Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 241 (2006) 233–240
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

The evolution of norms

F.A.C.C. Chalub[a,*], F.C. Santos[b], J.M. Pacheco[c]

[a]Centro de Matemática e Aplicações Fundamentais, Universidade de Lisboa, Av. Prof Gama Pinto 2, P-1649-003, Lisboa, Portugal
[b]IRIDIA, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, Belgium
[c]Centro de Física Teórica e Computacional and Departamento de Física da Faculdade de Ciências, P-1649-003 Lisboa Codex, Portugal

Received 29 July 2005; received in revised form 4 November 2005; accepted 22 November 2005
Available online 4 January 2006

Abstract

We develop a two-level selection model in the framework of evolutionary game theory, in which fitness selection at different levels is related to different games. We consider an archipelago of communities, such that selection operates at an individual level inside each community and at a group level whenever evolution of communities is at stake. We apply this model to the evolution of social norms, an open problem of ubiquitous importance in social science. Extensive statistical analysis of our results lead to the emergence of one common social norm, of which the evolutionary outcomes in different communities are simple by-products. This social norm induces reputation-based cooperative behavior, and reflects the evolutionary propensity to promote simple, unambiguous norms, in which forgiveness and repent are welcome, while punishment is implacable.

Keywords: Evolutionary game theory; Indirect reciprocity; Evolution of cooperation; Multi-level selection

1. Introduction

It is well known that biological entities typically compete and cooperate among themselves, being capable of building communities that also compete and cooperate among themselves. For example, genes within an organism tend to cooperate, although in some situations, e.g. meiotic drive (see Pomiankowski, 1999, and references therein) they become fierce competitors. As such, and independently of which is the *fundamental* unit of natural selection—the gene (Dawkins, 1976) or the individual (Gould, 1984; Sober and Lewontin, 1984)—it is clear that natural selection constitutes a multi-scale process acting through intricate networks of interactions under the combined effect of competition and cooperation. Indeed, a large body of work exists on multi-level selection, as reviewed, for instance, in Keller (1999).

The following example illustrates the problem. Consider a population of two different kinds of individuals, A and B, and suppose that the $A-A$ and $B-B$ interactions are much stronger than the $A-B$ interaction. Then, if individuals of both types play an iterated prisoner's dilemma (adequately parametrized), the strategy "cooperate with individuals of the same type, defect against the other type" will evolve. If, furthermore, there is any cohesion force (e.g. a kin selection mechanism, for instance resulting from genetic relatedness between A individuals and also between B individuals), then the A-population and the B-population may play as single units against each other. In other words, there will be two co-existing levels of selection, one at the individual level, the other at the group level.

In this work we develop a mathematical model incorporating two levels of selection, based on evolutionary game theory. This means that our *units* of selection interact among themselves and also organize in groups that interact with each other. Similarly to the example given, different issues will be at stake at each level. Therefore, they will be modeled using an appropriate (different) game, in contrast with other two-level models (Trausen et al., 2005; Paulsson, 2002). Group selection mechanisms for cultural evolution have been studied in Bowles et al. (2003), Bowles and Ginits (2004), Boyd and Richerson (1985, 1990), Henrich and Boyd (2001), and Boyd et al. (2003),

*Corresponding author. Tel.: +351 21 790 49 31; fax: +351 21 795 42 88.
E-mail addresses: chalub@cii.fc.ul.pt (F.A.C.C. Chalub), fsantos@ulb.ac.be (F.C. Santos), pacheco@cii.fc.ul.pt (J.M. Pacheco).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.11.028

whereas empirical data supporting group selection were examined in Soltis et al. (1995).

We shall apply our model to study the evolution of social norms of cooperation (in Kandori (1992) a game theoretical study of community enforcement of social norms has been carried out), an important concept in social science, where the existence of social norms is very often invoked, despite the fact that their origin and evolution remain big unsolved problems in this area. To this end we adapt to the present framework the seminal work recently developed in Brandt and Sigmund (2004, 2005), Ohtsuki and Iwasa (2004). Nonetheless, the framework proposed has a scope which extends well beyond the particular application to the evolution of social norms considered here.

More explicitly, in Brandt and Sigmund (2004, 2005) an "assessment module" (which we call a "norm") was introduced as a way of giving relative merit to the actions performed by the players, evaluating each action as "good" or "bad". From this definition, individual-based simulations compared the efficiency of some modules as promoters of cooperation. On the other hand, in Ohtsuki and Iwasa (2004) the same concept was called "reputation dynamics" and for all possible given "reputation dynamics" all possible Evolutionary Stable Strategies (ESS) were classified. Among these, eight were selected as the most efficient as promoters of high average payoff, and as such become crucial to the understanding of the evolution of cooperation ("the leading eight"). In this work we go further in this analysis and introduce a dynamics on the space of norms, allowing natural selection to choose, among all possible equilibra, the most robust, stable, and efficient. In particular, we associate the outcome of evolution at the lowest level of selection with group fitness.

2. The model

We consider an archipelago, each island being occupied by a group of individuals. These groups live in semi-isolation, i.e. most of the time individuals of a given island interact with their island's co-inhabitants. Let us define this process as "peace" time. From time to time, different islands interact. This we define as "war" time (although one should not take these names too seriously). During war-time, individual behavior becomes irrelevant, each island acting as a whole.

Each island has a social norm, which is shared by all its inhabitants. This social norm, or "moral", dictates how a certain action performed by an individual affects his/her reputation. On the other hand, individuals in the population behave according to their own strategies, which define how an individual acts when engaging with another individual in the social dilemma game described below. The fitness of each individual is associated with the total payoff accumulated as a result of his/her interactions with each co-inhabitant of the same island. Such fitness determines the probability to successfully pass the individual's strategy to the next generation. This constitutes the lowest level of selection.

Both norms and strategies change in time, but while strategies evolve under individual interactions within one island, at the lowest level of selection, the norms will evolve based on the performance of each island as a whole. This means that the time-scales associated with each mechanism are different, evolution of norms taking place at a slower rate than the evolution of individual strategies. For simplicity, we assume that social norms remain unchanged during peace time and possibly change as a result of confrontation between islands (war-time).

Since the actions of individuals depend on reputation, which in turn depends on the social norm, the average fitness (defined as the average payoff of the game in the lowest level) of an island achieved during peace time will depend on the social norm adopted by that island. We use the average fitness of an island, resulting from interactions between individuals from that island throughout peace time, as a feedback mechanism determining the evolution of the social norms. In this way, islands performing better during peace time will confer to their associated norms a fitness advantage, i.e. they are better equipped to win a conflict. During war time, after each conflict between islands, the defeated island will be forced to change its norm in the direction of the winner, if the victor island adopted aggressive strategies. In this case a small part of the population of the defeated island is eliminated and replaced by migrants from the victor island, which keep their strategies but change their norms to the one of their new homeland. In practice, this means that the individual absorption by the island norm is much faster than the typical norm changes (Soltis et al., 1995). In any case, norms are subject to possible mutations. We do not worry about the precise way in which norms and strategies are passed from generation to generation (cultural or genetic transmission).[1]

After a long succession of war-and-peace cycles, the archipelago will reach a stationary situation in which islands will adopt successful norms and individuals adopt successful strategies inside each island. The analysis of these norms and strategies will be the topic of next section. Now, we describe the two levels of selection and respective games in further detail.

2.1. Norms and individual behavior

Let us consider a given "island", and describe the evolution inside this island.

We consider a two-people reputation game (the "give-and-receive game", introduced in Nowak and Sigmund (1998a, b)), where in each interaction one of the players acts as *donor*, while the other acts as *recipient*. In each round, the donor should decide if (s)he shall provide (play

[1]Group selection seems, however, to be more adequate to study cultural evolution (Soltis et al., 1995).

GIVE) or not (play NOT GIVE) a certain help to the recipient. If he/she plays GIVE, then his/her own payoff is decreased by 1, while the recipient's payoff is increased by $b>1$. If he/she plays NOT GIVE, the individual payoffs remain unchanged (following common practice (Brandt and Sigmund, 2004; Nowak and Sigmund, 1998b; Leimar and Hammerstein, 2001), we increase the payoff of every interacting player by 1 in every round to avoid negative payoffs). This action will be witnessed by a third-party individual who, based on the island's social norm, will ascribe (subject to some small error probability $\mu_a \ll 1$) a new reputation to the donor, which we assume to spread efficiently without errors to the rest of the individuals in that island. This corresponds to the so-called "indirect observation model" defined in Ohtsuki and Iwasa (2004).

In Nowak and Sigmund (1998b), the idea of "image score" was introduced as a way to measure the reputation of each player. The score serves as an indication of how he/she behaved in the past, being higher for the ones who played GIVE and lower for those that systematically refused help. This idea seems too restrictive, as the score of a player will increase after playing GIVE irrespective of the score of the co-player. This rules out other possibilities, such as the central idea of punishment of bad players (Fehr and Fischbacher, 2004; Boyd et al., 2003).

A simple means to overcome this limitation was introduced in Brandt and Sigmund (2004) and Ohtsuki and Iwasa (2004). Each individual has a binary reputation—GOOD or BAD associated with the integers 0 and 1, respectively—which is ascribed to the individual based on his/her past behavior, evaluated according to the island's norm. The norm $\mathcal{N} = N(7)\cdots N(0)$, $N(i) = 0$ or 1, $i = 0,\ldots,7$, is an islands attribute, shared by all its inhabitants, and history of behavior resumes to the previous interaction of an individual, acting as a donor. As such, and because the norm must specify the reputation associated with all possible interaction scenarios within such a limited memory horizon, there will be a total of 256 possible norms, associated with a 8-bit binary digit encoding all possibilities, each associated with a possible combination of actions, as shown in Table 1.

As a result, each individual "is born" with a GOOD initial reputation and after each round where (s)he played as donor, his/her new reputation will be $N(i)$, with i matching the action taken and the reputations of the two individuals involved in that round. We could as well consider a ninth digit encoding the reputation at birth—initial condition. As extensive computations show, this is immaterial to the final result, and we will not consider it here.

Each individual adopts a well-defined strategy, corresponding to the "action module" of Brandt and Sigmund (2004) also called "behavioral strategy" in Ohtsuki and Iwasa (2004). This means that when two individuals, with given reputations, interact, the donor will 0 = NOT GIVE or 1 = GIVE according to his own strategy $\mathcal{A} = A(3)A(2)A(1)A(0)$, $A(i) = 0$ or 1, $i = 0,\ldots,3$, detailed in Table 2.

Table 1
String representation of a social norm

Donor (old score)	Recipient	Result of the game	Donor (new score)
GOOD	GOOD	GIVE	$N(7)$
GOOD	GOOD	NOT GIVE	$N(6)$
GOOD	BAD	GIVE	$N(5)$
GOOD	BAD	NOT GIVE	$N(4)$
BAD	GOOD	GIVE	$N(3)$
BAD	GOOD	NOT GIVE	$N(2)$
BAD	BAD	GIVE	$N(1)$
BAD	BAD	NOT GIVE	$N(0)$

Each norm is represented by a string of 8 bits, each of which determines the new reputation of an individual based on his decision to GIVE or NOT GIVE when acting as a donor towards a given recipient. The final assessment takes into account not only the action of the donor, but also the previous reputations of both donor and recipient, leading to the 8 possible combinations tabulated.

Table 2
String representation of an individual strategy

Donor's score	Recipient's score	Donor's behavior
GOOD	GOOD	$A(3)$
GOOD	BAD	$A(2)$
BAD	GOOD	$A(1)$
BAD	BAD	$A(0)$

Each 4-bits strategy determines the individual's action as a donor—to GIVE (1) or NOT GIVE (0)—to a given recipient, a decision based on the reputation of both individuals, which is determined by the social-norm under which they live in a given island. This leads to the 4 possible combinations tabulated.

Even when the strategy of an individual compels him to cooperate, he may fail to do so with a probability $\mu_e \ll 1$, which allows for the occurrence of execution errors.

Overall, there are 16 (2^4) different strategies. Initially all strategies are randomly distributed among players and after each generation these strategies reproduce according to their relative payoff (population size is kept constant).

In our simulations we have $I_0 = 64$ islands with $n_0 = 128$ individuals each. To each island we attribute a randomly generated 8-bits string, the "norm" $\mathcal{N}_I = N_I(7)\cdots N_I(0)$, $N_I(i) = 0$ or 1, $I = 1,\ldots,I_0$, $i = 0,\ldots,7$. These norms can be considered as resulting from the "founders effect", so we will not discuss the evolutionary path that leaded to them.

Inside each island, each player interacts once with every other player by means of the give-and-receive game with given and fixed parameter $b>1$, assuming with equal probability the role of donor or receiver. This applies to all islands.

After all interactions take place, one generation has passed. "Reproduction at the lowest level of selection is based on payoff in the following way (Brandt and Sigmund, 2004): to reproduce individual A choose randomly, proportional to payoff, a neighbor (say, B) of A (including A). A's offspring will then inherit B's strategy."

Furthermore, after one generation, we compute the normalized average payoff GDP_I for each island which will be used whenever pairs of islands engage in conflict, as described next.

2.2. Conflict between communities

Assume a complete graph for the Network of Contacts (NoC) of each island, i.e. every island is directly connected to every other island. (The same is assumed for the individuals in the previous section.) This assumption corresponds to the settings used in Brandt and Sigmund (2004) for finite populations as well as in Ohtsuki and Iwasa, 2004 for infinite populations.

We attribute a probability W_c, identical for all islands, that, during war-time, each island engages in a conflict with any other island. If island $\mathbb{A}$ goes to war, then we choose its adversary from the set of direct neighbors specified by the NoC. Let us call it $\mathbb{B}$. GDPs are GDP_A and GDP_B, respectively.

For each island there are two possible strategies, HAWK and DOVE, similar to the Hawk-and-Dove game described in Smith (1982). We give payoff values V for "victory", T for the investment of each player when both decide to play DOVE, W for the cost of fighting for the winner and L for the cost of fighting for the loser. We also introduce $p(\mathrm{GDP}_A, \mathrm{GDP}_B) = p(\mathrm{GDP}_A - \mathrm{GDP}_B)$ for the probability that $\mathbb{A}$ wins a contest against $\mathbb{B}$ (estimated by $\mathbb{A}$) when both play HAWK with given GDP. In particular, we shall adopt

$$p = p_\theta(x) = [1 + \exp(-x/\theta)]^{-1},$$

where $\theta \geqslant 0$ is the "temperature", assumed equal for all islands. (This "isothermal" assumption means that there is an universal way of estimating chances of victory, which, by chance, is the correct estimation.) Note that $p_\theta(x) + p_\theta(-x) = 1$.

Following Crowley (2000), our payoff matrix $\mathbf{M}$ (for player $\mathbb{A}$) is given by Table 3.

The most interesting scenario (Crowley, 2000) occurs whenever $L > W > 0$, $V > W > 0$, $L + W > V > 2T > 0$ and, in order to avoid negative payoffs, we add the absolute value of the minimum possible payoff, L, to all players after one conflict, a procedure which does not introduce any changes in the game.

We assume that islands are such that island $\mathbb{A}$ will play HAWK with probability $q(p_\theta(\mathrm{GDP}_A - \mathrm{GDP}_B))$ associated with the Nash equilibrium of the game's payoff matrix.

Defining $r := (V - W + L)p - L$ we write:

$$q(p) = \begin{cases} 1 & \text{if } r \geqslant 0, \\ \left[1 - \dfrac{r}{(V/2) + T}\right]^{-1} & \text{otherwise.} \end{cases}$$

To understand the above formula, we consider the payoff matrix $\mathbf{M}$ in Table 3. If $(V - W)p - L(1 - p) = r \geqslant 0$ then, it always pays to play HAWK. Now, consider $r < 0$. In this case the Nash equilibrium is given by a mixed strategy, which is equivalently determined (Hofbauer and Sigmund, 1998) as the attractor of the replicator dynamics

$$\dot{q} = q((\mathbf{M}\vec{q})_2 - \vec{q}\cdot\mathbf{M}\vec{q}) = q(1-q)\left(\left(r - \left(\frac{V}{2} + T\right)\right)q + \left(\frac{V}{2} + T\right)\right),$$

where $\vec{q} = (1 - q, q)$. The equilibrium points are given by $q = 0$, $q = 1$ (unstable) and the only stable point (ω-point) is given by $q = q(p)$.

Similarly, island $\mathbb{B}$ will play HAWK with probability $q(p_\theta(\mathrm{GDP}_B - \mathrm{GDP}_A)) = q(1 - p_\theta(\mathrm{GDP}_A - \mathrm{GDP}_B))$. It is important to understand q as a probability (of playing HAWK) and not as a frequency.

After conflict, the norms adopted by islands $\mathbb{A}$ and $\mathbb{B}$ will possibly change from what they were before. Let $Q(A)$ be the payoff obtained by $\mathbb{A}$ and $Q(B)$ that obtained by $\mathbb{B}$ as a result of the game. Then:

- If $\mathbb{A}$ played HAWK and $Q(A) > Q(B)$, then the 8-bit norm $\mathcal{N}_B = N_B(7)\cdots N_B(0)$, will change according to:

  $$N_B^{\mathrm{NEW}}(i) = N_B^{\mathrm{OLD}}(i) \text{ with probability } \frac{(1-\eta)Q(B)}{\eta Q(A) + (1-\eta)Q(B)},$$

  $$N_B^{\mathrm{NEW}}(i) = N_A^{\mathrm{OLD}}(i) \text{ with probability } \frac{\eta Q(A)}{\eta Q(A) + (1-\eta)Q(B)},$$

  if $N_A^{\mathrm{OLD}}(i) \neq N_B^{\mathrm{OLD}}(i)$ and is mutated by $\mu_N \ll 1$ if $N_A^{\mathrm{OLD}}(i) = N_B^{\mathrm{OLD}}(i)$. The parameter $\eta \in [0, 1]$ is the "will for change", such that $1 - \eta$ acts as an inertia for changing.

  The population (and its individual strategies) also changes in this case, eliminating at random a fraction $f = \eta' Q(A)/(\eta' Q(A) + (1 - \eta')Q(B))$, $\eta' \in [0, 1]$, of individuals of the defeated island, replacing them (again at random) by individuals (and their strategies) from the victor island.
- Same as before, swapping $\mathbb{A}$ and $\mathbb{B}$.
- If $\mathbb{A}$ played HAWK and $Q(A) \leqslant Q(B)$ or if $\mathbb{A}$ played DOVE, then norm entries $N_B(i)$ are mutated with probability $\mu_N \ll 1$ and the population strategies are mutated by μ_S.
- Same as before, swapping $\mathbb{A}$ and $\mathbb{B}$.

Table 3
Payoff matrix for the Island's game

	DOVE	HAWK
DOVE	$\frac{V}{2} - T$	0
HAWK	V	$(V - W)p - L(1 - p)$

Depending on the role played by each island—HAWK or DOVE—their payoff after the conflict is given by the matrix above, which also determines the partitioning of resources between the two intervening islands.

This provides an updating rule for the norms and for the population, and after this step we are back to peace, evolution being dominated by individual interactions within each island. During "peace" time, norms are kept constant.

The war-and-peace cycle is repeated until it reaches a stationary state.

It is important to state that it is not clear to what extent is this Hawk-and-Dove game a fundamental feature of the conclusions below. Other games can be easily imagined, as for example the "War of Attrition". In this case the GDP should be identified (as before) with the ability to win the contest and should also be clearly exhibited by the island, in the form of display. The result of the game is then settled without escalation, in such a way that the island with larger GDP is certainly the victor (Smith, 1982).

3. Numerical simulations and analysis

In our simulations, we adopted the following values: $I_0 = 64$, $n_0 = 128$, $V = 1$, $T = 0.01$, $W = 1/2$, $L = 3/4$, $\theta = 0.0005$, $\eta = 0.1$, $\eta' = 0.0005$, $\mu_N = 0.0001$, $\mu_S = 0.01$, $\mu_a = \mu_e = 0.001$, $W_c = 0.001$. The benefit b varied from a large range of parameters (see below). We ran the simulation for 5000 generations and computed the average using the last 1000 results. (As a cross-validation, results did not change if instead we ran simulations for 10000 generations, which means that—some kind of—steady state has been reached. No changes are observed if we choose another set of parameters, i.e, the results presented below seem to be very robust.)

The first result to be analysed is the (statistical) distribution of norms among all the islands in the archipelago. If the mutation rate were small enough (which does not seem to be the case), one could expect each bit of the norm—we call it "gene"—to be fixed or lost (as is well-known in population genetics; Kimura, 1962). Unfortunately small mutation rates could easily lead to entrapment in unwanted metastable states. Consequently, we decided to increase the mutation rate and extract statistical information from "genes" not necessarily fixed in the following way: We consider the final result (i.e. the last 1000 generations) and state that "gene" fixation occurs whenever it was present in more than 98% of the islands' norms in a given archipelago (we found the same results using as threshold for fixation 95%). In this way we created what we call a "Pseudo-Macro-Norm" (PMN) for every run, consisting of a 8-digits string of 0, 1 (if 0 or 1 were fixed, respectively) and X (if none was fixed). Typical results are provided in Table 4 for different values of b.

We repeat this analysis for 10 groups of 20 simulations, generating 10 tables of gene frequency in PMNs for different values of the benefit b. From these 10 tables, we compute the average (f_0, f_1 and f_X) and the standard deviation (d_0, d_1 and d_X) of the gene frequencies of 0, 1 or X in the PMN, on the basis of which we construct what we call a "Meta-Norm". This consists of a string of 8 genes, 0, 1 or · obtained from the following rules: if $|f_0 - f_1| > \max\{d_0, d_1, f_X\}$, we say that 0 is fixed (if $f_0 > f_1$) or 1 is fixed (if $f_1 > f_0$). If neither is fixed, we write · for that "gene" position. Results for the Meta-Norms as a function of b are provided in Table 5.

From Tables 4 and 5 it is clear that the Norm 10011001(= 153) is ubiquitous. More precisely, all norms in Table 5 are (at most) degeneracies of the norm 153. For $b = 20$, we show in Table 6 the frequencies of 0 and 1 (φ_0, φ_1) of the final norm of each island, for each gene, including all (200) simulations. The results are fully consistent with the dominance of the norm 153.

Despite the fact that genes 1, 2, 3, 4, 6 and 7 are not the most frequent they are the only ones to appear for every b in the studied range. On the other hand, for $b = 20$, genes 4, 7, 2 and 0 are the most frequent (in this order). As such, the strongest information (very frequent and valid for large range of benefits) is given by gene 4: GOOD individuals remain GOOD if they refuse to help BAD individuals. After comes gene number 7, which states that GOOD individuals remain GOOD by helping GOOD individuals. The next genes (in importance) states that when a BAD individual refuses help to a GOOD individual, he remains BAD whereas if the same person acts the same way against another BAD player, his/her reputation will change right away.

Table 4
PMN dependence on benefit

b	PMN	b	PMN	b	PMN
	1011101X		10111001		1101X001
1.25	10111101	5	10111X1X	15	10011010
	10111010		10011001		10X11001
	10111000		10011001		10011001
1.5	100110X0	8	100111XX	18	10111010
	10011001		10011000		10011000
	11011001		10111110		1X000001
2	X1011001	10	10X11001	20	110X1000
	10111X10		11010001		10111011

For each value of the benefit b, we illustrate typical results obtained for the Pseudo-Macro-Norm defined in main text. The results show how specific genes in this PMN persist for all valus of b.

Table 5
Meta-Norm dependence on benefit

b	Meta-Norm	b	Meta-Norm
1.25	1001100·	8	10011001
1.5	1001100·	10	10011001
2	10011001	20	10 · 11001
3	10011001	30	10 · 1100·
4	10011001	40	10 · 11001
5	10011001	100	10 · 1100·
6	10011001	200	10 · 1100·

For each value of the b, we show the results obtained for the Meta-Norm defined in main text. Irrespective of whether a given gene is fixed or not, all Meta-Norms are consistent with norm 153 for all values of b.

Table 6
Frequency of 0 and 1 for each gene

Gene	φ_0	φ_1	Gene	φ_0	φ_1
0	0.196	**0.804**	4	0.084	**0.916**
1	**0.729**	0.271	5	**0.593**	0.407
2	**0.849**	0.151	6	**0.766**	0.234
3	0.296	**0.704**	7	0.149	**0.851**

For a benefit $b = 20$, we tabulate the overall frequency of occurrence of bit-values 0 (φ_0) and 1 (φ_1) for each bit in the social norms occurring during the last 1000 generations of all islands in a total of 200 archipelago-simulations carried out.

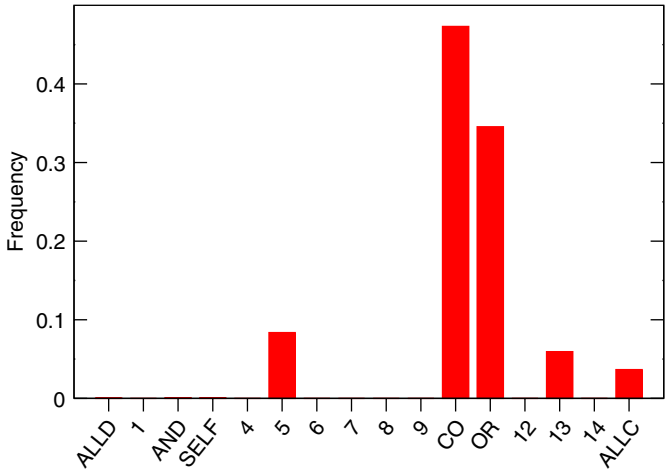

Fig. 1. Different strategies which co-evolve under the norm 153. We plot the frequency with which each strategy occurs in an island under the norm 153. The last 1000 generations of each of the 200 simulations carried out for $b = 20$ have been used. The largest share is split between the cooperative strategies CO and OR, followed by strategies 5 and 13 and, to a lesser extent, by strategy ALLC. The predominance of cooperative strategies is clear, strategies 5 and 13 being the opposite strategies of CO and OR, respectively.

When we fix the island norm to 153 and investigate the evolution of the strategies under such a norm, islands become dominated by the cooperative strategies CO and OR (Brandt and Sigmund, 2004), with over 1/3 of the population adopting each, as shown in Fig. 1, whereas strategies 5 and 13 occur with a frequency of 10% or less, ALLC completing the strategy distribution.

4. Discussion

Our model of the evolution of social norms naturally leads to the emergence of several features which are widely recognized as prototypical in what concerns the existence of social norms and their role in human cooperation. Whenever one of the two islands engaging in a conflict wins, a small fraction η' of individuals (and their strategies) from the victor island migrate to the loosing island. We used a very small value for this fraction ($\eta' = 0.0005$), but still this parameter is of fundamental importance. This number means that up to 3.2% of the island's population migrates after one round. The discussion below does not change for others small but positive values of η'. If $\eta' = 0$, not only it is harder for norms to evolve into a stationary state, but also individual strategies inside a given island often fail to maximize the fitness of that island. This happens because ALLD is the only ESS for all possible norms (Ohtsuki and Iwasa, 2004). This changes profoundly whenever $\eta' \neq 0$ in which case islands rapidly achieve a maximal possible fitness, and strategy evolution inside each island is much more consistent. This is in strong agreement with previous suggestions that a small migration seems to be essential for group evolution (Soltis et al., 1995; Rogers, 1990). Furthermore, norms also evolve into better defined stationary states. However, and in full accord with empirical evidence (Fehr and Fischbacher, 2004), it is possible that different strategy distributions maximize an island's fitness, under the same norm. Likewise, different norms may coexist with maximum fitness with an underlying distribution of strategies in the population which may be distinct. This, in turn, corresponds entirely to the existing anthropological evidence indicating that human groups differ greatly in their social norms (Sober and Wilson, 1998; Henrich et al., 2001) and that the existence of social norms favors the occurrence of few strategies within groups, but a great heterogeneity in the distribution of strategies among different groups abiding to different norms.

A very interesting feature of the norm 10011001 is that the only important information for moral judgments resides in the recipient's score and the final result, but not in the current status of the donor. In other words, the only GOOD behaviors are: help GOOD individuals and refuse help to BAD ones which makes it easier for anyone to achieve good standards after a single act. The last gene, $N(0)$, for example, is associated with the possibility of "forgiveness", which seems to be very important for a rapid achievement of a high level of cooperation. On the other hand, under this norm, society can readily punish a GOOD player after a single BAD move. This results in a strong pressure toward cooperation.

It is interesting to note that in the norms (“assessment modules”) studied in Brandt and Sigmund (2004) the possibility of forgiveness to a BAD player not helping other BAD players has been ruled out. More precisely, if a BAD player meets another BAD player in the JUDGING norm, he/she cannot change his/her score. On the other hand, a GOOD player who meets a BAD one in the STANDING norm is in an extremely comfortable situation: (s)he will be considered GOOD in the next round irrespective of the action decided.

In Axelrod (1984), the success of the strategy TIT-FOR-TAT in the iterated prisoner's dilemma was attributed to some simple facts, including not being the first to defect and reciprocating immediately both to cooperation and to defection. We can do the same analysis here. The success of the norm 153 can be attributed to never being morally dubious (to each encounter there is one GOOD move and one BAD one). Also, it is always possible for anyone to be promoted to the best standard possible in a single move. Conversely, one bad move will be readily punished with the reduction of the player's score. It is important to stress that this norm has zero history (meaning that the previous donor's score is immaterial to the result), while our simulations allowed up to one-round history (donor's score) and memory (recipient's score). Consequently, if we restrict the analysis to the subspace of zero-history norms then $N(7)N(6)N(5)N(4) = N(3)N(2)N(1)N(0)$. Imposing a clear definition of GOOD and BAD behaviors, we have $N(7)N(6) = 10$ or $N(7)N(6) = 01$ and the same for $N(5)N(4)$. We can go further and say that GOOD and BAD individuals should be treated differently, and then $N(7) \neq N(5)$ and $N(6) \neq N(4)$. Within this subspace, we have only 10011001 and 01100110, and only the first is able to promote cooperation (is economically viable). As with the iterated prisoner's dilemma, simplicity is the key to the success. This provides a direct interpretation of the bits 0 and 1, justifying why we called then BAD and GOOD, respectively.

As pointed out in Mackie (1995) and Alexander (1995), a better understanding of reputation dynamics should include group selection, leading to selection operating at different levels. Our results making use of the presently developed two-level selection framework show that the conjectures advanced in Mackie (1995) and Alexander (1995) are entirely justified. In Ohtsuki and Iwasa (2004), an evolutionary explanation of goodness was drafted. In this sense, this work confirms such a sound definition of good and bad behaviors. We further note that, in the notation of Ohtsuki and Iwasa (2004), norm 153 is one of the “leading eight”. For this norm the strategy CO is an ESS (Ohtsuki and Iwasa, 2004), being precisely one of the leading strategies for the norm 153 found in our simulation. It is important to remember that the concept of ESS in reference (Ohtsuki and Iwasa, 2004) is related only to invasion by a small amount of different strategists, while here we also consider a second level of competition, namely, for the islands. ESSs for two-level games were not computed in this work.

This work leaves many questions unanswered. In particular, we believe it is worth investigating how the intrinsic cognitive capacity of each individual will affect the evolution of social norms, since changes may have an impact not only in what concerns the overall pressure toward cooperation, but also in the capacity of each individual to be socially forgiven, thereby recovering a good reputation. After all, it is precisely the cognitive capacity of humans that allows them to set and enforce social norms (Stevens and Hauser, 2004), thereby distinguishing them from all other biological species. More specifically, a natural follow up of this work would be to include norms with longer memory and history capabilities. However, to the extent that the previous analysis is correct, we expect to observe no substantial changes in the result, that is, the resulting norm should also exhibit zero-history. On the other hand, one may allow different scores for different individuals in the same island, both generalizations being realizable in nature at the expense of additional cognitive capabilities (individual recognition and long memory and history). Another possible generalization would be to allow for intermediate values between 0 and 1 for each norm “bit”. Furthermore, the role of population structure, not only between individuals in each island but also in what concerns the (co-evolving) topology of the NoC between islands in an archipelago, should not be overlooked, in view of the recent findings concerning the importance of the NoC in promoting cooperation (Pacheco and Santos, 2005).

Acknowledgements

FACCC has been supported by the project POCTI/ISFL/209 (FCT/Portugal). FCS acknowledges the support of $COMP^2SYS$, a Marie Curie Early Stage Training Site, funded by the EC through the HRM activity.

References

Alexander, R.D., 1995. A biological interpretation of moral systems. In: Thompson, P. (Ed.), Issues in Evolutionary Ethics. State University of New York Press, Albany, NY, pp. 179–202.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York, USA.

Bowles, S., Ginits, H., 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor. Popul. Biol. 65, 17–28.

Bowles, S., Choi, J.-K., Hopfensitz, A., 2003. The co-evolution of individual behaviors and social institutions. J. Theor. Biol. 223, 135–147.

Boyd, R., Richerson, P.J., 1985. Culture and the Evolutionary Process. University of Chicago Press, Chicago.

Boyd, R., Richerson, P.J., 1990. Group selection among alternative evolutionarily stable strategies. J. Theor. Biol. 145 (3), 331–342.

Boyd, R., Gintis, H., Bowles, S., Richerson, P.J., 2003. The evolution of altruistic punishment. Proc. Natl Acad. Sci. 100 (6), 3531–3535.

Brandt, H., Sigmund, K., 2004. The logic of reprobation: assessment and action rules for indirect reciprocation. J. Theor. Biol. 231 (4), 475–486.

Brandt, H., Sigmund, K., 2005. Indirect reciprocity, image scoring, and moral hazard. Proc. Natl Acad. Sci. 102 (7), 2666–2670.

Crowley, P.H., 2000. Hawks, doves and mixed-symmetry games. J. Theor. Biol. 204 (4), 543–563.

Dawkins, R., 1976. The Selfish Gene. Oxford University Press, Oxford, UK.

Fehr, E., Fischbacher, U., 2004. Social norms and human cooperation. Trends Cogn. Sci. 8 (4), 185–190.

Gould, S.J., 1984. Caring groups and selfish genes. In: Sober, E. (Ed.), Conceptual Issues in Evolutionary Biology: An Anthology. MIT Press, Cambridge, MA, pp. 119–124.

Henrich, J., Boyd, R., 2001. Why people punish defectors. Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. J. Theor. Biol. 208, 79–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., 2001. In search of homo economics: behavioral experiments in 15 small-scale societies. Am. Econ. Rev. 91 (1), 73–78.

Hofbauer, J., Sigmund, K., 1998. Evolutionary Games and Population Dynamics. Cambridge University Press, Cambridge, UK.

Kandori, M., 1992. Social norms and community enforcement. Rev. Econ. Stud. 59, 63–80.

Keller, L. (Ed)., 1999. Levels of Selection in Evolution. Monographs in Behavior and Ecology, Princeton University Press, Princeton, NJ.

Kimura, M., 1962. On the probability of fixation of mutant genes in a population. Genetics 47, 713–719.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocation. Proc. R. Soc. London B 268, 745–753.

Mackie, J.L., 1995. The law of the jungle: moral alternatives and principle of evolution. In: Thompson, P. (Ed.), Issues in Evolutionary Ethics. State University of New York Press, Albany, NY, pp. 165–177.

Nowak, M., Sigmund, K., 1998a. The dynamics of indirect reciprocity. J. Theor. Biol. 194, 561–574.

Nowak, M., Sigmund, K., 1998b. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Ohtsuki, H., Iwasa, Y., 2004. How should we define goodness?—reputation dynamics in indirect reciprocity. J. Theor. Biol. 231 (1), 107–120 Erratum in: J. Theor. Biol. 232(4), 451 (2005).

Pacheco, J.M., Santos, F.C., 2005. Network dependence of the social dilemmas of cooperation. Mendes, J.F.F. (Ed.), Science of Complex Networks: From Biology to the Internet and WWW, vol. 776. AIP Conference Proceedings, New York, p. 90.

Paulsson, J., 2002. Multilevelled selection on plasmid replication. Genetics 161, 1373–1384.

Pomiankowski, A., 1999. Intragenomic conflict. In: Keller, L. (Ed.), Levels of Selection in Evolution. Monographs in Behavior and Ecology. Princeton University Press, Princeton, NJ, pp. 121–152.

Rogers, A.R., 1990. Group selection by selective emigration: the effects of migration and kin structure. Am. Nat. 135, 398–413.

Smith, J.M., 1982. Evolution and the Theory of Games. Cambridge University Press, Cambridge, UK.

Sober, E., Wilson, D.S., 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

Sober, E., Lewontin, R.C., 1984. Artifact, cause and gene selection. In: Sober, E. (Ed.), Conceptual Issues in Evolutionary Biology: An Anthology. MIT Press, Cambridge, MA, pp. 210–231.

Soltis, J., Boyd, R., Richerson, P.J., 1995. Can group-functional behaviors evolve by cultural group selection? An empirical test. Cur. Anthropol. 36 (3), 473–494.

Stevens, J.R., Hauser, M.D., 2004. Why be nice? Psychological constraints on the evolution of cooperation. Trends Cogn. Sci. 8, 60–65.

Trausen, A., Sengupta, A.M., Nowak, M.A., 2005. Stochastic evolutionary dynamics in two levels. J. Theor. Biol. 235 (3), 393–401.