribution of types in the population is $\alpha = (\alpha^f, \alpha^m)$. Let $B(\alpha)$ denote the set of Perfect Bayesian equilibria of the game when the distribution of types in the population is α.

[3] Note that $(b^O_{\theta,g}, b^O_{\tilde{\theta}_1,-g})$ depends on the realized outcomes of the resource acquisition game. These may be random if players are mixing. We have omitted the dependence for ease of notation.

A *configuration* is a bundle (α, b) where $\alpha \in A$ and $b \in B$. A configuration (α, b) is an *equilibrium configuration* if $b \in B(\alpha)$.

Evolution As the each generation plays the two-stage game the distribution of types and equilibrium play will evolve. Given an initial distribution of types α_0, players choose their actions according to some equilibrium strategy $b_0 \in B(\alpha_0)$. Each player produces offspring according to their payoffs in the fitness game and the distribution of types in the next generation will be determined by the inheritance rules. Let $\Omega(\alpha_0, b_0)$ the set of configurations after one generation has passed after the initial configuration (α_0, b_0).

Definition 1. A *evolutionary sequence starting from initial equilibrium configuration* (α_0, b_0) is a sequence of $\{(\alpha_k, b_k)\}_{k=0}^{\infty}$ such that $(\alpha_k, b_k) \in \Omega(\alpha_{k-1}, b_{k-1})$ for $k \geq 1$.

A *long run configuration of a distribution* α_0, (α^*, b^*), is the limit configuration of an evolutionary sequence starting from an initial equilibrium configuration (α_0, b_0) for some $b_0 \in B(\alpha_0)$. Note that a long run configuration may not exist if all the evolutionary sequences starting from equilibrium configurations containing α_0 do not have a limit. We refer to distribution of types α^* and strategies b^* as a long run distribution and a long run equilibrium play of initial distribution α_0, respectively.[4]

Invasions and stability We consider the possibility that the distribution of preferences is subject to mutations or small invasions by populations with different preferences. Let $\widehat{\delta} > 0$ and let θ be a preference type. We define a $(\widehat{\delta}, \theta)$-*invasion of a distribution* α as the distribution of types such that a proportion $(1 - \widehat{\delta})$ of the population has preferences according to the original distribution α, a proportion $\widehat{\delta}$ corresponds to type θ players of which half are male and half are female. For short, we will denote a $(\widehat{\delta}, \theta)$-invasion of α as $(1 - \widehat{\delta})\alpha + \widehat{\delta}\theta$.

An invasion of players who have different preferences could destabilize the original distribution of types as the newcomers preferences could become a higher

[4] Note that because (α^*, b^*) is a limit of equilibrium configurations it must also be an equilibrium configuration.

proportion of the population generation after generation if them and their descendants obtain higher payoffs in the two-stage game than the incumbent preference types. Distributions that are evolutionary stable must be robust to small invasions of new preference types. We define the fitness of a type θ as follows. Given the initial configuration (α_0, b_0), let $(\alpha_2, b_2) \in \Omega\left(\Omega(\alpha_0, b_0)\right)$. That is (α_2, b_2) is an equilibrium configuration after two generations starting from configuration (α_0, b_0). The *fitness of a preference type θ* given initial configuration (α_0, b_0) and $(\alpha_2, b_2) \in \Omega\left(\Omega(\alpha_0, b_0)\right)$ is given by

$$\Pi(\theta|(\alpha_0, b_0), \alpha_2) = \frac{\alpha_2(\theta)/\int \alpha_2(\theta)\,d\theta}{\alpha_0(\theta)/\int \alpha_0(\theta)\,d\theta}.$$

.

Thus, the fitness of a type θ is its proportion after two generations relative to its proportion in the initial distribution.

The proportion of the population of a preference type in the long run depends on the not only on the number of offspring that carry the preference but also how these children perform in their offspring game. As the choice of partner affects both variables the appropriate notion of fitness in our setting relates to the outcomes of preference types after two generations instead of one.

We define stability as follows.

Definition 2. An equilibrium configuration (α, b) is *stable* if for every preference type θ there is $\delta > 0$ such that if $\widehat{\delta} < \delta$, every $(\widehat{\delta}, \theta)$-invasion

1. There is $(\alpha_2, b_2) \in \Omega(\Omega((1-\widehat{\delta})\alpha + \widehat{\delta}\theta, b_0))$ such that $\Pi(\tilde{\theta}|(\alpha_0, b_0), \alpha_2) \geq \Pi(\theta|(\alpha_0, b_0), \alpha_2)$, for every $\tilde{\theta} \in \text{supp}(\alpha)$ and;
2. has a long run configuration, (α^*, b^*), such that $(b^*)^R = b^R$ and $(b^*)^O(o_1^R, o_2^R) = b^O(o_1^R, o_2^R)$ for every outcome (o_1^R, o_2^R).

That is, a configuration (α, b) is stable if in the long run distribution equilibrium play coincides with equilibrium play under (α, b), and the fitness of the preference θ is weakly less than the fitness of any preference in the support of α. The latter implies that in the long-run distribution the proportion of the population that is of the invader type is weakly less than its proportion right after the invasion. This

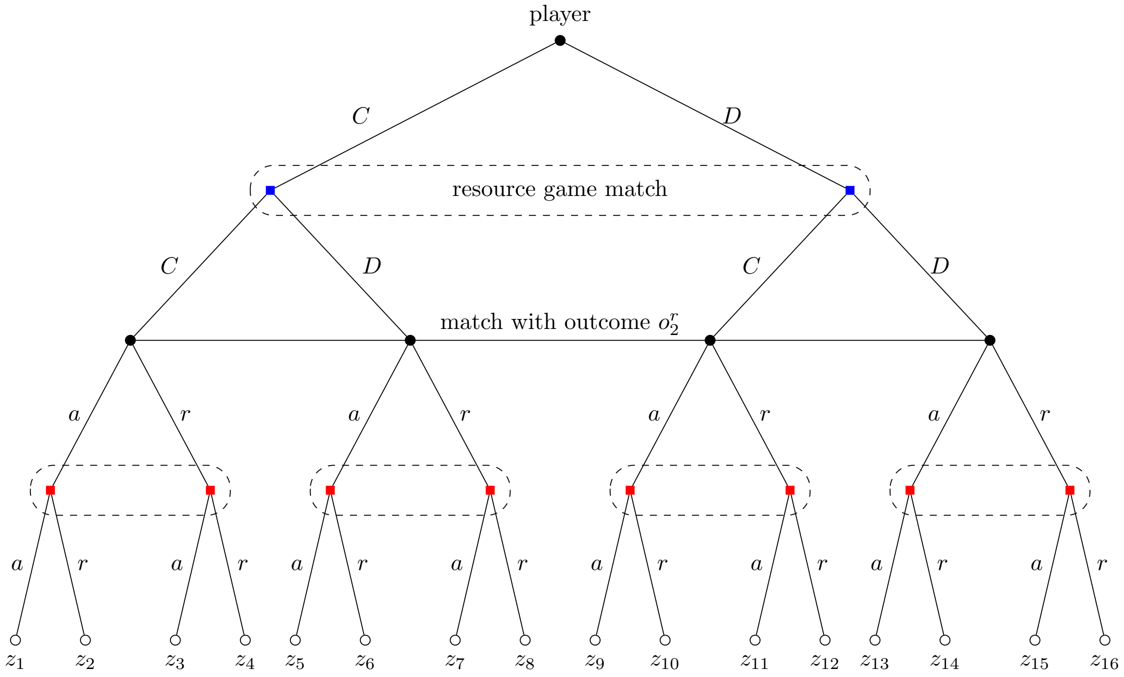

Figure 1: Partial game tree showing all information sets of a player. The player first chooses C or D in the resource game. He or she then finds a match in the offspring game and observes the match's outcome o_2^r. Players decide whether to to accept or reject their offspring game match conditional on their own and their match's outcome. The blue squares are nodes belonging to the player's match in the resource game and the red squares are the nodes belonging to the match in the offspring game. A type of a player is set of values associated to the terminal nodes of the tree.

means that there is no selective pressure for the invader types as there will be fewer of them, as proportion of the population, in the long run. The remaining invaders play the same actions as the incumbents on the equilibrium path. The difference in their preferences is only relevant when playing against preference types that do not exist in the long-run distribution, that is, preference types who are either not present among the invaders or whose preferences are selected against.

We ask whether altruistic preferences for cooperation can be sustained in a stable equilibrium of this game.

3 Analysis

We say that players *have preferences for offspring at outcome* o_1^R if their type θ is such that

$$\theta(o_1^R, o_2^R, (a,a), g) > \theta(\tilde{o}_1^R, o_2^R, (a,r), g) \tag{3}$$

for every gender $g \in \{f, m\}$, $\tilde{o}_1^R \in \Phi^R$ and $o_2^R \in \Phi^R$. That is, when a player's outcome in the resource acquisition game is o_1^R, the payoff when accepted by the potential partner exceeds the payoff when rejected.

We say that players *have a preference for partners who play action* $x \in \{C, D\}$ if their type θ is such that

$$\theta(o_1^R, (x, s_2), (a,a), g) > \theta(o_1^R, (x, s_2), (r,a), g) \text{ and } \theta(o_1^R, (x', s_2), (a,a), g) < \theta(o_1^R, (x', s_2), (r,a), g) \tag{4}$$

for every $x' \neq x$, $s_2 \in \{C, D\}$, $o_1^R \in \Phi^R$ and gender g. That is, the player accepts a potential partner who plays actions x but rejects a player who plays any other action in the resource acquisition game.

Let b^c be the strategy such that $(b^c)^R = C$, $(b^c)^O \left(o_1^R, (C, s_2)\right) = a$ and $(b^c)^O \left(o_1^R, (D, s_2)\right) = r$ for every $s_2 \in \{C, D\}$ and $o_1^R \in \Phi^R$. b^c is the behavioral strategy such that players choose C and only accept partners who played C in the resource acquisition game.

The following Theorem shows that cooperation is the unique stable if all preference types prefer offspring and prefer partners who play C.

Proposition 1. *If α is a distribution in which all players have a preference for offspring at (C, C) and also a preference for partners who play C then (α, b^c) is the unique stable equilibrium configuration.*

To understand this result note that the players who defect have low fitness in a population that only accepts players who cooperate. As they are rejected by almost all their potential mates they have very few offspring.

An invasion of players who behave differently depending on their gender does not have a relative advantage either. Consider a small invasion of players who defect when of gender g and cooperate when of gender $-g$. Suppose that both genders accept all potential partners in the offspring game. Because the invaders are few relative to the general population the players who defect have a small probability of being accepted in the offspring game. The players who accept defectors and play C have more offspring when matched to defectors. They also have zero probability of remaining unmatched while the players who play C and only accept C players remain unmatched with positive probability. However, half of their children will be of gender g, defect and face rejection from the incumbent population. Theorem 1 shows that for a small enough invasion the invader's disadvantage from having the gender g children rejected dominates the advantage of having more offspring when matched to a player of their own type.

Finally, an invasion of players who accept defectors but choose cooperation does not have a relative advantage. The fitness of the invaders is equal to the fitness of the incumbent population in the absence of defecting players.

Proof. Let's see that the set $b^c \in B(\alpha)$. First, all types in the population have a preference for partners who play C and, therefore, they reject players who play D. Due to the players' preference for offspring it is a best response for all types of players is to choose C if the opponents choose C.

Consider and invader type θ and let $\alpha_0 = \alpha(1-\widehat{\delta}) + \widehat{\delta}\theta$. Let's see that for any preference type θ there is an evolutionary sequence $\{(\alpha_k, b_k)\}_{k=0}^{\infty}$ with limit (α^*, b^*) such that conditions 1 and 2 in Definition 2 hold.

Step 1. Any type $\tilde{\theta}$ that has preference for offspring at (C,C) would choose C in the resource acquisition game for $\widehat{\delta}$ small enough if incumbent types continue to play C after the initial invasion.

Let δ^g denote the proportion of the gender g population with preference $\tilde{\theta}$ (in this case $\delta^g = \delta^{-g} = \frac{\widehat{\delta}}{2}$). The payoff of type $\tilde{\theta}$ of gender g from cooperation is at least

$$\tilde{\theta}((C,C),(C,C),(a,a),g)(1-\delta^{-g})^2(1-\delta^g),$$

as with probability $(1-\delta^{-g})(1-\delta^g)$ type $\tilde{\theta}$ matches with incumbents in both stages of the fitness game and with probability $(1-\delta^{-g})$ type $\tilde{\theta}$'s partner matched with an incumbent in the resource game.

Now, with probability $(1-\delta^{-g})^2$ the player finds incumbent matches in both stages and with probability at least $\frac{\widehat{\delta}}{2} = \max\{\delta^{-g}, \delta^g\}$ finds a match among the invaders in either of the two stages, the payoff from defection is at most

$$\tilde{\theta}((D,C),(C,C),(r,a),g)(1-\delta^{-g})^2+\frac{\widehat{\delta}}{2}\max\left\{\tilde{\theta}((D,s_2),o_2^R,j,g)|(s_2,o_2^R,j)\in\{C,D\}\times\Phi^R\times\Phi^O\right\}.$$

Now, as $\tilde{\theta}((C,C),(C,C),(a,a),g) > \tilde{\theta}((D,C),(C,C),(r,a),g)$ there is δ such that if $\widehat{\delta}\leq\delta$ type $\tilde{\theta}$ chooses cooperation.[5] This shows that (α_0,b) where $b_{\tilde{\theta}}=b^c$ for each $\theta\in\text{supp}\alpha$ is an equilibrium configuration. That is, there is an equilibrium in which the incumbents continue to play according to b^c after the invasion.

Step 2. If the incumbents keep playing C then the percentage of the invader population decreases with each generation.

Given distribution of types α, let $p_{\theta_1,\theta_2,\theta_3}$ be the probability that a player meets a player of type θ_3 in the offspring game, that the potential partner met a type θ_2 in the resource game and that the player met a type θ_1 in the resource game. Define

$$p_{\theta,\theta,I}=\sum_{\tilde{\theta}\in\text{supp }\alpha}p_{\theta,\theta,\tilde{\theta}}.$$

$p_{\theta,\theta,I}$ is the probability with which a player meets an incumbent in the offspring game, and the player and the player's potential partner met a type θ player in the resource game. The probabilities $p_{J,K,L}$ with $J,K,L\in\{\theta,I\}$ is defined analogously. We have $p_{\theta,\theta,I}=p_{I,\theta,\theta}=p_{\theta,I,\theta}=\delta_0^2(1-\delta_0)$, $p_{\theta,\theta,\theta}=\delta_0^3$, $p_{\theta,I,I}=(1-\delta_0)^2\delta_0$ and $p_{I,I,I}=(1-\delta_0)^3$.

Case 1. If θ has a preference for offspring at (C,C) by Step 1 for small enough $\widehat{\delta}$ there is an evolutionary sequence $\{(\alpha_k,b_k)\}_{k=0}^{\infty}$ in which θ and all the incumbents types continue to play according to b^c in each generation. The relative proportion of all types, including $\tilde{\theta}$ remains constant and therefore the sequence has a limit. Let α^* denote its limit distribution. The relative proportion of θ stays constant at $\widehat{\delta}$ in each generation. Thus $1=\Pi(\theta|(\alpha_0,b_0),\alpha^*)\leq\Pi(\tilde{\theta}|(\alpha_0,b_0),\alpha^*)=1$ for every $\tilde{\theta}\in\text{supp}\alpha$.

Case 2. Suppose that the invaders of both genders do not have a preference for

[5] For this step we use the fact that the distribution α has support on finitely many types. This assumption could be relaxed.

offspring at (C, C). That is

$$\theta((D, C), (C, C), (r, a), g) \geq \theta((D, C), (C, C), (a, a), g)$$

for $g \in \{f, m\}$. If the invaders play C they obtain at most the same fitness as the incumbents. Suppose that the invaders are willing to play D. By the previous discussion we know there is an equilibrium in which the incumbents play C as long as the probability of encountering an invader is sufficiently small. If $\theta((D, C), o_2^R, (a, a), -g) < \theta((D, C), o_2^R, (a, r), -g)$ the g invaders obtain zero fitness as they are rejected by all $-g$ players and vanish after one generation.

Suppose that $\theta((D, C), o_2^R, (a, a), g) \geq \theta((D, C), o_2^R, (a, r), g)$ for $g \in \{f, m\}$. That is, the invaders are willing to accept players who defect. Let's compute the number of offspring that carry preference θ one generation after the invasion if incumbents play according to b^c after the invasion (which by Step 1 is part of an equilibrium configuration). Since only $-g$ invaders accept g invaders the number of offspring of a type θ of gender g who inherit preference type θ is denoted $\pi^{D,D}(\theta, g)$ and is given by

$$\pi^{D,D}(\theta) = \pi^{D,D}(\theta, g) = \pi^{D,D}(\theta, -g) = \frac{1}{2}\left(2ap_{I,I,\theta} + (d+a)p_{\theta,I,\theta} + 2dp_{\theta,\theta,\theta}\right) \quad (5)$$

The number of offspring carrying the preferences of an incumbent player is

$$\pi_1^I = \frac{1}{2}\left(2cp_{I,I,I} + cp_{\theta,I,I} + cp_{I,\theta,I}\right) \quad (6)$$

The proportion of type θ players after one generation is

$$f^{D,D}(\theta) = \frac{2\delta_0 \pi^{D,D}(\theta)}{2\delta_0 \pi^{D,D}(\theta) + 2(1-\delta_0)\pi_1^I}.$$

For small enough δ_0 we have $f(\theta) \leq \frac{1}{1+\frac{(1-\delta_0)^2}{\delta_0^2}\frac{c}{a}} \leq \frac{\delta_0^2}{(1-\delta_0)^2}\frac{a}{c} < \frac{1}{2}\delta_0$.[6] Thus, from Step 1 there is an equilibrium in which incumbent players behave according to b^c after the first generation. By the same argument, the proportion of invaders is less than

[6] The first inequality holds if $\delta_0 \leq 1 - \delta_0$ and the last inequality holds if $\frac{\delta_0}{(1-\delta_0)^2} < \frac{1}{2}$.

or equal than $\frac{\delta_0}{4}$ and under distribution α_2 and there is an equilibrium in which the incumbents behave according to b^c. Arguing recursively, we obtain that the fitness of type θ is $\Pi(\theta|(\alpha_0, b_0), \alpha^*) = 0$.

Case 3. Suppose that

$$\theta((D,C),(C,C),(r,a),g) \geq \theta((D,C),(C,C),(a,a),g)$$

for some $g \in \{f, m\}$, while type θ of gender $-g$ has a preference for offspring at (C,C). The number of offspring of type θ of a type θ player of gender g is $\pi^{D,C}(\theta, g) = \pi^{D,D}(\theta)$, where $\pi^{D,D}(\theta)$ is defined in equation (5). The number of offspring of a type of gender $-g$ is given by

$$\begin{aligned} \pi^{D,C}(\theta,-g) &= \frac{1}{2}\left((c+a)p_{I,I,\theta} + (c+a)p_{\theta,I,\theta} + (c+d)p_{\theta,\theta,\theta} + 2cp_{I,I,I} + 2cp_{\theta,I,I} + cp_{I,\theta,I}\right) \\ &= \pi_1^I + \tilde{\pi}^{D,C}(\theta), \end{aligned} \tag{7}$$

where $\tilde{\pi}^{D,C}(\theta) = \frac{1}{2}\left((c+a)p_{I,I,\theta} + (c+a)p_{\theta,I,\theta} + (c+d)p_{\theta,\theta,\theta}\right)$ and $\tilde{\pi}_1^I = 2cp_{I,I,I} + 2cp_{\theta,I,I} + cp_{I,\theta,I}$.

The number of offspring that carry the preferences of an incumbent of gender g is given by

$$\pi_1^I(g) = \frac{1}{2}\left(2cp_{I,I,I} + 2cp_{I,I,\theta} + cp_{\theta,I,I} + cp_{I,\theta,I} + cp_{\theta,I,\theta} + cp_{I,\theta,\theta}\right). \tag{8}$$

The number of offspring that carry the preferences of an incumbent of gender $-g$ is $\pi_1^I(-g) = \tilde{\pi}_1^I$.

The proportion of type θ players is

$$\begin{aligned} f^{D,C}(\theta) &= \frac{\delta_0(\tilde{\pi}_1^I + \pi^{D,D}(\theta) + \tilde{\pi}^{D,C}(\theta))}{\delta_0\pi^{D,D}(\theta) + \delta_0\tilde{\pi}^{D,C}(\theta) + (1-\delta_0)\left(\pi_1^I(g) + \pi_1^I(-g)\right) + \delta_0\tilde{\pi}_1^I} \\ &\leq \frac{\delta_0(\pi_1^I + 2\pi^{D,D}(\theta))}{2\delta_0\pi^{D,D}(\theta) + (1-\delta_0)\left(\pi_1^I(g) + \pi_1^I(-g)\right) + \delta_0\tilde{\pi}_1^I} \leq \frac{1}{1 + \frac{2(1-\delta_0)\pi_1^I}{\delta_0\pi_1^I + 2\pi^{D,D}(\theta)\delta_0}}. \end{aligned}$$

where the first inequality comes from $\tilde{\pi}^{D,C} \leq \pi^{D,D}$ and the second inequality from $\pi_1^I(g), \pi_1^I(-g) \leq \tilde{\pi}_1^I$.

Now, for small enough δ_0 we have

$$\frac{\delta_0\pi_1^I + 2\pi^{D,D}(\theta)\delta_0}{2(1-\delta_0)\pi_1^I} = \frac{2\pi^{D,D}(\theta)\delta_0}{2(1-\delta_0)\pi_1^I} + \frac{\delta_0}{2(1-\delta_0)} \leq \frac{\delta_0^2}{(1-\delta_0)^2}\frac{a}{c} + \frac{\delta_0}{2(1-\delta_0)} \leq \frac{\delta_0}{4} + \frac{\delta_0}{2(1-\delta_0)},$$

where the first inequality is justified by $\pi^{D,D} \leq a\delta_0(1-\delta_0)^2$ and $\pi_1^I \leq c(1-\delta_0)^3$, and the last inequality holds if δ_0 is small enough so that $\frac{\delta_0}{(1-\delta_0)^2}\frac{a}{c} < \frac{1}{4}$. Replacing in the expression for $f^{D,C}(\theta)$ we obtain $f^{D,C}(\theta) \leq \frac{\delta_0(3-\delta_0)}{\delta_0(3-\delta_0)+4(1-\delta_0)}$. Note that as $\delta_0 \to 0$, $\frac{(3-\delta_0)}{\delta_0(3-\delta_0)+4(1-\delta_0)} \to \frac{3}{4}$. Thus, for δ_0 small enough $f^{D,C}(\theta) \leq \frac{3}{4}\delta_0$. As in Case 2, this implies that $\Pi(\theta|(\alpha_0, b_0), \alpha^*) = 0$. □

Theorem 1 shows that in order to sustain cooperation the players do not need to have a preference for cooperation in the resource acquisition game as in the observable or partially observable preferences case (see [2]). Cooperation is, instead, sustained by the judgement of players who do not conform to the social norm.

Corollary 1. *Let α be a distribution of types such that players prefer partners who cooperate and have a preference for offspring at outcome (C, C), then there is p^* such that for $p \geq p^*$ every distribution of the form $p\alpha + (1-p)\mu$ is stable and has a long run equilibrium configuration (α^*, b^*) with $b^* = b^c$. The threshold probability p^* is decreasing in $c - a$.*

Remarks

1. If a society is not composed exclusively of cooperators who like cooperators but the group of cooperators is isolated in some way such that they have probability of at least p^* of interacting with each other, from Theorem 1, cooperators would have greater fitness relative to non-cooperators.

2. In a society with social norms that sustain cooperation as in Theorem 1, invasions or mutations that involve higher fitness when cooperating have a relative advantage and take over the entire population in the long run. Thus, technologies that favor cooperation such as an increasingly complex language, will be favored by evolution. Mutations that increase the value of cooperation also reduce the threshold probability p^* and the threshold probability p^* that sustains cooperation is diminished. A drop in a also lowers

the threshold p^*. Mechanisms that punish deviators would lower the value of defection. It is arguably easier to coordinate on such a system of punishments when members of society have evolved to disapprove of defectors.

3. In our setting we assumed that the actions taken in the resource acquisition game are perfectly observed. A result analogous to Theorem 1 holds when players imperfectly observe the outcomes of the resource acquisition game as long as the signal is sufficiently informative. Cooperation can be sustained if players have a preference for players who show signals that indicate they cooperated with higher probability.

4 Summary of further results

The remainder of the paper considers settings in which the actions are imperfectly observed, in which players are allowed to condition on a signal of the family reputation (whether defection signals have been observed in a potential partner's nuclear family) as well as richer models of the offspring game in which players receive more than one option. In the latter settings players who do not reject defectors are at a relative disadvantage and after a small invasion the society returns to its original distribution.

References

[1] COLE, H. L., G. J. MAILATH, AND A. POSTLEWAITE (1992): "Social Norms, Savings Behavior, and Growth," *Journal of Political Economy*, 100, 1092–1125.

[2] DEKEL, E., J. C. ELY, AND O. YILANKAYA (2007): "Evolution of Preferences," *The Review of Economic Studies*, 74, 685–704.

[3] FISHER, R. A. (1915): "The evolution of sexual preference," *The Eugenics Review*, 7, 184–192.

[4] GRAFEN, A. (1990): "Sexual selection unhandicapped by the fisher process," *Journal of Theoretical Biology*, 144, 473–516.

[5] GÜTH, W. AND M. YAARI (1992): "An evolutionary approach to explain reciprocal behavior in a simple strategic game," *U. Witt. Explaining Process and Change–Approaches to Evolutionary Economics. Ann Arbor*, 23–34.

[6] HEIFETZ, A., C. SHANNON, AND Y. SPIEGEL (2007): "What to maximize if you must," *Journal of Economic Theory*, 133, 31–57.

[7] KIRKPATRICK, M. (1982): "Sexual Selection and the Evolution of Female Choice," *Evolution*, 36, 1–12.

[8] ROBSON, A. AND L. SAMUELSON (2011): "The evolution of decision and experienced utilities," *Theoretical Economics*, 6, 311–339.

[9] SETHI, R. AND E. SOMANATHAN (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

[10] TRIVERS, R. L. (1971): "The evolution of reciprocal altruism," *Quarterly review of biology*, 35–57.