

# The Evolution of Moral Codes of Behavior

Christoph Kuzmics\* and Carlos Rodriguez-Sickert†

June 19, 2006

## Abstract

Within a standard stochastic evolutionary framework, we study the evolution of morality, both at the level of behavior and at the level of codes of behavior. These moral codes involve sanctioning deviant agents. We provide conditions under which the presence of inter-group conflict allows the emergence of moral codes which improve social efficiency. The result depends on both the efficacy of the available sanctioning technology and on the ratio of the number of societies and the number of members which societies are composed of. We also consider the possibility that a moral code involves rewards rather than sanctions. We show that, in contrast with sanctioning moral codes, no system of rewards will evolve and therefore social efficiency cannot be improved via this mechanism.

**Keywords:** Stochastic Stability, Multi-level Selection, Evolution of Morality, Rule-guided Behavior.

**JEL CLASSIFICATION:** C62, C73, Z13.

---

\*MEDS, Kellogg School of Management, Northwestern University, c-kuzmics@kellogg.northwestern.edu.

†Department of Sociology, Universidad Catolica de Chile and Santa Fe Institute, crodrigs@puc.cl.

# 1 Introduction

Adam Smith, who is mainly known to economists for defending the virtues of self-love, describes human nature in the *Theory of Moral Sentiments* [1776] as possessing moral dispositions opposed to selfish behavior. In the Smithian model, the moral sense operates via the figure of the impartial spectator, who allows us to define what actions are proper or appropriate (from propriety in the work of Smith) and thus meritorious (worthy of reward) and which actions are not and, thus, are demeritorious (worthy of punishment). Morality, in this scheme, can be described at two different levels, i) at the level of rules of behavior against which a social response is observed; and ii) at the level of effective behavior, i.e., the actual degree of compliance to the set of prevalent norms in the community.

Here we develop an evolutionary approach to understand the emergence of morality in these two domains. The evolutionary, or biological, interest in the human sense of morality, can be traced back to Darwin himself. In his *The Descent of Man* [1871] he concludes that

[P]rimeval man, at a very remote period, was influenced by the praise and blame of his fellows. It is obvious, that the members of the same tribe would approve of conduct which appeared to them to be for the general good, and would reprobate that which appeared evil [Darwin, *The Descent of Man*, ch.5].

Since then, most of the work in evolutionary approaches to morality deal only with the evolution of moral behavior (generally understood as apparent or genuine self-sacrifice) and not with the evolution of rules of behavior. In his *The Biology of Moral Systems*, Alexander (1987) argues that beyond Hamilton (1964)'s world of kin-selection, moral systems are sustained by what he calls *indirect reciprocity*. Specifically, selection could favor strategies that involve self-sacrifice due to returned benefits in the future –associated with the individual's reputation and status in his community\*. In his framework, there is no space for altruistic behavior when interaction is non-repeated. Sober and Wilson (1994), among others have pointed to multi-level selection arguments to understand how genuine altruism could have evolved in our species.

The evolution of the rules themselves is not only important in itself to the extent that it defines one community's culture, but also in its impact on behavior. Experimental economists have shown that altruistic punishment (sanctioning free-riders) plays a crucial role in the achievement of social

---

\*Indirect reciprocity can be seen as an extension of the concept of reciprocal altruism developed by Trivers (1971) in the context of dyadic interaction.

cooperation in collective action problems (See Ostrom, Walker & Gardner (1992) for evidence in the context of common pool resources and Fehr & Gächter (2000, 2002 for evidence in the context of public good games). Falk et al (2005), who experimentally investigate the driving forces behind informal sanctions, conclude that the sanctioning of non-cooperators is grounded in the desire to harm those who committed unfair acts (rather than a more general motive for action, e.g., inequity aversion). The conditions under which altruistic punishment can evolve in collective action problems have been studied by Sethi & Somanathan (1996, 2004), Henrich and Boyd (2001) and Boyd, Gintis, Bowles and Richerson (2003). They suggest that punishment allows cooperation to emerge when altruism alone cannot.

Rather than studying a specific setting, here we seek to specify general conditions, albeit in a specific model of evolution, under which a moral code can evolve. Moreover, we attempt to typify the structural properties of the feasible moral codes which could emerge once these conditions are satisfied. For this purpose, we posit a model which accounts for the aforementioned dual character of morality in which an agent's strategy is composed of two components: (i) a moral code, which specifies the set of actions which will be disapproved and sanctioned (therein its observability) and by (ii) the action itself, which she will choose to implement<sup>†</sup>. Then, we configure a dynamic process where (i) agents learn to play more successful strategies within the community where they belong; (ii) agents' behavior has a stochastic component (associated with either random experimentation or simple mistakes); and (iii) there exists between-group competition.

Moral codes will emerge only under specific circumstances. There are two necessary conditions for a moral code to emerge. In the first place, the existence of multiple societies that compete among each other is required for a moral code to survive. Second, a moral code will emerge only within spheres of social action in which a member of the society can implement an action which gives him a payoff-edge over those who act in accordance with the societal norm. In such cases, whether a moral code will evolve and social efficiency improved will depend on two factors: i) the efficacy of the sanctioning technology; and ii) the ratio of the number of societies which can enter into conflict with respect to the number of members which these societies are composed of. When no moral code can sustain the societal optimal, it may well be the case that a moral code allows the society to

---

<sup>†</sup>It is important to notice that in our model it is assumed that the choice of the moral code and the action the agent implements are independent decisions. This approach differs from higher-order preferences models (e.g., Schelling, 1984) which account for the dual character of morality as a struggle within the individual. Under our model, the higher order "preferences" embodied in the moral code of the agents only determine social sanctions towards deviants.

achieve an intermediate outcome which strengthens the society with respect to the no moral code scenario. We also show that moral codes are not redundant. No moral code will sustain an outcome which is an equilibrium in the absence of the moral code. Likewise, no moral code will sanction actions which are harmful for the individual who implements such action, even if that action weakens the society as a whole. We also consider the possibility that rather than sanctioning demeritorious actions a moral code can reward meritorious actions. Against the intuition that similar achievements could be associated with such mechanism, we show that no rewarding moral code can evolve.

In the final section, we discuss a number of additional issues. Higher order morals (moral codes over moral codes); religion as an inducement to the evolution of morality; the interaction of punishing and rewarding moral codes; the effect of having non-observable moral codes; and the evolvability of asymmetric outcomes via hierarchical structures. We also relate our approach to the literature on group selection, cheap talk and the evolution of moral preferences.

## 2 The Game

Consider a society of a given number of individuals. Interaction between members of this society is modelled as a 2-player normal form game. This we call the basic society game. The understanding is that this game will be played frequently over time by randomly paired individuals from the society. In this section we will focus on a single society and define and analyze morality in this context.

Let the basic society game be denoted by  $\Gamma = (I, S, u)$ , where  $I = \{1, 2\}$  is the set of players,  $S$  is the finite set of pure actions available to both players with  $|S| \geq 2$ , and  $u : S \times S \rightarrow \mathbb{R}^2$  is the payoff function. The game is assumed to be symmetric, i.e. the payoff to a player depends only on the actions chosen and not on the role of the player, i.e. whether she is a row or a column player. In other words  $u(s, t) = u_1(s, t) = u_2(t, s)$  for all  $s, t \in S$ , if  $u_1$  and  $u_2$  were to denote the row and the column player's payoff function, respectively.

Every individual can hold a different moral code. A moral code is defined to be a set  $M \subset S$  of pure actions. The interpretation of this set  $M$  is that the holder of this moral code  $M$  regards actions in  $M$  as morally reprehensible, as "demeritorious", and will indeed reprehend, or punish, opponents who play actions in  $M$ . The set of all moral codes is the power set, the set of all subsets, of  $S$ , denoted by  $\mathcal{P}(S)$ .

Our model does not allow for the moral code to prescribe sanctions over

the choice of the moral code. It only operates over the actions. In the prisoners' dilemma (Table i) game for instance sanctions over cooperators who do not sanction defectors are not allowed. See section 7 for a discussion about relaxing this assumption.

We assume that once an instance of the basic society game  $\Gamma$  is played, an individual observes (at least with some positive probability) the realized pure action taken by her opponent. The individual will then act upon her moral code and punish actions in  $M$  if indeed one such action was played by her opponent. This leads to an extended society game  $\Gamma_e = (I, S_e, u_e)$ , presented in normal form, where  $I$  is as before, and where  $S_e = S \times \mathcal{P}(S)$ . An element  $s_e$  in  $S_e$  will be called a (pure) strategy. In order to define  $u_e$  nicely it is useful to introduce some notation. Let  $a : S_e \rightarrow S$  be such that for  $s_e = (s, M)$   $a(s_e) = s$ , i.e.  $a(s_e)$  is the action played in  $\Gamma$  induced by strategy  $s_e$  in  $\Gamma_e$ . Let  $P : S_e \rightarrow \mathcal{P}(S)$  be such that for  $s_e = (s, M)$  we have  $P(s_e) = M$ , i.e.  $P(s_e)$  is the moral code induced by  $s_e$ , the set of actions in  $\Gamma$  punished by a player playing strategy  $s_e$  in  $\Gamma_e$ . The payoff function  $u_e$  shall then be such that for any  $s_e, t_e \in S_e$  we have  $u_e(s_e, t_e) = u(a(s_e), a(t_e)) - \alpha \mathbf{1}_{\{a(s_e) \in P(t_e)\}} - \beta \mathbf{1}_{\{a(t_e) \in P(s_e)\}}$ . The interpretation is as follows. If the action played by a player, call him Bob, is in the moral code of the opponent, i.e. gets punished, then Bob suffers a payoff-loss of  $\alpha$ , assumed to be positive. If the action played by Bob's opponent is in Bob's moral code, i.e. Bob punishes it, then Bob suffers a payoff-loss of  $\beta$ , also assumed to be positive<sup>‡</sup>. Typically we would also assume that  $\alpha > \beta$  but this will not be necessary for our results. Note that holding any non-empty moral code is a weakly dominated strategy. It is dominated by the strategy which induces the same action coupled with the empty moral code. In the context of the underlying extensive form game, in which players choose an action first and a moral code only after they observe the opponent's action, holding a moral code is never subgame perfect.

We will abuse notation slightly. For  $s \in S$  let  $s$  also denote  $s = (s, \emptyset)$ , i.e. the strategy in which action  $s \in S$  is played and no action is punished. For  $x \in \Delta(S)$  let also  $x \in \Delta(S_e)$  such that  $x_{s_e} = x_{a(s_e)}$  for all  $s_e = (s, \emptyset)$  for some  $s \in S$ , and  $x_{s_e} = 0$  otherwise.

A few results are immediate.

**Lemma 1** *If  $x \in \Delta(S)$  is such that  $(x, x)$  is a Nash equilibrium in  $\Gamma$  then  $(x, x)$  is also a Nash equilibrium in  $\Gamma_e$ .*

---

<sup>‡</sup>Note that in contrast to the literature on the evolution of preferences (e.g. Ely and Yilankaya, 2001; Dekel, Ely, and Yilankaya, 2006; and Herold, 2006), where individuals are assumed to maximize their subjective payoff and only nature chooses subjective payoffs such as to maximize the individuals' material payoffs, in our model individuals do simply maximize their objective material payoffs.

Proof: Let  $s_e \in S_e$  be arbitrary. Then  $u_e(s_e, x) \leq u(a(s_e), x)$  by the assumption that punishing is costly, i.e.  $\beta > 0$ . In fact  $s_e$  is weakly dominated by  $a(s_e)$ . But  $a(s_e) \in S$  and, hence, by the fact that  $(x, x)$  is a Nash equilibrium in  $\Gamma$  we have  $u(a(s_e), x) \leq u(x, x)$ . Hence,  $u_e(s_e, x) \leq u(x, x)$  for any  $s_e \in S_e$ . QED

For  $s \in S$  let  $C(s) = \{s_e \in S_e | a(s_e) = s \wedge s \notin P(s_e)\}$  denote the set of all strategies in  $S_e$  in which action  $s$  is played and not punished.

**Lemma 2** *If  $s \in S$  is such that  $(s, s)$  is a (pure) Nash equilibrium in  $\Gamma$  then every  $x \in \Delta(C(s))$  is such that  $(x, x)$  is a Nash equilibrium in  $\Gamma_e$  with identical payoff.*

Proof: First note that  $u_e(x, x) = u(s, s)$  for all  $x \in \Delta(C(s))$  as action  $s$  goes unpunished by every pure strategy which receives positive probability in  $x$ . Let  $t_e \in S_e$  be arbitrary. Then  $u_e(t_e, x) = u(a(t_e), s) - \alpha \sum_{s_e \in C(s)} x_{s_e} \mathbf{1}_{\{a(t_e) \in P(s_e)\}} - \beta \sum_{s_e \in C(s)} x_{s_e} \mathbf{1}_{\{a(s_e) \in P(t_e)\}}$ . The two sums are both non-negative. Hence,  $u_e(t_e, x) \leq u(a(t_e), s) \leq u(s, s) = u_e(x, x)$  by the fact that  $(s, s)$  is a Nash equilibrium of  $\Gamma$ . But then  $u_e(t_e, x) \leq u_e(x, x)$ . QED

This Lemma states that every pure Nash equilibrium of  $\Gamma$  induces a whole component of (payoff-equivalent) Nash equilibria in  $\Gamma_e$ .

**Lemma 3 (No Hypocrisy)** *Let  $x_e \in \Delta(S_e)$  be such that there are two pure strategies  $s_e, t_e \in S_e$  such that  $(x_e)_{s_e} > 0$  and  $(x_e)_{t_e} > 0$ , and  $a(s_e) \in P(t_e)$ . Then  $(x_e, x_e)$  is not a Nash equilibrium in  $\Gamma_e$ .*

Proof: Suppose not. Suppose  $(x_e, x_e)$  is a Nash equilibrium. Then any  $s_e \in S_e$  such that  $(x_e)_{s_e} > 0$  must yield maximal payoff against  $x_e$ . Let  $s_e, t_e \in S_e$  be as in the statement of the Lemma. Let  $s'_e \in S_e$  be such that  $a(s'_e) = a(s_e)$  and  $P(s_e) = \emptyset$ . Then  $u_e(s_e, x_e) = u_e(s'_e, x_e) - \beta \sum_{t_e \in S_e} (x_e)_{t_e} \mathbf{1}_{\{a(t_e) \in P(s_e)\}}$ . By assumption this sum is strictly positive. Hence,  $u_e(s_e, x_e) < u_e(s'_e, x_e)$  and we arrive at a contradiction. QED

This lemma states that individuals in a symmetric Nash equilibrium cannot be hypocrites. They do not play an action which they themselves find morally reprehensible. This, somewhat pompously called, No Hypocrisy Lemma is a simple consequence of the general statement that there cannot be punishment (on the equilibrium path) in a Nash equilibrium. While this general statement is also true for asymmetric Nash equilibria, however, the No Hypocrisy Lemma is not. This will be discussed in section 7.6.

When moving from the basic to the extended society game we may create new symmetric Nash equilibria with a payoff not obtainable in any symmetric Nash equilibrium of the basic game.

To see this consider the following example, which we will come back to often throughout the paper. Let  $\Gamma$  be the Prisoner's Dilemma with payoffs given in Table i.

	C	D
C	3,3	0,4
D	4,0	1,1

Table i: The prisoners' dilemma.

The extended society game is given in Table ii. The punishment parameters are  $\alpha = 2$ , and  $\beta = 1$ .

Note that strategy  $(C, \{D\})$  constitutes a symmetric Nash equilibrium. This NE yields a payoff of 3, a payoff which is not attainable in the basic society game.

Note also that the No Hypocrisy Lemma (Lemma 3) does indeed not extend to asymmetric Nash equilibria. In the example strategy profile  $(C, \emptyset); (D, \{D\})$  constitutes an asymmetric Nash equilibrium in which player 2 uses an action which she herself finds morally reprehensible.

### 3 The dynamic model

There are  $n$  societies with  $m$  members each. Members play pure strategies in  $S_e$  only. Let  $\omega$  be a specification of which strategy every member in every society is using; i.e.  $\omega \in S_e^{m \times n}$  is a vector of strategies, one for each individual. Let  $\Omega$  denote the state-space, the set of all such specifications. For  $\omega \in \Omega$  let  $x^k(\omega)_{s_e}$  denote the proportion of members in society  $k$  who play pure strategy  $s_e \in S_e$ . The vector  $x^k(\omega) \in \Delta(S_e)$  can then be interpreted as a mixed strategy, i.e.  $x^k(\omega)_{s_e} \geq 0$  for all  $s_e \in S_e$  for all  $k$  and  $\sum_{s_e \in S_e} x^k(\omega)_{s_e} = 1$  for all  $k$ .

In the dynamic model, described below, we have occasional wars between societies. For a society with actual aggregate behavior  $x \in \Delta(S_e)$  let  $\sigma(x)$  denote the societies strength (ability to fight wars). In examples we will typically assume that  $\sigma(x) = u_e(x, x)$ , i.e. the strength is the average payoff (or GDP) in the society, but this is not necessary for the results. To make the analysis easier we will assume, however, that  $\sigma$  has maxima in  $\Delta(S_e)$ , say it is a continuous function, and that all strategies which maximize  $\sigma(x)$  are such that they induce the same pure action, which will be denoted by  $c \in S$ .

We will now define a stochastic evolutionary model which will give rise to an ergodic Markov chain on  $\Omega$ .

	$C, \emptyset$	$C, S$	$C, \{D\}$	$C, \{C\}$	$D, \emptyset$	$D, S$	$D, \{D\}$	$D, \{C\}$
$C, \emptyset$	3,3	1,2	3,3	1,2	0,4	-2,3	<span style="border: 1px solid black;">0,4</span>	-2,3
$C, S$	2,1	0,0	2,1	0,0	-1,2	-3,1	-1,2	-3,1
$C, \{D\}$	3,3	1,2	<span style="border: 1px solid black;">3,3</span>	1,2	-1,2	-3,1	-1,2	-3,1
$C, \{C\}$	2,1	0,0	2,1	0,0	0,4	-2,3	<span style="border: 1px solid black;">0,4</span>	-2,3
$D, \emptyset$	4,0	2,-1	2,-1	4,0	<span style="border: 1px solid black;">1,1</span>	-1,0	-1,0	<span style="border: 1px solid black;">1,1</span>
$D, S$	3,-2	1,-3	1,-3	3,-2	0,-1	-2,-2	-2,-2	0,-1
$D, \{D\}$	<span style="border: 1px solid black;">4,0</span>	2,-1	2,-1	<span style="border: 1px solid black;">4,0</span>	0,-1	-2,-2	-2,-2	0,-1
$D, \{C\}$	3,-2	1,-3	1,-3	3,-2	<span style="border: 1px solid black;">1,1</span>	-1,0	-1,0	<span style="border: 1px solid black;">1,1</span>

Table ii: The extended society game in normal form when the basic society game is the prisoners' dilemma (Table i); with  $\alpha = 2$  and  $\beta = 1$ . Pure Nash equilibria are indicated by boxes.

Time shall be discrete and between two time periods  $t$  and  $t+1$  the following four events may take place. First, every member observes all strategies<sup>§</sup> chosen at time  $t$ . Second, there is a dynamic, much like in Kandori, Mailath, and Rob (1993), within a society, which gradually pushes members to play better actions (against the behavior in their society). In particular, any member in any society has a probability  $\lambda > 0$  of revising his/her strategy. If a member revises his/her strategy the member is assumed to choose another strategy, which is a better reply against the current (time- $t$ ) aggregate behavior in his/her society than his/her previous strategy. The likelihood with which this new strategy is chosen is proportional to the payoff-difference between it and the previous strategy. This implies among other things that a member who is already playing a best reply will continue to play this strategy (even if there are alternative best replies). Note that we here assume that strategy choices, in particular moral codes, are observable to everyone. We discuss the consequences of relaxing this assumption in section 7.4.

Third, after this potential revision each member has a chance  $\mu$ , positive, but small, of changing his/her strategy to something arbitrary (according to a fixed probability distribution with full support).

Fourth and finally, with probability  $\tau > 0$  one randomly chosen society meets another randomly chosen society to fight a war. Only the stronger society (the society with higher  $\sigma(x^k)$ ) can win this war. The probability of winning is proportional to the difference in strength of the two societies. If there is no winner nothing changes. If there is a winner the defeated society is replaced by a copy (member for member) of the victorious society.

<sup>§</sup>For the case in which members only observe actions see section 7.



This assumption is intended to capture the presence of competition between societies. This is indeed the most crucial assumption in our model. We discuss the consequences of relaxing this assumption in section 7.7.

## 4 The evolutionary solution concept

Let  $\pi^\mu : \Omega \rightarrow \mathbb{R}_+$  denote the unique invariant distribution of the Markov chain outlined in the previous section. We will be interested in the unique limiting distribution  $\pi = \lim_{\mu \rightarrow 0} \pi^\mu$ . Following Young (1993) and Kandori, Mailath, and Rob (1993) we call a state  $\omega$  stochastically stable if  $\pi(\omega) > 0$ . A set of states  $\Psi \subset \Omega$  is stochastically stable if it satisfies  $\pi(\Psi) = 1$  and there is no proper subset of  $\Psi$  with the same property. By the finiteness of  $\Omega$  we know that such a stochastically stable set exists, it is obviously unique, and it encompasses all stochastically stable states. Let the stochastically stable set be denoted  $\Psi^*$ .

We are not so much interested in the stochastically stable set  $\Psi^*$  per se, but rather in the actions and moral codes used in  $\Psi^*$ . Let

$$S^* = \left\{ s \in S \mid \exists \omega \in \Psi^*, \exists s_e \in S_e \text{ with } a(s_e) = s, \exists k \in \{1, \dots, n\} : x^k(\omega)_{s_e} > 0 \right\}$$

denote the set of all actions used by at least one member in at least one society for at least one  $\omega$  in the stochastically stable set. Let

$$M^* = \left\{ s \in S \mid \forall \omega \in \Psi^*, \exists s_e \in S_e \text{ with } s \in P(s_e) : x^k(\omega)_{s_e} > 0 \forall k \right\}$$

denote the set of all actions which, if played by somebody, are punished by at least one person in every stochastically stable state. In other words  $S^*$  is the set of all actions which are played in **some** states in  $\Psi^*$ , while  $M^*$  is the set of all actions which are punished in **every** state in  $\Psi^*$ . We shall term  $S^*$  the evolutionary stable action set and  $M^*$  the essential evolutionary stable moral code. Essential, because it does not contain those actions which are punished only in some but not all the states in  $\Psi^*$ . It is, for instance, never in anybody's interest to use an action which is strictly dominated in  $\Gamma$ . Yet individuals may hold this action in their moral code. This, however, will not have any effect on the behavior of others simply because the action was not individually rational to begin with. So while evolutionary drift may introduce this action into individuals' moral codes, it may also eliminate it again from individuals' moral codes. In any case, the outcome of play is the same even if no-one holds this action in their moral code.

## 5 Results

The dynamics described in the previous section is, in the terminology of Hart and Mas-Colell (2005), uncoupled and exhibits 1-recall. Removing mistakes (setting  $\mu = 0$ ) from the model, their Theorem 1 implies that we will not, for general games  $\Gamma$  (and induced games  $\Gamma_e$ ), obtain almost sure convergence to a Nash equilibrium. Introducing  $\mu$  and taking the limit when  $\mu$  tends to zero does not help either. Under certain conditions, however, the game analyzed here has sufficient structure to enable us to make strong predictions.

The proof of the following lemma is immediate and omitted.

**Lemma 4** *Let  $\omega \in \Psi^*$ . Let  $\omega' \in \Omega$  be such that the transition from  $\omega$  to  $\omega'$  is possible with learning only, i.e. without any mistakes, i.e. has positive probability even in the limit when  $\mu \rightarrow 0$ . Then  $\omega' \in \Psi^*$ .*

The next lemma is Lemma 4 from Nöldeke and Samuelson (1993).

**Lemma 5** *Let  $\omega \in \Psi^*$ . Let  $\omega' \in \Omega$  be such that the transition from both  $\omega$  to  $\omega'$  as well as from  $\omega'$  to  $\omega$  is not possible with learning only, but both are possible with a single mistake and then learning only afterwards. Then  $\omega' \in \Psi^*$ .*

Let us first illustrate that no essential moral code need necessarily emerge, i.e.  $M^* = \emptyset$  is well possible under some circumstances. Consider the case of only one society. Recall that strategies which involve punishing are weakly dominated. One might think that evolution should eliminate such weakly dominated strategies in this case. Samuelson (1993, 1994), however, shows that evolution, modelled much like here, will not necessarily eliminate weakly dominated strategies. Samuelson (1994) even provides an example of a game and a reasonable evolutionary dynamic in which the only stochastically stable state is a dominated pure strategy profile. In the present context, weakly dominated strategies (i.e. strategies which involve punishing) are definitely going to be in the support of the limiting distribution. Proposition 1, however, demonstrates that for every strategy which involves punishing there must be at least 1 state in  $\Psi^*$  in which nobody uses it.

**Proposition 1** *Let  $n = 1$ . Then  $M^* = \emptyset$ .*

Proof: Suppose not. Then there must be an  $l \geq 1$ , where  $l$  denotes the minimum number of members who punish  $a$  over all  $\omega \in \Psi^*$ . Let  $\omega \in \Psi^*$  be such that this minimum is achieved, i.e. there are exactly  $l$  members who punish  $a$ . Suppose that not every member who punishes  $a$  plays a best response to their belief in  $\omega$ . Then with positive probability exactly one

such member will choose a different strategy  $s_e$  (while no-one else changes theirs). As punishing  $a$  is weakly dominated, there is a positive probability that in the new chose strategy  $a$  is not punished. But then in the new state (which by Lemma 4 is in  $\Psi^*$ ) the number of members who punish  $a$  is less than  $l$ , a contradiction. Hence, everyone who punishes  $a$  must already play a best-response to  $\omega$ . Now suppose that every other member who does not already play a best response switches to one. With positive probability these switches will be to a strategy which does not involve punishing  $a$  and, in addition, induces the same action as one of those members who punish  $a$ . Now in this new state (which by Lemma 4 must be in  $\Psi^*$ ) we either have that everyone plays a best reply already or that some member who does not punish  $a$  does not play a best-reply. The latter case would lead to a contradiction just as before. If everyone now plays a best response then this new state  $\omega$  is stationary under the learning dynamics only. I.e. without mistakes no member will change strategy or belief. In this state now there are also exactly  $l$  members who punish  $a$ . But if now one member who punishes  $a$  switches to not punishing  $a$  by mistake we get to a new state which is again stationary without mistakes. Hence, it takes at least one mistake to get back to the original state with  $l$  members punishing  $a$ . But then, by Lemma 5, the state with only  $l - 1$  members punishing  $a$  is also in  $\Psi^*$ , which finally provides the conclusive contradiction. QED

The above proposition states that if there is only one society no essential moral code emerges.

Before we analyze the case of multiple societies a few preliminaries are in order. We will make the following additional assumptions.

**Assumption 1** *The set  $\operatorname{argmax}_{y \in \Delta(S_e)} \sigma(y)$  is non-empty and  $a(x) = c \in S$  for every  $x \in \operatorname{argmax}_{y \in \Delta(S_e)} \sigma(y)$ .*

Assumption 1 says that there is a unique (and pure) action,  $c \in S$  which maximizes the strength function  $\sigma$ .

**Assumption 2**  $\min_{s \in S} [\alpha - (u(s, c) - u(c, c))] > 0$ .

Assumption 2 guarantees that punishment ( $\alpha$ ) is severe enough such that the action-pair  $(c, c)$  can be sustained as a Nash equilibrium in  $\Gamma_e$ ; i.e. there is a symmetric Nash equilibrium  $(s_e, s_e)$  of  $\Gamma_e$  such that  $a(s_e) = c$ . In fact under assumption 2 there will typically be a whole component of Nash equilibria in which action  $c$  is played exclusively.

Let  $\Theta = \{x \in \Delta(S_e) | \forall s_e > 0 : a(s_e) = c \wedge c \notin P(s_e) \wedge u_e(c, x) \geq u_e(s, x) \forall s \in S\}$  denote the Nash equilibrium component in  $\Gamma_e$  induced by  $c \in S$ . The set  $\Theta$  is non-empty under Assumption 2.

Let  $\Psi = \{\omega \in \Omega | x^k(\omega) \in \Theta \forall k\}$ . We will provide sufficient conditions under which  $\Psi$  is the stochastically stable set, i.e.  $\Psi^* = \Psi$ .

**Lemma 6** *If  $\omega \in \Psi$  is stochastically stable then so is any  $\omega' \in \Psi$ .*

Proof: For any two  $\omega, \omega' \in \Psi$  there is a sequence of single mutations that allows the transition from  $\omega$  to  $\omega'$  and vice versa. Along this sequence learning alone does not allow any transitions. In fact  $\Psi$  is what Nöldeke and Samuelson (1993) in their Definition 2 term a single-mutation neighborhood. It is a maximal such set if  $n \geq 2$ . The result then follows from Lemma 5. QED

Now let us give sufficient conditions under which a particular essential moral code emerges.

**Proposition 2** *Let  $\Gamma$  be a basic society game. Let Assumptions 1 and 2 be satisfied. Then if  $n > m$  (the number of societies exceeds the number of members in any given society) the set  $\Psi$  is the stochastically stable set ( $\Psi^* = \Psi$ ).*

Proof: By Lemma 6 we know that if an  $\omega$  in  $\Psi$  is stochastically stable then so must all  $\omega$  in  $\Psi$ . The transition from  $\Psi$  to any other stationary set of states requires at least  $n$  mutations. In fact it requires one mutation in each society, otherwise wars will inevitably lead back to  $\Psi$ . The transition from any state not in  $\Psi$  to  $\Psi$  requires at most  $m$  mutations (e.g. every member in one society switching to  $(c, M^*)$ ). If  $n > m$  the mutation-counting argument of Young (1993) and Kandori, Mailath, and Rob (1993) implies the result. QED

**Proposition 3** *Let  $\Gamma$  be a basic society game. Let Assumptions 1 and 2 be satisfied. Then if  $n > m$  (the number of societies exceeds the number of members in any given society) the evolutionary stable action set is  $S^* = \{c\}$  and the essential moral code is  $M^* = \{s \in S | u(s, c) > u(c, c)\}$ .*

Proof: By proposition 2 we know that  $\Psi^* = \Psi$ . Note that for all  $\omega \in \Psi^*$ , for all  $k$ , for all  $s_e \in S_e$  such that  $x^k(\omega)_{s_e} > 0$  we have that  $a(s_e) = c$ . Hence,  $S^* = \{c\}$ . To see that  $M^* = \{s \in S | u(s, c) > u(c, c)\}$  requires two arguments. First, let  $s' \in S$  be such that  $u(s', c) \leq u(c, c)$ . We need to show that  $s' \notin M^*$ . Let  $\omega \in \Omega$  be such that  $x^k(\omega)_{s'_e} = 1$  for  $s'_e = (c, S \setminus \{c, s'\})$  and for all  $k$ . Hence,  $a(s'_e) = c$  and  $c \notin P(s'_e)$  and  $u_e(c, x^k) = u(c, c) \geq u(s', c) = u_e(s', x^k)$ . Also, by assumption 2, we have that  $u_e(c, x^k) = u(c, c) \geq u_e(s, x^k)$  for any  $s \neq s'$  ( $s \in S$ ). Hence,  $x^k(\omega) \in \Theta$  for all  $k$ ; hence,  $\omega \in \Psi^*$ ; hence,  $s' \notin M^*$ . Second, let  $s' \in S$  be such that  $u(s', c) > u(c, c)$ . Let  $\omega \in \Omega$  be such that, for some  $k$ ,  $x^k(\omega)_{s_e} = 0$  for all

$s_e \in S_e$  such that  $s' \in P(s_e)$ , i.e.  $s'$  is not punished by anybody in society  $k$ . But then  $u_e(c, x^k) \leq u(c, c) < u(s', c) = u_e(s', x^k)$ ; hence,  $x^k(\omega) \notin \Theta$ ; hence,  $\omega \notin \Psi^*$ ; hence,  $s' \in M^*$ . QED

Propositions 2 and 3 say the following. Suppose there is a unique strength maximizing action  $c \in S$ . Suppose punishing is sufficiently severe. Then if there are sufficiently many societies, compared to the number of members in each society, a non-empty essential moral code will emerge and the unique action played by everybody is the strength maximizing one. Note that not all strategies which are harmful to society (lowers the society's strength) must be punished. Only those must be punished which give the player a payoff-edge over those who play the societal optimum must be punished. So for example, drug-abuse, arguably harmful for society but also for the user, need not be punished in the evolutionary stable moral code, while drug-dealing, arguably harmful for society but beneficial for the dealer, must be punished by at least some members of the society.

Note that while the essential moral code is the same in all societies, individual moral codes can vary to a great extent. One thing one can say is that, according to the No Hypocrisy Lemma 3 nobody will find the societal strength-maximizing strategy morally reprehensible.

In general  $n > m$  is not necessary for  $\Psi$  to be the stochastically stable set. In fact it suffices that  $n > \kappa m$ , for some  $\kappa \in (0, 1)$  which can well be sufficiently smaller than 1. The question is how many individuals in one society (say society  $k$ ) have to make a mistake such that learning alone can then induce everybody to switch so that the new state is in  $\Psi$ . A more likely transition (than everybody in society  $k$  making a mistake) is the one in which a minimal fraction  $\kappa$  of individuals in one society switch to a strategy, denoted  $\tilde{s}_e = (\tilde{a}, \tilde{M})$  with moral code  $\tilde{M} = S \setminus \{c\}$  (punish everything other than  $c$ ) and with an action  $\tilde{a}$ , which is most favorable for  $c$ -strategists. Let  $\kappa$  then be the smallest  $\kappa'$  for which  $u_e(c, M^*) > u_e(s_e, x^k)$  for all  $s_e \in S_e : a(s_e) \neq c$  and for all  $x^k \in \Delta(S_e) : x_{\tilde{s}_e}^k \geq \kappa'$ . This  $\kappa$  will in general depend on the payoff matrix of the basic society game and the punishment parameters. Further below we use the prisoners' dilemma example to illustrate what  $\kappa$  can be like.

Suppose the conditions of propositions 2 and 3 are not met. This does not imply that no moral code need emerge. Suppose the societal optimum (in terms of strength) is too inviting to deviations (punishment is not severe enough). It may well be that there is another strategy which improves strength over the equilibria in the basic society game and is such that deviations are not too profitable. In this case a social norm will emerge which induces people to punish profitable deviations from this sub-optimal but relative to the basic game strength-improving action. It is clear, however, from Proposition 1, that if there are not sufficiently many societies no essential

moral code will emerge at all.

Another case where no essential moral emerges is given in the following

**Corollary 1** *Suppose that in addition to the assumptions of Propositions 2 and 3,  $(c, c)$  is a Nash equilibrium of the basic society game  $\Gamma$ . Then  $M^* = \emptyset$ .*

Proof: Follows immediately from  $M^* = \{s \in S | u(s, c) > u(c, c)\}$  in Proposition 3. QED

The prisoners' dilemma example from before may illuminate Propositions 2 and 3 better. The intuition comes across better if we restrict attention to a somewhat partially extended society game only. The game given in Table iii only has one additional strategy compared to the basic society game, the prisoners' dilemma (Table i). This strategy has action  $C$  and moral code  $\{D\}$ , i.e. a person using this strategy cooperates and punishes defectors. This is the game studied in Sethi and Somanathan (1996) for a single society.

	$C, \emptyset$	$D, \emptyset$	$C, \{D\}$
$C, \emptyset$	3,3	0,4	3,3
$D, \emptyset$	4,0	1,1	2,-1
$C, \{D\}$	3,3	-1,2	3,3

Table iii: The partially extended society game in normal form when the basic society game is the prisoners' dilemma (Table i).

This partially extended society game has two Nash-equilibrium components. One is a singleton, consisting of the state in which everyone chooses  $D$ , while the other is the set of states in which there are no  $D$ -types, and there are 50% or more  $(C, \{D\})$ -types. These two components are depicted in the strategy simplex in Figure 1 as a fat dot at  $D$  and a fat line from  $P = (C, \{D\})$  halfway to  $C$ , respectively.

Let us first consider the single society case ( $n = 1$ ) and let  $P = (C, \{D\})$ . The value for  $\kappa$  is here  $1/2$ . Any state  $\omega$  which induces a mixed strategy  $x = x(\omega)$  below the line where  $x_P = 1/2$  must lead to everyone playing  $D$  if there are no mutations or mistakes ( $\mu = 0$ ). On the other hand, from any state  $\omega$  above this  $x_P = 1/2$  line there is a positive probability, under the dynamics with  $\mu = 0$ , that the dynamics will lead to the Nash-equilibrium component in which there are no  $D$ -types. This positive probability is sufficient for the mutation counting argument to go through. In the area in the simplex, denoted  $C \succ P \succ D$ , this probability is even equal to 1.

Note that  $\kappa$ , i.e. the fraction of people in the society who need to make a mistake in order for the dynamic to lead to the cooperation NE component,

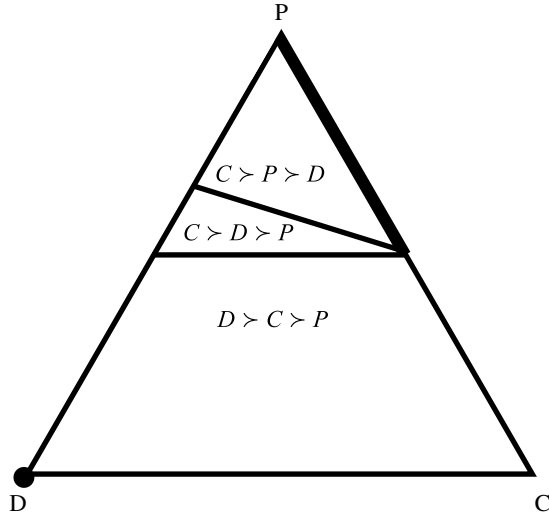


Figure 1: Strategy simplex for the partially extended society game given in Table iii, where  $P = (C, \{D\})$ .

does not depend on  $\beta$ , the cost of punishing to the punisher herself. A higher value for  $\beta$  would do two things. First it would reduce the area denoted by  $C \succ P \succ D$  plus it would reduce the positive probability of a transition from the area denoted  $C \succ D \succ P$  to the cooperation NE component. However small this positive probability is, though, the mutation argument is unaffected.

So, in this case it takes at most  $m/2$  mistakes to allow a transition with then positive conditional probability from everyone playing  $D$  to the other NE component, in which everyone cooperates, while at least half the population punishes defectors. On the other hand the easiest way to leave the cooperating NE component is from the state in which exactly half the population plays  $P$  while the other plays  $D$ . Now one mistake can induce the dynamic to lead to everyone playing  $D$ . Hence, only everyone playing  $D$  can be stochastically stable as claimed in Proposition 1.

Now for the case of multiple societies ( $n > 1$ ) let's assume the strength function is given by  $\sigma(x) = u_e(x, x)$ , which is here maximized at any  $x \in \Delta(S_e)$  which induces action  $C$  and does not punish  $C$ . The transition from everyone in every society playing  $D$  to every society playing in the cooperative NE component then still only requires  $m/2$  mistakes. This is so because once one society plays in the cooperative NE component it will have had wars with every other society much before there is another single mistake

in this society. Hence, once one society plays in the cooperative NE component very soon all societies will do so. On the other hand, a transition from the cooperative NE component in all societies to the state where everyone plays  $D$  now requires a simultaneous mistake in all  $n$  societies, inducing all societies to just fall of the edge to just fall below half the population playing  $P$ . Hence, which of the two NE components is stochastically stable now depends on whether the number of societies is greater or smaller than half the number of members in any given society.

## 6 Rewarding Moral Codes

In our setting, the moral code of an agent is described by the set of actions he disapproves of and is revealed by the set of actions he sanctions. By doing this we describe the agent's moral code in a negative way. Alternatively, the moral code could be defined by the set of (meritorious) actions which are rewarded by the agent who holds such moral code. If the set of meritorious and demeritorious actions would exhaust the action space (i.e., there are no neutral actions), these two options seem to be equivalent. However, this is not the case. First, evolutionary psychologists Cosmides and Tooby (1992) have shown that human beings are comparatively more skilled at identifying individuals who break social rules than at identifying those who adhere to the norm. Furthermore, in the experimental work of Gülerk, Irlenbusch and Rockenbach (2005), they report that in the context of collective action problems, the availability of sanctioning mechanisms has a stronger effect on social cooperation than the availability of rewarding mechanisms.

In the previous sections we defined a moral code,  $M$ , as the subset of actions an individual classifies as "demeritorious". A strategy not in the moral code  $M$  can then be interpreted as "meritorious". Now, while demeritorious strategies are punished, meritorious strategies are unaffected. In this section we want to do the opposite. Here meritorious strategies will be rewarded, while demeritorious strategies remain unchallenged. Starting from the basic society game  $\Gamma = (I, S, u)$ , as before, a moral code is again defined to be a set  $L \subset S$  of pure actions. The interpretation of this set  $L$ , however, is that the holder of this moral code  $L$  regards actions in  $L$  as morally sound, as "meritorious", and will indeed reward opponents who play actions in  $L$ . Again the set of all moral codes, defined this way, is the power set, the set of all subsets, of  $S$ , denoted by  $\mathcal{P}(S)$ .

This leads to a, what we shall term, **positively** extended society game  $\Gamma_e = (I, S_e, u_e)$ , presented in normal form, where  $I$  is as before, and where  $S_e = S \times \mathcal{P}(S)$ . An element  $s_e$  in  $S_e$  will be called a (pure) strategy. Let  $a : S_e \rightarrow S$  be such that for  $s_e = (s, L)$   $a(s_e) = s$ , i.e.  $a(s_e)$  is the action



played in  $\Gamma$  induced by strategy  $s_e$  in  $\Gamma_e$ . Let  $R : S_e \rightarrow \mathcal{P}(S)$  be such that for  $s_e = (s, L)$   $R(s_e) = L$ , i.e.  $R(s_e)$  is the moral code induced by  $s_e$ , the set of actions in  $\Gamma$  rewarded by a player playing strategy  $s_e$  in  $\Gamma_e$ . The payoff function  $u_e$  shall then be such that for any  $s_e, t_e \in S_e$   $u_e(s_e, t_e) = u(a(s_e), a(t_e)) + \alpha \mathbf{1}_{\{a(s_e) \in R(t_e)\}} - \beta \mathbf{1}_{\{a(t_e) \in R(s_e)\}}$ . The interpretation is as follows. If the action played by Bob is in the moral code of the opponent, i.e. gets rewarded, then Bob receives a payoff-gain of  $\alpha$ , assumed to be positive. If the action played by Bob's opponent is in Bob's moral code, i.e. Bob rewards it, then Bob suffers a payoff-loss of  $\beta$ , also assumed to be positive.

As in previous sections for  $s \in S$  let  $s$  also denote  $s = (s, \emptyset)$ , i.e. the strategy in which action  $s \in S$  is played and no action is rewarded. For  $x \in \Delta(S)$  let also  $x \in \Delta(S_e)$  such that  $x_{s_e} = x_{a(s_e)}$  for all  $s_e = (s, \emptyset)$  for some  $s \in S$ , and  $x_{s_e} = 0$  otherwise.

A few results are immediate.

**Lemma 7** *If  $x \in \Delta(S)$  is such that  $(x, x)$  is a Nash equilibrium in  $\Gamma$  then  $(x, x)$  is also a Nash equilibrium in  $\Gamma_e$ .*

Proof: Let  $s_e \in S_e$  be arbitrary. Then  $u_e(s_e, x) \leq u(a(s_e), x)$  by the assumption that rewarding is costly, i.e.  $\beta > 0$ . In fact  $s_e$  is weakly dominated by  $a(s_e)$ . But  $a(s_e) \in S$  and, hence, by the fact that  $(x, x)$  is a Nash equilibrium in  $\Gamma$  we have  $u(a(s_e), x) \leq u(x, x)$ . Hence,  $u_e(s_e, x) \leq u(x, x)$  for any  $s_e \in S_e$ . QED

**Lemma 8 (No Self-Congratulation)** *Let  $x_e \in \Delta(S_e)$  be such that there are two pure strategies  $s_e, t_e \in S_e$  such that  $(x_e)_{s_e} > 0$  and  $(x_e)_{t_e} > 0$ , and  $a(s_e) \in R(t_e)$ . Then  $(x_e, x_e)$  is not a Nash equilibrium in  $\Gamma_e$ .*

Proof: Suppose not. Suppose  $(x_e, x_e)$  is a Nash equilibrium. Then any  $s_e \in S_e$  such that  $(x_e)_{s_e} > 0$  must yield maximal payoff against  $x_e$ . Let  $s_e, t_e \in S_e$  be as in the statement of the Lemma. Let  $s'_e \in S_e$  be such that  $a(s'_e) = a(s_e)$  and  $R(s_e) = \emptyset$ . Then  $u_e(s_e, x_e) = u_e(s'_e, x_e) - \beta \sum_{t_e \in S_e} (x_e)_{t_e} \mathbf{1}_{\{a(t_e) \in R(s_e)\}}$ . By assumption this sum is strictly positive. Hence,  $u_e(s_e, x_e) < u_e(s'_e, x_e)$  and we arrive at a contradiction. QED

This lemma states that individuals in a symmetric Nash equilibrium do not reward an action which they themselves play. This is, as is the No Hypocrisy Lemma, an immediate consequence from the fact that in equilibrium there can be no punishment.

**Lemma 9** *Let  $(s_e, t_e)$  be a NE of  $\Gamma_e$ . Then  $(a(s_e), a(t_e))$  is a NE of  $\Gamma$  and  $a(t_e) \notin R(s_e)$  and  $a(s_e) \notin R(t_e)$ .*

Proof: Let  $(s_e, t_e)$  be a NE of  $\Gamma_e$ . First it is immediate that  $a(t_e) \notin R(s_e)$  and  $a(s_e) \notin R(t_e)$ , otherwise  $(a(s_e), \emptyset)$  would do better than  $s_e$  (and the same for  $t_e$ ). Now suppose  $a(s_e), a(t_e)$  is not a NE of  $\Gamma$ . Without loss of generality suppose there is an action  $s \in S$  such that  $u(s, a(t_e)) > u(a(s_e), a(t_e)) = u_e(s_e, t_e)$ . But  $u_e(s, t_e) \geq u(s, a(t_e))$  as  $R(s) = \emptyset$ . Hence,  $u_e(s_e, t_e) < u_e(s, t_e)$ . We thus arrive at a contradiction. QED

This lemma states that there are no new pure Nash equilibria in the positively extended society game. The idea is very straightforward. Suppose action  $s \in S$  is never a best-reply in  $\Gamma$ . To make it a best reply in  $\Gamma_e$  the opponent must reward  $s$ , which, given  $s$  is actually played, is not in the opponent's interest.

An example of a positively extended society game, based on the prisoners' dilemma in Table (i) is given in Table (iv).

	$C, \emptyset$	$C, S$	$C, \{D\}$	$C, \{C\}$	$D, \emptyset$	$D, S$	$D, \{D\}$	$D, \{C\}$
$C, \emptyset$	3,3	5,1	3,3	5,1	0,4	2,2	0,4	2,2
$C, S$	1,5	3,3	1,5	3,3	-2,6	0,4	-2,6	0,4
$C, \{D\}$	3,3	5,1	3,3	5,1	-2,6	0,4	-2,6	0,4
$C, \{C\}$	1,5	3,3	1,5	3,3	0,4	2,2	0,4	2,2
$D, \emptyset$	4,0	6,-2	6,-2	4,0	1,1	3,-1	3,-1	1,1
$D, S$	2,2	4,0	4,0	2,2	-1,3	1,1	1,1	-1,3
$D, \{D\}$	4,0	6,-2	6,-2	4,0	-1,3	1,1	1,1	-1,3
$D, \{C\}$	2,2	4,0	4,0	2,2	1,1	3,-1	3,-1	1,1

Table iv: The positively extended society game in normal form when the basic society game is the prisoners' dilemma (Table i) and rewarding costs and benefits are  $\alpha = \beta = 2$ . Pure Nash equilibria are indicated by boxes.

This game illustrates that while there are no new pure Nash equilibria in the positively extended society game, there are some new mixed Nash equilibria. Indeed something like Lemma 2 (in the punishment context) holds here too. It is true that, as in Lemma 2 in the punishment context, a pure Nash equilibrium in the basic society game induces a whole component of Nash equilibria in the positively extended society game. However, compared to Lemma 2 this component may not be as large. The game in Table iv illustrates this. The only pure NE is the  $(D, \emptyset); (D, \emptyset)$ , which follows from Lemmas 7 and 9. It comes with a component of symmetric Nash equilibria. Any  $\delta(D, \emptyset) + (1 - \delta)(D, \{C\})$  with  $\delta \geq 1/2$  constitutes a symmetric NE. If  $\delta < 1/2$  action  $D$ , however, is no longer a best-reply,  $C$  is.

Let  $\Psi^*$  be the stochastically stable set given the dynamic outlined before

when the game now is the positively extended society game  $\Gamma_e$ . Let

$$S^* = \left\{ s \in S \mid \exists \omega \in \Psi^*, \exists s_e \in S_e \text{ with } a(s_e) = s, \exists k \in \{1, \dots, n\} : x^k(\omega)_{s_e} > 0 \right\}$$

be, as before, the set of all actions used by at least one member in at least one society for at least one  $\omega$  in the stochastically stable set. Let

$$L^* = \left\{ s \in S \mid \forall \omega \in \Psi^*, \exists s_e \in S_e \text{ with } s \in R(s_e) : x^k(\omega)_{s_e} > 0 \forall k \right\}$$

denote the set of all actions which are rewarded by at least one person in every stochastically stable state.

**Proposition 4** *Let  $n \geq 1$ . Then  $L^* = \emptyset$ .*

Proof: Suppose not. I.e. suppose there is a  $s \in S$  such that for all  $\omega \in \Psi^*$  there is a  $s_e \in S_e$  such that  $s \in R(s_e)$  and  $x^k(\omega)_{s_e} > 0$ . Consider two cases. First, suppose there is an  $\omega \in \Psi^*$  such that action  $s$  is actually played by at least one person in society  $k$ . But then rewarding  $s$  is not best and, therefore, there is a positive probability that everyone in society  $k$  who rewards  $s$  will review his/her strategy to something that does not involve rewarding  $s$ . This new state, by Lemma 4, must then also be in  $\Psi^*$ , a contradiction to the supposition. Second, suppose  $s$  is not played by anyone in any  $\omega \in \Psi^*$ . Let  $\omega \in \Psi^*$  be the state with the minimal number of  $s$ -rewarders. But now rewarding  $s$  is payoff irrelevant and if one member who rewards  $s$  mutates to not rewarding  $s$  it would again take a single mutation to revert back as well. But then this new state, by Lemma 5, must also be in  $\Psi^*$ , which again yields a contradiction. QED

This result, that no essential (rewarding) moral code can evolve is consistent with our cognitive specialization on the identification of deviants rather than adherents to a particular norm as reported by Cosmides and Tooby (1992).

## 7 Discussion

In this final section we discuss some related literature and various extensions of the model.

### 7.1 Higher order morals

In the previous sections we ruled out higher order moral codes. I.e. person A cannot punish person B solely on the basis that person B does not punish certain types of behaviors. Punishments (or rewards) are only conditional on the chosen actions. One could, however, easily incorporate any higher

order moral codes. Consider the basic society game and the extended society game based, say, on punishment only. Now call this extended society game the basic society game. The induced new extended society game now allows for second order morals. However, by force of Corollary 1 and provided the initial basic society game and punishment technology satisfies the assumptions of Proposition 2, we immediately obtain that the essential second order moral will be empty. There may be evolutionary drift towards some second order moral codes sometimes, but these will not be essential to sustain the final evolutionary stable action set. If we assume that holding a second order moral code is just a little costly, then these second order moral codes will not appear at all. One interesting question which we have not tackled here is whether in the case in which the number of societies is too small (relative to the number of individuals in any given society) such that no essential moral code emerges higher order morals might yet evolve.

## 7.2 Religion as an inducement to the evolution of morality

It was Emile Durkheim, in his *Elementary Forms of Religious Life* [1912], the first author who called our attention to the collective secular utility a community unified by religion could achieve. More recently, Irons (1996) calls our attention to "...the role of religion in creating, sustaining and communicating motivations to behave in ways that are reciprocally altruistic, and thus serving to create large cooperative groups among unrelated individuals." (Irons 1996, p.51). In this section we consider the possibility that religion (or religious leaders) influence the emergence of moral codes.

Suppose that between any two dates  $t$  and  $t + 1$  in addition to three possible events described in Section 3 a fourth event can occur. With probability  $\nu$ , positive and potentially very small, in any given society a prophet appears and announces a moral code  $M \subset S$ . Prophets are rare but convincing. So every member in this society has a probability  $\rho$ , positive and fixed, of adopting this moral code  $M$ . This induces a new dynamic with unique invariant distribution  $\pi^{\mu, \nu} : \Omega \rightarrow \mathbb{R}_+$ . The limiting distribution  $\pi = \lim_{\mu, \nu \rightarrow 0} \pi^{\mu, \nu}$  now depends on the relative speeds of convergence to zero of  $\mu$  and  $\nu$ .

If  $\nu > \mu^k$  then religion, as modelled here, enhances the likelihood of a moral code to emerge. The inducement process modelled here is much in concordance with the one described in Wilson (2002) when he shows how John Calvin, without any formal authority, was able to introduce a new set of religious traits in the city of Geneva in the 1530s, affecting thus social behavior<sup>¶</sup>.

---

<sup>¶</sup>Bowles, Choi and Hopfensitz (2003), also consider mutations at the group level, but they take the form of an institutional change, specifically changes in a food-sharing rule.

### 7.3 Punishing and Rewarding Moral Codes

In this subsection we discuss the consequences of having potentially both rewarding and punishing behavior. While, again, rewarding behavior will not evolve in the long-run, it may allow an easier transition to the equilibrium punishing moral code, which will again be the same as in section 5. This implies that the possibility of rewarding behavior may lead to a non-trivial moral code, again the same as in section 5, even when there are only 2 societies.

Starting again from the basic society game  $\Gamma = (I, S, u)$ , a moral code is now a pair of sets  $M, L \subset S$  of pure actions. As before, actions in  $M$  are punished, while actions in  $L$  are rewarded. For convenience we will assume that  $M \cap L = \emptyset$ , i.e. no-one is allowed to punish and reward the same action. The set of all moral codes, denoted by  $\mathcal{M}(S)$ , is then the set of all such tri-partitions of  $S$ .

This leads to a, what we shall term, **completely** extended society game  $\Gamma_e = (I, S_e, u_e)$ , presented in normal form, where  $I$  is as before, and where  $S_e = S \times \mathcal{M}(S)$ . An element  $s_e$  in  $S_e$  will again be called a (pure) strategy. Let  $a : S_e \rightarrow S$  be such that for  $s_e = (s, M, L)$   $a(s_e) = s$ , i.e.  $a(s_e)$  is the action played in  $\Gamma$  induced by strategy  $s_e$  in  $\Gamma_e$ . Let  $P : S_e \rightarrow \mathcal{P}(S)$  be such that for  $s_e = (s, M, L)$   $P(s_e) = M$ , i.e.  $P(s_e)$  is the punishing moral code induced by  $s_e$ , the set of actions in  $\Gamma$  punished by a player playing strategy  $s_e$  in  $\Gamma_e$ . Let  $R : S_e \rightarrow \mathcal{P}(S)$  be such that for  $s_e = (s, M, L)$  we have  $R(s_e) = L$ , i.e.  $R(s_e)$  is the rewarding moral code induced by  $s_e$ , the set of actions in  $\Gamma$  rewarded by a player playing strategy  $s_e$  in  $\Gamma_e$ .

The payoff function  $u_e$  shall then be such that for any  $s_e, t_e \in S_e$

$$\begin{aligned} u_e(s_e, t_e) &= u(a(s_e), a(t_e)) \\ &- \alpha^P \mathbf{1}_{\{a(s_e) \in P(t_e)\}} - \beta^P \mathbf{1}_{\{a(t_e) \in P(s_e)\}} \\ &+ \alpha^R \mathbf{1}_{\{a(s_e) \in R(t_e)\}} - \beta^R \mathbf{1}_{\{a(t_e) \in R(s_e)\}}. \end{aligned} \tag{1}$$

The interpretation is the same as in the previous sections. The parameters  $\alpha^P$  and  $\beta^P$  are the punishment costs, denoted  $\alpha$  and  $\beta$  in section 2. The parameters  $\alpha^R$  and  $\beta^R$  are the costs and benefits associated to rewarding, i.e. the  $\alpha$  and  $\beta$  from section 6.

Versions of the intermediate results still hold in this model. In particular, the No Hypocrisy Lemma as well as the No Self Congratulation Lemmas still hold.

Consider the prisoners' dilemma, which will illustrate that the presence of potential rewards will allow an easier transition from the state where everyone defects in every society to the component of states in which the strength-maximizing action is played by everyone in every society. Let the parameter values be  $\alpha^R = \beta^R = 2$  and  $\alpha^P = 2, \beta^P = 1$ .

Fix society  $k$ . Consider the state in which everybody in every society plays  $(D, \emptyset, \emptyset)$ , i.e. everyone defects and there are no morals. This is a Nash equilibrium. Notice that the strategy  $(D, \emptyset, \{C\})$  is an alternative best reply to  $(D, \emptyset, \emptyset)$ . A person using this strategy defects but would reward cooperators. There could now be evolutionary drift towards more and more people using  $(D, \emptyset, \{C\})$ . This drift will go unchecked as long as the proportion of people playing  $(D, \emptyset, \emptyset)$  is more than a half. This is so because in this case all best replies involve playing  $D$  and not punishing  $D$ . So we could have a chain of single (and unchecked) mutations such that finally we arrive at a state in which a fraction  $\frac{1}{2} - \frac{1}{m}$  people in society  $k$  play  $(D, \emptyset, \emptyset)$  and the remaining fraction plays  $(D, \emptyset, \{C\})$ . But now, given the particular choices of the punishment and reward parameters, all best replies involve playing action  $C$ , not punishing  $D$ , and either rewarding or not rewarding  $C$ . There is now a positive probability that everybody in society  $k$  switches to play  $(C, \{C\}, \emptyset)$ . In this new state the best replies are now to play  $C$  and not reward  $C$  but either punish or not punish  $D$ . There is now a positive probability that everyone in society  $k$  switches to play  $(C, \emptyset, \{D\})$ , which again is a Nash equilibrium and selection (learning) forces alone cannot change the state. This means that there is a sequence of single mutations that takes play from everyone playing  $(D, \emptyset, \emptyset)$  to everyone playing  $(C, \emptyset, \{D\})$ .

If there is one society only then there is also a sequence of single mutations taking play the opposite direction. This implies that even if there is only one society we could see the strength maximizing action sometimes by force of Lemmas 5 and 4. The essential moral code, however, is still the empty set. In fact it is  $(\emptyset, \emptyset)$ .

If there are at least 2 societies, while there is still a sequence of single mutations to take play from everyone playing  $(D, \emptyset, \emptyset)$  to everyone playing  $(C, \emptyset, \{D\})$  (in one society, but then through the possibility of war, quickly also to all societies), it takes at least one mutation simultaneously in all societies (i.e. at least 2) to take play the opposite direction. Hence, by the usual mutation-counting arguments of Kandori, Mailath, and Rob (1993) and Young (1993), we have a non-empty moral code even if there are only 2 societies and it consists of the essential punishing moral code  $M^*$ , the same as in section 5, and the essential rewarding moral code  $L^* = \emptyset$ , the same as in section 6.

## 7.4 Non-observability of moral codes

In the previous sections we assumed that not only actions but also moral codes are observable to everyone. We believe that changing this assumption would not lead to significantly different results. Consider the following model in the spirit of Nöldeke and Samuelson (1993).

Each member not only plays an action but also holds a belief about which pure strategy everyone else in her society is playing. Now call  $\omega$  a specification of which strategy every member in every society is using and which belief she is holding. Let  $\Omega$  denote the state-space, the set of all such specifications. For  $\omega \in \Omega$  we can still define  $x^k(\omega)_{s_e}$ , the proportion of members in society  $k$  who play pure strategy  $s_e \in S_e$ . Just as before the vector  $x^k(\omega) \in \Delta(S_e)$  then looks like a mixed strategy. For  $\omega \in \Omega$  and for a given member  $l$  in society  $k$  we can now also define  $y_l^k(\omega) \in \Delta(S_e)$ , this members belief. I.e.  $y_l^k(\omega)_{s_e}$  denotes the proportion of members in society  $k$  who this member  $l$  believes to play pure strategy  $s_e \in S_e$ . Again  $y_l^k(\omega)$  then looks like a mixed strategy.

When members review their strategy they now first adapt their belief to the true behavior wherever possible. I.e. if action  $s \in S$  is played by at least 1 person in her society this member will know who exactly holds this action in their moral code. On the other hand, if an action  $s \in S$  is not played by anybody in her society, she does not know exactly who would potentially punish (or reward) this action, and, hence, she will stick to her old belief about who does and who does not hold this action in their moral code. Mutation is now assumed to be to an arbitrary action/belief combination with, again, full support.

This again leads to a well-defined Markov Chain, whose limiting distribution can be analyzed just as before.

The definition of the evolutionary stable action set,  $S^*$ , is just the same as before, but the definition of the essential moral code would have to be adapted. It is now not so important what moral codes are held in the society, but what members in the society believe these moral codes to be. This does no longer have to be the same. In the PD game, for example, we could well have the situation that in a given society, while everyone plays the societal optimum  $c$ , and no-one would actually punish any deviations from it, everyone in this society believes that a sufficient number of people would punish deviations. This sustains the outcome just as much as when people actually would punish. It still takes a mutation to take this society out of this outcome. This society, in fact, from an aggregate point of view, is playing a self-confirming equilibrium (Fudenberg and Levine, 1993). While the beliefs may be incorrect, they are at least not contradicted by the actual outcome.

The appropriate definition of an essential moral code is then the following.

$$M^* = \left\{ s \in S \mid \forall \omega \in \Psi^* \forall \text{ members } l \exists s_e \in S_e \text{ with } s \in P(s_e) : y_l^k(\omega)_{s_e} > 0 \right\}$$

In other words  $S^*$  is now the set of all actions which are **played** in **some**

states in  $\Psi^*$ , while  $M^*$  is the set of all actions which are **believed to be punished in every** state in  $\Psi^*$ .

We strongly believe that none of the results in the previous sections would be any different in this model. Consider the PD example and only punishing moral codes. In fact consider only the partially extended society game given in Table i. The stationary set of states for one society now consists of two components, the singleton Nash equilibrium  $(D, \emptyset; D, \emptyset)$  as before, and the set of all strategy profiles in which the action C is played and everyone believes that  $C, \{D\}$  is at least just as likely as  $C, \emptyset$ .

The transition from everyone defecting to the set in which everyone cooperates still takes  $m/2$  mutations (in one society). Wars will then lead to all societies playing in this set eventually with positive probability. The reverse transition is again achieved by a simultaneous mutation in all societies. So the reasoning remains completely unchanged.

## 7.5 Related Literature on Group-Selection, Evolution of Preferences, Cheap Talk, and the Secret Hand-Shake

The group-selection argument, as put forward by Wynne-Edwards (1962), very coarsely is the following. If something is good for the group (species, society) then it will evolve because of that. Maynard-Smith (1964) and others refuted this idea completely<sup>||</sup>. Their argument is that evolutionary selection is carried by the individual and not the group. Hence, we will only see behavior evolve which is good for the individual relative to whatever everyone else (in the group, species, society) is doing. This, in particular, implies that if the behavior which is good for the group is not in the interest of the individual this behavior cannot evolve. This, furthermore implies that if we see the societal optimum evolve it must be that this societal optimum is also in the interest of the individual.

In this paper we argue that if a punishing mechanism is available the evolution of the societal optimum may well be achieved. In fact we prove the following argument: If something is good for the group then it will evolve, not just because of the fact that it is good for the group, but also because there are (potentially only partially observable) strategies available which make this societal optimum a Nash equilibrium, and, hence, make this societal optimum such that no individual will want to deviate from it if the optimum is actually played. Note that in this Nash equilibrium we may not actually be able to observe those strategies which make the societal optimum a Nash equilibrium. This means that a biologist observing a particular

---

<sup>||</sup>See Bergstrom (2002) for an overview of the group-selection literature. See also Weibull and Salomonsson (2005) for another approach to modelling that allows group-selection



species or group of animals or plants may potentially be puzzled as to how a certain societal optimum is really in the interest of the individual given that off-equilibrium behavior is not observable. Similarly an anthropologist may also wonder why an individual behaves in the societal optimal fashion when it does not necessarily seem in this individual's interest. This is again due to the fact that the anthropologist only observes equilibrium behavior. In both cases it will be hard for the biologist as well as the anthropologist to comprehend the full game the respective individuals are playing.

That group-selection or similar models of "voting-with-ones-feet" may pick out the Pareto-optimal strict Nash equilibrium in coordination games (as opposed to the risk-dominant equilibrium) is demonstrated in Boyd and Richardson (1990), Oechssler (1997, 1999), Ely (2002), and Kuzmics (2003). Kuzmics (2003) also illustrates that group-selection does not pick out the dominated outcome of cooperation in the prisoners' dilemma. This is due to the fact that it is indeed never in the interest of the individual to cooperate and, hence, while group-selection might push towards cooperation, individual selection pushes against it. Only if group-selection is at least an order faster than individual selection can cooperation in the simple prisoners' dilemma occur (Sjöstrom and Weitzman, 1996).

Groups can also be seen as a message in a round of cheap-talk before the game, which allows individuals to coordinate deviations from the status quo. Evolution in these cheap-talk games was first investigated in Matsui (1991), and Kim and Sobel (1995). The reason why cheap-talk can lead to the Pareto-dominant outcome is that a formerly unused message can be seen as a secret hand-shake (Robson, 1990). If both parties use it they can coordinate to deviate from the status quo. If only one uses it nobody deviates.

This secret hand-shake idea is also central to the argument why individuals might evolve to have preferences, assumed observable to everyone, which will implement a Pareto-optimal outcome (Dekel, Ely, and Yilankaya, 2005).

The paper closest to ours within the evolution of preference literature is probably Herold (2005). Herold's (2005) model has a large population of individuals randomly allocated to groups. As long as they are in their group individuals play a game with other individuals in their group. This game is such that a proposer chooses one of 2 actions, to each of which a responder can react, with fitness-cost to her, by punishing or rewarding this action. Individuals can hold preferences different to material (fitness) preferences, which might induce them to rationally, given their preferences, punish or reward. When the groups are formed individuals within a group observe (or learn) the preferences of all other individuals in their group. After a while material (fitness) payoffs are accrued and groups dissolve only

to then again re-form randomly. Herold (2005) shows a few results which are different from but nicely related to ours. First, if individuals (when they are responders) only have the opportunity to reward, the only evolutionary stable state is where there is a mix of rewarders and non-rewarders. In contrast to this in our model we obtain that rewarding behavior will not evolve. Second, if individuals (when they are responders) only have the opportunity to punish, then there are two stable states, one in which nobody punishes and one in which everyone punishes, with obviously different implications for the proposer behavior. In our model, this corresponds somewhat to our finding that if there are not sufficiently many groups no punishment moral code can evolve, while if there sufficiently many a punishing moral code does evolve, again also with an impact on behavior. Third and finally, if individuals in Herold's (2005) model have the opportunity to both punish and reward, evolution can lead away from everyone "defecting" as rewarders can come in up to a point after which punishers will do better and we eventually end up in a state in which everyone punishes and proposers behave accordingly. The corresponding result in our model, for the prisoners' dilemma, is the possibility that the transition from everyone defecting to everyone cooperating and punishing defectors can be made easier by the presence of rewarding strategies. However, while in Herold's (2005) model this process is driven by selection, in our model it is driven by mutation.

## 7.6 Asymmetric groups

If members in a society have distinguishable features and can be identified as belonging to subgroups, such as men and women, evolutionary moral codes may potentially be asymmetric. Consider again the prisoners' dilemma as the basic society game and suppose the strength function is such that it is maximal at the action-profile  $C, D$ . In this case the essential evolutionary moral code can be different for different parts of the society. For example, evolution could lead to the following. Men defect and punish women who defect, but do not punish men who defect, while women cooperate and do not punish anybody. This means the essential evolutionary stable moral code for men is different than for women. I.e. we have the evolution of a double moral standard.

In fact, if the strength function is indeed maximized at an asymmetric strategy profile then one would expect evolution to find a way to allow individuals to form subgroups with distinguishable characteristics. One could easily have masters and slaves, various castes, etc., and different moral standards towards these different sub-groups of society.

## 7.7 About wars

In this section we discuss the fourth and perhaps most controversial assumption in our dynamic model from section 3. This fourth assumption has at least three parts. That wars are substantially more frequent than individual mutations; that in any war only the stronger society can win the war; and that if the stronger society wins the war, the weaker one is changed such that it resembles a copy of the stronger one (individual by individual). We will discuss these three assumptions (and the consequences of relaxing them) in some detail in this section.

There is a lot of anthropological and archeological evidence that warfare was and is common among past and present hunter-gatherer societies. For instance, Chagnon's (1968, 1997) anthropological study of the Yanomamö people of Brazil and Venezuela was originally subtitled "the fierce people" owing to the observation that groups of the Yanomamö engage in perpetual warfare. Another example is given by LeBlanc (1999) who demonstrates, based on archeological findings, that warfare was commonplace in prehistoric societies of the American southwest. The frequency of war can be estimated from Cioffi-Revilla's Long-Range Analysis of War (LORANOW) Project. Cioffi-Revilla (1996) reports a recorded 2300 wars in various geographic areas from as early as 3000 BC up until 1600 AD. He reports an "onset rate" of wars per year in any given geographic area of 0.2 on average and an average of 5 years between wars. If a period in our model denotes a generation, say 20 years, it seems safe to assume that  $\tau$ , the probability of war, is rather large (much larger than the mutation rate). It would be desirable to have data on inter-tribal conflict before 3000 BC, which is the time-period we consider our model to apply to more precisely. It is, however, rather difficult to find direct evidence of such proto-warfare from archeological findings. For Europe, Christensen (2004) demonstrates, though, that there is significant evidence of warfare even in the so-called Neolithic period (5500 BC - 2200 BC). Given modern anthropological studies (such as Chagnon, 1997) of "similar" hunter-gatherer societies, and assuming some sort of continuity of human behavior over time, we feel that our first assumption of  $\tau > 0$  (i.e. that the probability of war is significantly larger than the probability of an individual deviating from the norm) is somewhat justified.

Our second assumption, that only the stronger society can win the war, is very important for our results and very difficult to assess empirically. It is not enough to assume that the stronger society is just more likely to win the war. To be more precise suppose that the extended society game is based only on punishing moral codes (the model of section 5). If we then assume that any society can win any war with positive probability then no moral code would evolve. It is as if there was no between group competition. If we

add the possibility of rewarding moral codes as well as punishing ones the results may be different. In fact we conjecture the following to be true. If there are rewarding and punishing moral codes available, if stronger societies are more likely to win wars, and if there are many societies, a moral code would evolve and it would be the same as in section 5.

There is evidence that prehistoric warfare may well have been quite severe. See e.g. Milner et al. (1991) who, based on excavations of an Illinois burial site from about AD 1300, find that roughly a third of the adult population in west-central Illinois around this time died in wars or at least from violent causes. While this sort of evidence provides a partial justification for our third assumption that the losing society is replaced by a member-for-member copy of the winning society, this third assumption is, in any case, not necessary for the results. It would be sufficient to assume that the society which forms in place of the one which lost the war with some positive probability has a composition which, again with positive probability, can evolve to the composition of the winning society.

There are societies in which inter-tribal conflict is rare if not completely absent. See, for instance, Barnard (1992) for a study of the Bushmen of the Kalahari who do not seem to engage in warfare among "bands" at all. Yet, moral systems evolved there too. It is indeed not necessary to have wars for the evolution of moral codes. There are other forces of inter-group competition which give the same results. Consider the following model. Let  $\tau > 0$ , instead of denoting the probability of war, denote the probability of a prolonged drought, a famine, or just a bad year or period in which resources are more limited than usual. Strength  $\sigma$  now denotes the society's ability to extract resources (especially when they are in short supply). Survival of a society may now very well depend on its strength as it relates to other societies' strengths. If so, this model induces the same dynamics as the model with wars and will lead to the same conclusions.

## 8 Conclusion

In this paper we propose a model of inter-societal conflict which allows the evolution of both moral behavior and, more importantly, of moral codes or rules of behavior. These moral systems are not sustained by indirect reciprocity and not based on kin-selection. They evolve, if they do, even in societies in which people are not related with each other.

In this paper a moral code is understood as an individual's judgement as to what is good and what is bad. If an individual considers a certain action to be morally reprehensible, i.e. "bad", this individual punishes any individual using this action against him/her. This punishment is costly to

both the punisher as well as to the recipient of the punishment.

A non-trivial moral code can only evolve if there is at least some form of inter-group conflict which favors "stronger" societies (Proposition 1). A non-trivial moral code will definitely evolve if there are more societies than individuals in any given society (Propositions 2 and 3). We then proceed to demonstrate that a non-trivial moral code may well evolve even if there are significantly fewer societies than members in any given society. Religion, modelled as the rare occurrence of convincing prophets, can improve the chances of a moral code to evolve (section 7.2). While a moral code on the basis of rewarding cannot evolve, whatever the number of societies, the possibility of rewarding can make it easier for a moral code based on punishment to evolve (section 7.3).

Whenever a non-trivial moral code indeed evolves it typically has the following characteristics. First, all individuals in every society agree that the societal optimum (the unique action played in the evolutionary stable set of states) is a "good" action and will not be punished. Second, all societies "agree" that those action which give the individual a payoff-edge over the societal optimum must be punished by at least some individuals in the society. Third, there is a tremendous amount of potential heterogeneity among individual moral codes. Any given individual may well consider even those actions "good" which the society as a whole "agrees" to be "bad". Any given individual may well also consider those actions to be "bad" which the society as whole regards as "good", with the exception of the societal optimal action, which all individuals consider to be "good".

## References

- [1] ALEXANDER, R. (1987). *The Biology of Moral Systems*, New York, Aldine de Gruyter.
- [2] AXELROD, R. (1986). An Evolutionary Approach to Norms, *The American Political Science Review*, **80**, 4, 1095-111.
- [3] BARNARD, A. (1992). *Hunters and Herders of Southern Africa*, Cambridge University Press.
- [4] BERGSTROM, T. (2002). Evolution of Social Behavior: Individual and Group Selection, *Journal of Economic Perspectives*, **16**, 67-88.
- [5] BOEHM, C. (1982). The Evolutionary Development of Morality as an Effect of Dominance Behavior and Conflict Interference, *Journal of Social and Biological Structures*, **5**, 413-421.

- [6] BOWLES, S., J-K. CHOI AND A. HOPFENSITZ (2003). The Coevolution of Individual Behaviors and Social Institutions, *Journal of Theoretical Biology*, **223**, 135-47.
- [7] BOYD, R., H. GINTIS, S. BOWLES AND P. J. RICHERSON (2003). The Evolution of Altruistic Punishment, *Proceedings of the National Academy of Science (PNAS)*, **100**, 6, 3631-35.
- [8] BOYD, R. AND P. J. RICHERSON (1990). Group Selection among Alternative Evolutionarily Stable Strategies. *Journal of Theoretical Biology*, **145**, 331-42.
- [9] CANALS, J. AND VEGA-REDONDO, F. (1998). Multi-level evolution in population games. *International Journal of Game Theory*, **27**, 21-35.
- [10] CHAGNON, N. (1968). *Yanomamö: The Fierce People*, Holt, Rinehart and Winston.
- [11] CHAGNON, N. (1997). *Yanomamö*, 5th ed., Harcourt Brace College Publishers, Fort Worth.
- [12] CHRISTENSEN, J. (2004). Warfare in the European Neolithic. *Acta Archaeologica*, **75**, 128-56.
- [13] CIOFFI-REVILLA, C. (1996). Origins and evolution of war and politics. *International Studies Quarterly*, **40**, 1-22.
- [14] COSMIDES, L. AND J. TOOBY (1992). Cognitive Adaptations for Social Exchange. In Barkow, Cosmides and Tooby (eds.), *The Adapted Mind; Evolutionary Psychology and the Generation of Culture*. Oxford University Press.
- [15] DARWIN, C. (1981). *The Descent of Man*, Princeton University Press. Originally published in 1871.
- [16] DEKEL, E., ELY, J. C., AND YILANKAYA, O. (2006), Evolution of Preferences, forthcoming in *Review of Economic Studies*.
- [17] DURKHEIM, E. (1995). *The Elementary Forms of Religious Life*. Free Press, New York. Originally published in 1912.
- [18] ELY, J. C. (2002). Local conventions. *Advances in Theoretical Economics*, **2**, 1, 1.
- [19] ELY, J. C., & YILANKAYA, O. (2001). Nash Equilibrium and the Evolution of Preferences, *Journal of Economic Theory*, **97**, 255-272.

- [20] FALK, A., E. FEHR AND U. FISCHBACHER (2005). Driving Forces behind Informal Sanctions, *Econometrica*, **7**, 6, 2017-2030.
- [21] FEHR, E. & S. GACHTER (2000). Cooperation and Punishment in Public Good Games, *American Economic Review*, **90**, 4, 980-994.
- [22] FEHR, E. & S. GACHTER (2002). Altruistic Punishment in Humans, *Nature*, **415**, 10, 137-140.
- [23] FUDENBERG, D. & LEVINE, D. K. (1993). Self-confirming equilibrium, *Econometrica*, **61**, 523-45.
- [24] GINTIS, H. & S. BOWLES (2006). *A Cooperative Species: Human Sociality and its Evolution*, manuscript.
- [25] GURERK, O., B. IRLENBUSCH & B. ROCKENBACH (2005). The Competitive Advantage of Sanctioning Institutions, *mimeo*.
- [26] HAMILTON, W. D. (1964). The genetical evolution of social behavior, *Journal of Theoretical Biology*, **37**, 1-52.
- [27] HART, S. & A. MAS-COLELL (2005). Stochastic Uncoupled Dynamics and Nash Equilibrium, *mimeo*.
- [28] HENRICH, J. & R. BOYD (2001). Why people Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas, *Journal of Theoretical Biology* **208**, 79-89.
- [29] HEROLD, F. (2005). Carrot or Stick: The Evolution of Reciprocal Preferences in a Haystack Model, *mimeo*.
- [30] IRONS, W. (1996). In Our Own Self Image: The Evolution of Morality, Deception, and Religion, *Skeptic* **4**, 2, 50-61.
- [31] KANDORI, M., MAILATH, G. AND ROB, R. (1993). Learning, mutation, and long-run equilibria in games, *Econometrica*, **61**, 1, 29-56.
- [32] KIM, Y.-G., AND SOBEL, J. (1995). An evolutionary approach to pre-play communication, *Econometrica*, **63**, 1181-93.
- [33] KUZMICS, C. (2003). Individual and group selection in symmetric 2-player games, *mimeo*.
- [34] LEBLANC, S. A. (1999). *Prehistoric Warfare in the American Southwest*, University of Utah Press.
- [35] MATSUI, A. (1991). Cheap talk and cooperation in a society, *Journal of Economic Theory*, **54**, 245-58.

- [36] MILNER, G. R., ANDERSON, E., AND SMITH, V. G. (1991). Warfare in Late Prehistoric West-Central Illinois. *American Antiquity*, **56**, 4, 581-603.
- [37] NÖLDEKE, G. AND SAMUELSON, L. (1993). An Evolutionary Analysis of Backward and Forward Induction, *Games and Economic Behavior*, **5**, 425-54.
- [38] OECHSSLER, J. (1997). Decentralization and the Coordination Problem, *Journal of Economic Behavior and Organization*, **32**, 119-135.
- [39] OECHSSLER, J. (1999). Competition among conventions, *Computational & Mathematical Organization Theory*, **5**, 31-44.
- [40] OSTROM, E., J. WALKER & R. GARDNER (1992). Covenants With and Without a Sword: Self-Governance is Possible, *The American Political Science Review*, **86**, 2, 404-417.
- [41] ROBSON, A. J. (1990). Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, **144**, 379-96.
- [42] SAMUELSON, L. (1993). Does evolution eliminate dominated strategies?, in *Frontiers of Game Theory*, Eds. Binmore, K., Kirman, A., Tani, P., MIT Press, Cambridge, 213-35.
- [43] SAMUELSON, L. (1994). Stochastic Stability in Games with Alternative Best Replies, *Journal of Economic Theory*, **64**, 35-65.
- [44] SETHI, R. AND E. SOMANATHAN (1996). The Evolution of Social Norms in Common Property Resources, *American Economic Review* **86**, 4, 766-88.
- [45] SETHI, R. AND E. SOMANATHAN (2004). Collective Action in the Commons: A Theoretical Framework for Empirical Research, *mimeo*.
- [46] SMITH, A. (1976). *The Theory of Moral Sentiments*, Oxford University Press. Originally published in 1759.
- [47] SJOSTROM, T. AND WEITZMAN, M. L. (1996). Competition and the evolution of efficiency, *Journal of Economic Behavior and Organization*, **30**, 25-43.
- [48] SCHELLING, T. (1984). *Choice and Consequence*, Harvard University Press.



- [49] TRIVERS, R. L. (1971). The evolution of reciprocal altruism, *Quarterly Review of Biology*, **46**, 35-57.
- [50] YOUNG, H.P. (1993). The Evolution of Conventions, *Econometrica*, **61**, 1, 57-84.
- [51] WEIBULL, J. W. & SALOMONSSON, M. (2006). Natural selection and social preferences, *Journal of Theoretical Biology*, **239**, 79-92.
- [52] WILSON, D.S. (2002). *Darwin's Cathedral: Evolution, Religion and the Nature of Society*, University of Chicago Press.