Stern-Judging: A Simple, Successful Norm Which Promotes Cooperation under Indirect Reciprocity

Jorge M. Pacheco[1], **Francisco C. Santos**[2], **Fabio A. C. C. Chalub**[3]*

1 Centro de Física Teórica e Computacional and Departamento de Física da Faculdade de Ciências, Lisbon, Portugal, **2** IRIDIA, CoDE, Université Libre de Bruxelles, Brussels, Belgium, **3** Departamento de Matemática da Universidade Nova de Lisboa and Centro de Matemática e Aplicações, Caparica, Portugal

We study the evolution of cooperation under indirect reciprocity, believed to constitute the biological basis of morality. We employ an evolutionary game theoretical model of multilevel selection, and show that natural selection and mutation lead to the emergence of a robust and simple social norm, which we call *stern-judging*. Under stern-judging, helping a good individual or refusing help to a bad individual leads to a good reputation, whereas refusing help to a good individual or helping a bad one leads to a bad reputation. Similarly for tit-for-tat and win-stay-lose-shift, the simplest ubiquitous strategies in direct reciprocity, the lack of ambiguity of stern-judging, where implacable punishment is compensated by prompt forgiving, supports the idea that simplicity is often associated with evolutionary success.

Citation: Pacheco JM, Santos FC, Chalub FACC (2006) Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. PLoS Comput Biol 2(12): e178. doi:10.1371/journal.pcbi.0020178

Introduction

Many biological systems employ cooperative interactions in their organization [1]. Humans, unlike other animal species, form large social groups in which cooperation among nonkin is widespread. This contrasts with the general assumption that the strong and selfish individuals are the ones who benefit most from natural selection. This being the case, how is it possible that unselfish behaviour has survived evolution? Adopting the terminology resulting from the seminal work of Hamilton, Trivers, and Wilson [2–4], an act is altruistic if it confers a benefit b to another individual in spite of accruing a cost c to the altruist (where it is assumed, as usual, that $b > c$). In this context, several mechanisms have been invoked to explain the evolution of altruism, but only recently an evolutionary model of indirect reciprocity (using the terminology introduced by [5]) has been developed by Nowak and Sigmund [6], which, with remarkable simplicity, addressed "unique aspects of human sociality, such as trust, gossip, and reputation" [7]. As a means of community enforcement, indirect reciprocity had been investigated earlier in the context of economics, notably by Sugden [8] and Kandori [9] (see below). More recently, many studies [7,8,10–17] have been devoted to investigating how altruism can evolve under indirect reciprocity. Indeed, according to Alexander [5], indirect reciprocity presumably provides the mechanism that distinguishes us humans from all other living species on Earth. Moreover, as recently argued in [10], "indirect reciprocity may have provided the selective challenge driving the cerebral expansion in human evolution." In the indirect reciprocity game, any two players are supposed to interact at most once with each other, one in the role of a potential donor, with the other as a potential receiver of help. Each player can experience many rounds, but never with the same partner twice, direct retaliation being unfeasible. By helping another individual, a given player may increase (or not) her reputation, which may change the predisposition of others to help her in future interactions. However, her new reputation depends on the social norm used by her peers to assess her action as a donor. Previous studies of reputation-based models of cooperation reviewed recently [10] indicate that cooperation outweighs defection whenever, among other factors, assessment of actions is based on norms that require considerable cognitive capacities [10,12,13], even when individuals are capable of making binary assessments only, in a "world in black and white" [10], as assumed in most recent studies (see, however, [6]). Furthermore, stable cooperation hinges on the availability of reliable reputation information [6]. Such high cognitive capacity contrasts with technology-based interactions, such as e-trade, which also rely on reputation-based mechanisms of cooperation [18–20]. Indeed, anonymous one-shot interactions between individuals loosely connected and geographically dispersed usually dominate e-trade, raising issues of trust-building and moral hazard [21]. Reputation in e-trade is introduced via a feedback mechanism which announces the ratings of sellers. Despite the success and high levels of cooperation observed in e-trade, it has been found [18] that publicizing a detailed account of the seller's feedback history does not improve cooperation, as compared with publicizing only the seller's most recent rating. In other words, practice shows that simple reputation-based mechanisms are capable of promoting high levels of cooperation. In view of the previous discussion, it is

Editor: Simon A. Levin, Princeton University, United States of America

Received September 26, 2006; **Accepted** November 8, 2006; **Published** December 29, 2006

* To whom correspondence should be addressed. E-mail: chalub@cii.fc.ul.pt

Synopsis

Humans, unlike other animal species, form large social groups in which cooperation among non-kin is widespread. This contrasts with the general assumption that the strong and selfish individuals are the ones who benefit most from natural selection. Among the different mechanisms invoked to explain the evolution of cooperation, indirect reciprocity is associated with cooperation supported by reputation: I help you and someone else helps me. However, how did reputation evolve and which type of moral is encapsulated in those social norms that are evolutionary successful? Suggesting a simple scenario for the evolution of social norms, Pacheco, Santos, and Chalub propose a reputation-based multilevel selection model, where individual behaviour and moral systems co-evolve, governed by competition and natural selection. Evolution leads to the emergence of a simple and robust social norm, which the authors call *stern-judging,* where implacable punishment goes side-by-side with prompt forgiving. The low level of complexity of this norm, which is supported by empirical observations in e-trade, conveys the idea that simplicity is often associated with evolutionary success.

hard to explain the success of e-trade on the basis of the results obtained so far for reputation-based cooperation in the context of indirect reciprocity.

A Model of Multilevel, Multigame Selection

Let us consider a world in black and white consisting of a set of tribes, such that each tribe lives under the influence of a single norm, common to all individuals (see Figure 1). Each individual engages once in the indirect reciprocity game (cf. Methods) with all other tribe inhabitants. Her action as a donor will depend on her individual strategy, which dictates whether she will *provide help* or *refuse to do it* depending on her and the recipient's reputation. Reputations are public: this means that the result of every interaction is made available to everyone through the "indirect observation model" introduced in [13] (see also [15]). This allows any individual to know the current status of the co-player without observing all of her past interactions. On the other hand, this requires a way to spread the information (even with errors) to the entire population (communication/language). Consistently, language seems to be an important cooperation promoter [22], although recent mechanisms of reputation-spreading rely on electronic databases (e.g., in e-trade, where reputation of sellers is centralized). Since reputations are either *GOOD* or *BAD,* there are $2^4 = 16$ possible strategies. On the other hand, the number of possible norms depends on their associated order. The simplest are the so-called *first-order norms,* in which all that matters is the action taken by the donor. In *second-order norms,* the reputation of one of the players (donor or recipient) also contributes to decide the new reputation of the donor. And so on, in increasing layers of complexity (and associated requirements of cognitive capacities from individuals) as shown in Figure 2, which illustrates the features of third-order norms such as those we shall employ here. Any individual in the tribe shares the same norm, which in turn raises the question of how each inhabitant acquired it. We do not address this issue here. However, inasmuch as indirect reciprocity is associated with "community enforcement" [9,10], one may assume, for simplicity, that norms are acquired through an educational process. Moreover, it is likely that a common norm contributes to the overall cohesiveness and identity of a tribe. It is noteworthy, however, that if norms were different for different individuals, the "indirect observation model" would not be valid, as it requires trust in judgments made by co-inhabitants. For a norm of order n, there are 2^{2^n} possible norms, each associated with a binary string of length 2^n. We consider third-order norms (8-bit strings, Figure 2): in assessing a donor's new reputation, the observer has to make a contextual judgment involving the donor's action, as well as her and the recipient's

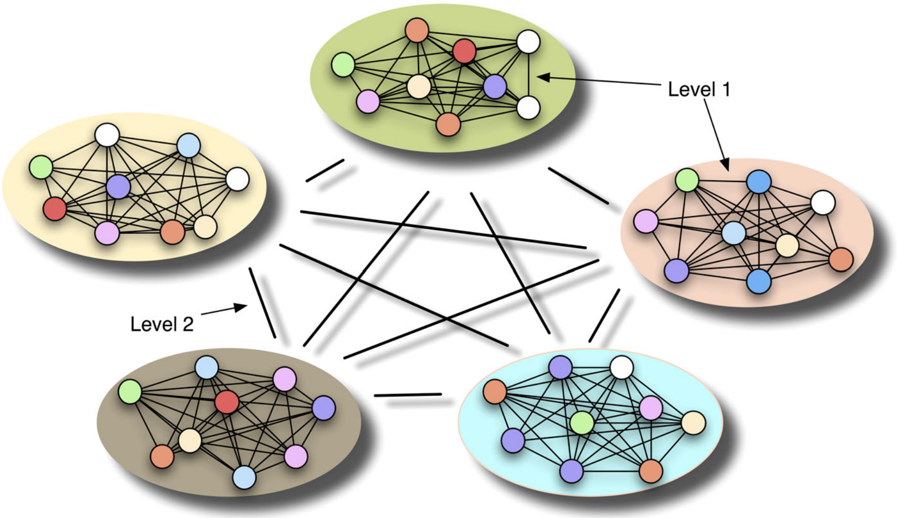

Figure 1. Multilevel Selection Model for the Evolution of Norms

Each palette represents a tribe in which inhabitants (coloured dots) employ different strategies (different colours) to play the indirect reciprocity game. Each tribe is influenced by a single social norm (common background colour), which may be different in different tribes. All individuals in each tribe undergo pairwise rounds of the game (lower level of selection, Level 1), whereas all tribes also engage in pairwise conflicts (higher level of selection, Level 2), as described in the main text. As a result of the conflicts between tribes, norms evolve, whereas evolution inside each tribe selects the distribution of strategies that best adapt to the ruling social norm in each tribe.

doi:10.1371/journal.pcbi.0020178.g001

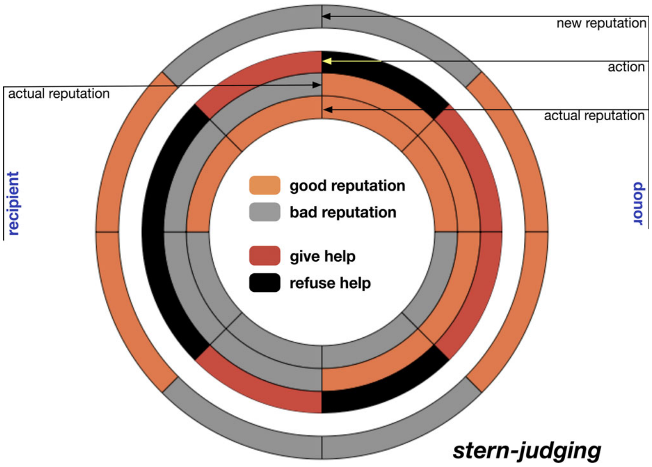

Figure 2. Norm Complexity

The higher the order (and complexity) of a norm, the more "inner" layers it acquires. The outer layer stipulates the donor's new reputation based on the three different reputation/action combinations aligned radially layer by layer: inward, the first layer identifies the action of the donor. The second layer identifies the reputation of the recipient; the third the reputation of the donor. Out of the 2^8 possible norms, the highly symmetric norm shown as the outer layer emerges as the most successful norm. Indeed, stern-judging renders the inner layer (donor reputation) irrelevant in determining the new reputation of the donor. This can be trivially confirmed by the symmetry of the figure with respect to the equatorial plane (not taking the inner layer into account, of course). All norms of second order will exhibit this symmetry, although the combinations of 1 and 0 bits will be, in general, different. We use this representation in Protocol S1 to depict other popular norms, namely, the *leading-eight, standing, simple-standing,* and *image-scoring*.
doi:10.1371/journal.pcbi.0020178.g002

reputations scored in the previous action. We introduce the following evolutionary dynamics in each tribe: during one generation all individuals interact once with each other via the indirect reciprocity game. When individuals "reproduce," they replace their strategy by that of another individual from the same tribe, chosen proportional to her accumulated payoff [12]. The most successful individuals in each tribe have a higher reproductive success. Since different tribes are "under the influence" of different norms, the overall fitness of each tribe will vary from tribe to tribe, as well as the plethora of successful strategies that thrive in each tribe (Figure 1). This describes individual selection in each tribe (Level 1 in Figure 1).

Tribes engage in pairwise conflicts with a small probability, associated with selection between tribes. After each conflict, the norm of the defeated tribe will change toward the norm of the victor tribe, as detailed in the Methods section (Level 2 in Figure 1). We consider different forms of conflict between tribes, which reflect different types of inter-tribe selection mechanisms: *group selection* [5,23–28] based on the average global payoff of each tribe (involving different selection processes and intensities; imitation dynamics, a Moran-like process; and the pairwise comparison process, the latter discussed in Protocol S1) as well as selection resulting from inter-tribe conflicts modeled in terms of games—the display game of war of attrition, and an extended hawk–dove game [14] (see Protocol S1). We perform extensive computer simulations of evolutionary dynamics of sets of *64* tribes, each with *64* inhabitants. Once a stationary regime is reached, we collect information for subsequent statistical analysis (cf. Methods). We compute the frequency of occurrence of bits ***1*** and ***0*** in each of the 8-bit locations. A bit is said to fixate if its frequency of occurrence exceeds or equals *98%*. Otherwise, no fixation occurs, which we denote by ***X***, instead of by ***1*** or ***0***. We analyze *500* simulations for the same value of ***b***, subsequently computing the frequency of occurrence φ_1, φ_0, and φ_X of the bits ***1***, ***0***, and ***X***, respectively. If $\varphi_1 > \varphi_0 + \varphi_X$, the final bit is ***1***; if $\varphi_0 > \varphi_1 + \varphi_X$, the final bit is ***0***; otherwise we assume it is indeterminate, and denote it by **•**. It is noteworthy that our bit-by-bit selection/transmission procedure, though artificial, provides a simple means of mimicking biological evolution, where genes are interconnected by complex networks and yet evolve independently. Certainly, a co-evolutionary process would be more appropriate (and more complex), and this will be explored in future work.

Results/Discussion

The results for different values of ***b*** are given in Table 1, showing that a unique, ubiquitous social norm emerges from these extensive numerical simulations. This norm is of *second-order,* which means that all that matters is the action of the donor and the reputation of the receiver. In other words, even when individuals are equipped with higher cognitive capacities, they rely on a simple norm as a key for evolutionary success. In a nutshell, helping a good individual or refusing help to a bad individual leads to a good reputation, whereas refusing help to a good individual or helping a bad one leads to a bad reputation. Moreover, we find that the final norm is independent of the specifics of the second-level selection mechanism, i.e., different second-level selection mechanisms will alter the rate of convergence, but not the equilibrium state. In this sense, we conjecture that more realistic procedures will lead to the same dominant norm.

The success and simplicity of this norm relies on never being morally dubious: to each type of encounter, there is one *GOOD* move and one *BAD* one. Moreover, 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 [29,30] with the reduction of the player's score. This prompt forgiving and implacable punishment leads us to call this norm *stern-judging*.

Long before the seminal work of Nowak and Sigmund [6], several social norms have been proposed as a means to promote (economic) cooperation. Notable examples are the *standing* norm, proposed by Sugden [8], and the norm proposed by Kandori [9], as a means to allow community enforcement of cooperation. When translated into the present formulation, *standing* constitutes a third-order norm, whereas a fixed-order reduction of the social norm proposed by Kandori (of variable order, dependent on the benefit-to-cost ratio of cooperation) would correspond to *stern-judging*. Indeed, in the context of community enforcement, one can restate *stern-judging* as: "Help good people and refuse help otherwise, and we shall be nice to you; otherwise, you will be punished."

It is, therefore, most interesting that the exhaustive search carried out by Ohtsuki and Iwasa [13,15] in the space of up to third-order norms found that these two previously proposed norms were part of the so-called *leading-eight* norms of

Table 1. Emerging Social Norm

b	Imitation Dynamics	Moran	Pairwise Comparison*	War of Attrition	Hawk–Dove*
2	1001 1001	1●01 1001	1001 1001	●●●● ●●●●	1001 1001
≥3	1001 1001	1001 1001	1001 1001	1001 1001	1001 1001

For each value of the benefit b (c = 1), each column displays the 8-bit norm emerging from the analysis of 500 simulations employing the selection method between tribes indicated as column headers. Irrespective of the type of selection, the resulting norm that emerges is always compatible with stern-judging. Details of the different selection processes are given in the Methods section (those marked with an * are provided in Protocol S1). For the pairwise comparison rule, the inverse temperature used was $\beta = 10^5$ (strong selection, cf. Protocol S1).

doi:10.1371/journal.pcbi.0020178.t001

cooperation. On the other hand, *image-score,* the norm emerging from the work of Nowak and Sigmund [6], which has the attractive feature of being a simple, second-order norm (like *stern-judging*) does not belong to the *leading-eight*. Indeed, the features of *image-scoring* have been carefully studied in comparison with *standing* [7,16,17], showing that *standing* performs better than *image-scoring,* mostly in the presence of errors [12].

Among the *leading-eight* norms discovered by Ohtsuki and Iwasa [13,15], only *stern-judging* [10] and the so-called *simple-standing* [31] constitute second-order norms (see below). Our present results clearly indicate that *stern-judging* is favored compared with all other norms. Nonetheless, in line with the model considered here, the performance of each of these norms may be evaluated by investigating how each norm performs individually, *taking into account all 16 strategies simultaneously*. In Protocol S1, we carry out such comparison: we consider *standing* among the third-order norms, as well as *stern-judging, image-scoring,* and *simple-standing* among second-order norms. Note that, in spite of fixating the norm, errors of assessment and of implementation as well as errors of strategy update are taken into account. The results show that the overall performance of *stern-judging* is better than that of the other norms over a wide range of values of the benefit ***b***. Furthermore, both *standing* and *simple-standing* perform very similarly, again pointing out that reputation-based cooperation can successfully be established without resorting to higher-order (more sophisticated) norms. Finally, *image-scoring* performs considerably worse than all the other norms, a feature already addressed [7,16,17]. Within the space of second-order norms, similar conclusions have been found recently by Ohtsuki and Iwasa [31]. Note, however, that the result obtained here is stronger than the analysis carried out in Protocol S1, since *stern-judging* emerges as the most successful norm surviving selection and mutation with other norms, irrespective of the selection mechanism. In other words, *stern-judging*'s simplicity and robustness to errors may contribute to its evolutionary success, since other well-performing strategies may succumb to invasion of individuals from other tribes who bring along strategies that may affect the overall performance of a given tribe. In this sense, robustness plays a key role when evolutionary success is at stake. We believe that *stern-judging* is the most robust norm promoting cooperation.

The present result correlates nicely with the recent findings in e-trade, where simple, reputation-based mechanisms ensure high levels of cooperation. Indeed, *stern-judging* involves a straightforward and unambiguous reputation assessment, decisions of the donor being contingent only on the previous reputation of the receiver. We argue that the absence of constraining environments acting upon the potential customers in e-trade, for whom the decision of buying or not buying is free from further ado, facilitates the adoption of a *stern-judging* assessment rule. Indeed, recent experiments [32] have shown that humans are very sensitive to the presence of subtle, psychologically constraining cues, their generosity depending strongly on the presence or absence of such cues. Furthermore, under simple unambiguous norms, humans may escape the additional costs of conscious deliberation [33].

As conjectured by Ohtsuki and Iwasa [13] (cf. also [5,23]), group selection might constitute the key element in establishing cooperation as a viable trait. The present results show that even when more sophisticated selection mechanisms operate between tribes, the outcome of evolution still favors *stern-judging* as the most successful norm under which cooperative strategies may flourish.

Materials and Methods

We considered sets of *64* tribes, each tribe with *64* inhabitants. Each individual engages in a single round of the following indirect reciprocity game [10] with every other tribe inhabitant, assuming with equal probability the role of donor or recipient. The donor decides *YES* or *NO,* if she will provide help to the recipient, following her individual strategy encoded as a 4-bit string [12–14] (which takes into account the current donor and recipient status—see Protocol S1). If *YES*, then her payoff decreases by ***1***, while the recipient's payoff increases by $b > 1$. If *NO*, the payoffs remain unchanged (following common practice [6,12,14,16,21], 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 tribe's social norm, will ascribe (subject to some small error probability $\mu_a = 0.001$) a new reputation to the *donor,* which we assume to spread efficiently without errors to the rest of the individuals in that tribe [12–14]. Moreover, individuals may fail to do what their strategy compels them to do, with a small execution error probability $\mu_e = 0.001$. After all interactions take place, one generation has passed, simultaneously for all tribes. Individual strategies in each tribe replicate to the next generation in the following way: for every individual ***A*** in the population, we select an individual ***B*** proportional to fitness (including ***A***) [12]. The strategy of ***B*** replaces that of ***A***, apart from bit mutations occurring with a small probability $\mu_s = 0.01$. Subsequently, with probability $p_{CONFLICT} = 0.01$, all pairs of tribes may engage in a conflict, in which each tribe acts as an individual unit. Different types of conflicts between tribes have been considered.

Imitation selection. We compare the average payoffs Π_A and Π_B of the two conflicting tribes ***A*** and ***B***, the winner being the tribe with highest score.

Moran process. In this case the selection method between tribes mimics that used between individuals in each tribe; one tribe ***B*** is chosen at random, and its norm is replaced by that of another tribe ***A*** chosen proportional to fitness.

War of attrition. We choose at random two tribes ***A*** and ***B*** with average payoffs Π_A and Π_B. We assume that each tribe can display its strength for a time, which is an increasing function of its average payoff. To this end, we draw two random numbers, R_A and R_B, each following an exponential probability distribution given by $exp(-t/\Pi_A)/\Pi_A$ and $exp(-t/\Pi_B)/\Pi_B$, respectively. The larger of the two numbers identifies the winning tribe.

As a result of inter-tribe conflict (two additional conflicts are discussed in Protocol S1), the norm of the losing tribe (***B***) is shifted in the direction of the victor norm (***A***). Convergence of such a nonlinear evolutionary process dictates a smooth norm crossover. Hence, each bit of norm ***A*** will replace the corresponding bit of norm ***B*** with probability

$$p = \frac{\eta\Pi_A}{\eta\Pi_A + (1-\eta)\Pi_B}$$

which ensures good convergence whenever $\eta \leq 0.2$, independently of the type of conflict (a bit mutation probability $\mu_N = 0.0001$ has been used). Furthermore, a small fraction of the population of tribe ***A*** replaces a corresponding random fraction of tribe ***B****:* each individual of tribe ***A*** replaces a corresponding individual of tribe ***B*** with a probability $\mu_{migration} = 0.005$. Indeed, if no migration takes place, a tribe's population may get trapped in less cooperative strategies, compromising the global convergence of the evolutionary process [26].

Each simulation runs for *9,000* generations, starting from randomly assigned strategies and norms, to let the system reach a stationary situation, typically characterized by all tribes having maximized their average payoff, for a given benefit $b > c = 1$. The subsequent *1,000* generations are then used to collect information on the strategies used in each tribe and the norms ruling the tribes in the stationary regime. We ran *500* evolutions for each value of ***b***, subsequently performing a statistical analysis of the bits that encode each norm, as detailed before.

The results presented are quite robust to variations of the different mutation rates introduced above, as well as to variation of population size and number of tribes. Furthermore, reducing the threshold from *98%* to *95%* does not introduce any changes in the results shown.

Supporting Information

Protocol S1. Supplementary Information

Found at doi:10.1371/journal.pcbi.0020178.sd001 (1.0 MB DOC).

Acknowledgments

JMP would like to thank Yoh Iwasa and Hisashi Ohtsuki for helpful discussions.

Author contributions. JMP, FCS, and FACCC conceived and designed the experiments, performed the experiments, analyzed the data, and wrote the paper.

Funding. JMP acknowledges support from FCT, Portugal, and the Program for Evolutionary Dynamics at Harvard University. FCS acknowledges the support of COMP^{2}SYS, a Marie Curie Early Stage Training Site, funded by the European Community through the HRM activity. FACCC was supported by project POCI/MAT/57546/2004.

Competing interests. The authors have declared that no competing interests exist.

References

1. Smith JM, Szathmáry E (1995) The major transitions in evolution. Oxford: Freeman.
2. Hamilton WD (1996) Narrow roads of gene land. Volume 1. New York: Freeman. 568 p.
3. Trivers R (1985) Social evolution. Menlo Park (California): Benjamin Cummings. 479 p.
4. Wilson EO (1975) Sociobiology. Cambridge (Massachusetts): Harvard University Press.
5. Alexander RD (1987) The biology of moral systems. NewYork: Aldine deGruyter. 300 p.
6. Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature 393: 573–577.
7. Panchanathan K, Boyd R (2003) A tale of two defectors: The importance of standing for evolution of indirect reciprocity. J Theor Biol 224: 115–126.
8. Sugden R (1986) The economics of rights, co-operation and welfare. Oxford: Basil Blackell. 256 p.
9. Kandori M (1992) Social norms and community enforcement. Rev Econ Studies 59: 63–80.
10. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. Nature 437: 1291–1298.
11. Fehr E, Fischbacher U (2003) The nature of human altruism. Nature 425: 785–791.
12. Brandt H, Sigmund K (2004) The logic of reprobation: Assessment and action rules for indirect reciprocation. J Theor Biol 231: 475–486.
13. Ohtsuki H, Iwasa Y (2004) How should we define goodness?—Reputation dynamics in indirect reciprocity. J Theor Biol 231: 107–120.
14. Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. J Theor Biol 241: 233–240.
15. Ohtsuki H, Iwasa Y (2006) The leading eight: Social norms that can maintain cooperation by indirect reciprocity. J Theor Biol 239: 435–444.
16. Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. Proc Biol Sci 268: 745–753.
17. Panchanathan K, Boyd R (2004) Indirect reciprocity can stabilize cooperation without the second-order free rider problem. Nature 432: 499–502.
18. Dellarocas C (2003) Sanctioning reputation mechanisms in online trading environments with moral hazard. Cambridge (Massachusetts): MIT Sloan School of Management. Working paper 4297–4203.
19. Bolton GE, Katok E, Ockenfels A (2004) How effective are electronic reputation mechanisms? An experimental investigation. Management Sci 50: 1587–1602.
20. Keser C (2002) Trust and reputation building in e-commerce. IBM-Watson Research Center. CIRANO working paper 2002s–2075k.
21. Brandt H, Sigmund K (2005) Indirect reciprocity, image scoring, and moral hazard. Proc Natl Acad Sci U S A 102: 2666–2670.
22. Brinck I, Gardenfors P (2003) Co-operation and communication in apes and humans. Mind Language 18: 484–501.
23. Mackie JL (1995) The law of the jungle: Moral alternatives and principle of evolution. In: Thompson P, editor. Albany: State University of New York Press. pp. 165–177.
24. Bowles S, Gintis H (2004) The evolution of strong reciprocity: Cooperation in heterogeneous populations. Theor Popul Biol 65: 17–28.
25. Bowles S, Choi JK, Hopfensitz A (2003) The co-evolution of individual behaviors and social institutions. J Theor Biol 223: 135–147.
26. Boyd R, Richerson PJ (1985) Culture and the evolutionary process. Chicago: University of Chicago Press. 340 p.
27. Boyd R, Richerson PJ (1990) Group selection among alternative evolutionarily stable strategies. J Theor Biol 145: 331–342.
28. Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proc Natl Acad Sci U S A 100: 3531–3535.
29. de Quervain DJ, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, et al. (2004) The neural basis of altruistic punishment. Science 305: 1254–1258.
30. Gintis H (2003) The hitchhiker's guide to altruism: Gene–culture coevolution, and the internalization of norms. J Theor Biol 220: 407–418.
31. Ohtsuki H, Iwasa Y (2007) Global analysis of evolutionary dynamics and exhaustive search for social norms that maintain cooperation and reputation. J Theor Biol. In press.
32. Haley KJ, Fessler DMT (2005) Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evol Hum Behaviour 26: 245–256.
33. Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB (2006) On making the right choice: The deliberation-without-attention effect. Science 311: 1005–1007.

SUPPLEMENTARY INFORMATION

Stern-judging : A simple, successful norm which promotes cooperation under indirect reciprocity

J. M. Pacheco[1], F. C. Santos[2] and F. A. C. C. Chalub[3 *]

Simulation details

In our simulations, we adopted the following values: η=0.1, μ_N=0.0001, μ_S=0.01, μ_a=μ_e=0.001. The benefit ***b*** varied from ***b*** =2 to ***b***=36.

We ran each simulation for *9000* generations and computed the average using the subsequent *1000* results. As a cross validation, results did not change if instead we ran simulations for *14000* generations, accumulating information over the subsequent *1000* generations. This indicates that a steady state has been reached. Finally, results are robust to reasonable changes of the parameters above.

Each individual, in each tribe, has a strategy (chosen randomly at start) encoded as a four-bit string, which determines the individual's action (**N**=no, do not provide help; **Y**=yes, provide help) as a donor, knowing hers and the recipient's reputation, as detailed in Table S1. This results in a total of 16 strategies, ranging from unconditional defection (***ALLD***) to unconditional cooperation (***ALLC***), as detailed in Table S2. These two extreme strategies are however, norm-independent. Hence, our statistical analysis only takes into account those steady states in which the prevalence of any of these strategies is below a given threshold. The results shown correspond to a maximum threshold of 10%, although results did not change by reducing or increasing this threshold by a factor of two.

Pairwise comparison and norm evolution for different intensities of selection

The pairwise comparison rule [1] provides a convenient framework to study how the intensity of selection between tribes affects the emergence of stern-judging. It corresponds to introduce the following dynamics: Given two tribes chosen for a conflict, say ***A*** and ***B***, with average payoffs Π_A and Π_B, respectively, then norm of tribe ***B*** will replace that of ***A*** with a probability given by

$$p = \left[1 + e^{-\beta(\Pi_B - \Pi_A)}\right]^{-1}$$

whereas the inverse process will occur with probability $(1-p)$. In physics this function corresponds to the well-known Fermi distribution function, in which the inverse temperature β determines the sharpness of transition from $p = 0$, whenever $\Pi_B < \Pi_A$, to $p = 1$, whenever $\Pi_A < \Pi_B$. Indeed, in the limit $\beta \rightarrow +\infty$ we obtain imitation dynamics (strong selection), whereas whenever $\beta \rightarrow 0$ ***B*** replaces ***A*** with the same probability that ***A*** replaces ***B*** (½ - neutral drift). As we change β between these two extreme limits, we can infer the role of selection intensity on the emergence of stern-judging. In Table S3 we show results for different values of β, which testify for the robustness of stern-judging. In other words, in spite of the fact that, with decreasing β (decreasing selection intensity), it becomes increasingly difficult for all 8 bits to fixate whenever ***b=2***, in no case do we get results which deviate from stern-judging as the emerging social norm. These results (together with the analysis carried out in the following for inter-tribe selection determined by a Hawk-Dove game), reinforce the conclusion that stern-judging is robust and ubiquitous.

Hawk-Dove Tribal Conflict

This method of tribal conflict has been developed in Ref. [2] and is based on an extended Hawk-Dove game introduced in Ref. [3]. If tribe ***A*** goes to war, then we choose at random its adversary (***B***) from the remaining tribes. Average payoffs of both tribes are denoted, as usual, by Π_A and Π_B respectively.

For each tribe there are two possible strategies, HAWK and DOVE, similar to the Hawk-and-Dove game described in [3]. The payoff matrix (for player *A*) reads

	DOVE	**HAWK**
DOVE	**V**/2 - **T**	0
HAWK	**V**	(**V**-**W**)p-**L**(1-p)

where $p(\Pi_A,\Pi_B) = p(\Pi_A\text{-}\Pi_B)$ is the probability that *A* wins a contest against *B* (estimated by *A*) when both play HAWK with given average payoff. In particular, we shall adopt $p=p_\beta(x) = [1+ \exp(-\beta x)]^{-1}$, where the inverse temperature $\beta>0$ is assumed to be the same for all tribes. The most interesting scenario [3] occurs whenever **L**>**W**>0, **V**>**W**>0, **L**+**W**>**V**>2**T**>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. Hence we adopted the values **V**=1, **T**=0.01, **W**=1/ 2, **L**=3/4 and $\beta=10^4$.

We assume that tribes are rational players, such that tribe *A* will play HAWK with probability $q(p_\theta(\Pi_A\text{-}\Pi_B))$ associated with the Nash equilibrium of the game's payoff matrix. Defining r:= (**V**-**W**+**L**)p-**L** we have $q(p)=1$ if $r=0$, and $q(p)=[1-r/(\mathbf{V}/2+\mathbf{T})]^{-1}$ otherwise. Similarly, tribe *B* will play HAWK with probability $q(p_\theta(\Pi_B\text{-}\Pi_A)) =q(1- p_\theta(\Pi_A\text{-}\Pi_B))$. After conflict, the norms adopted by tribes *A* and *B* will possibly change from what they were before. Let **Q**(*A*) be the payoff obtained by *A* and **Q**(*B*) that obtained by *B* as a result of the game. Then:

- If **A** played HAWK and **Q**(*A*) > **Q**(*B*), then each bit of norm of *B* will change with the probability defined in the methods section, incuding a mutation probability μ_N.
- Same as before, swapping *A* and *B*.

- If *A* played HAWK and **Q(*A*)** < **Q(*B*)** or *A* played DOVE, then norm entries $N_{\mathbf{B}}(i)$ are mutated with probability μ_N«1 and the population strategies are mutated by μ_S.
- Same as before, swapping ***A*** and ***B***.

Results for this update rule, shown in Table 1, provide clear evidence for the robustness and ubiquity of stern-judging. In this case, we obtain fixation of bits even for values of $b < 3$.

Cooperation under selected social norms

In order to better understand the success of ***stern-judging***, we carry out in the following a study of how tribes perform under the influence of a specific norm *which we now fix from the outset*. We shall compare the performance of ***stern-judging*** with the popular norms *standing* and *image-scoring*, as well as with the other second-order norm which incorporates the leading eight, coined strict standing [4]. We shall maintain mutation errors in strategy update, as well as errors of implementation. As a result, and given a fixed (immutable) norm, selection and mutation dictates the simultaneous evolution of all the 16 strategies in a given tribe. We are not aware of any study which undertook such a comparison. Indeed, in Ohtsuki and Iwasa's seminal work [5], they searched for well defined combinations of one norm which would constitute a non-trivial Evolutionary Stable Strategy in a monomorphic population with an associated cooperative strategy. Hence they discovered the leading eight. In Fig. S1 we depict the leading eight norms, using the convention of Fig. 2. The white "slices" correspond to places where both *GOOD* (orange) or *BAD* (grey) reputations can be freely assigned, the remaining norm being on of the leading eight. Since a second order norm, in this representation, is simply a norm which exhibits a mirror symmetry with respect to the equatorial plane, it is obvious that there are only two second order norms which incorporate the leading eight: Besides ***stern-judging***, also *simple-standing* belongs to the leading eight. Both norms form the first row of Fig. S2, whereas image-scoring and standing, the original norm proposed by Sugden, complete the lower row in Fig. S2.

Brandt and Sigmund [6] have carried an individual based model analysis in which evolution took place under selection and mutation between individuals whose norm (in the sense defined here) was individually assigned, as well as the strategy. Moreover, information was private, not public. Finally, Ohtsuki and Iwasa have recently [4] examined which strategies thrive under the presence of a single, second-order norm, now in a (infinite) polymorphic population in which individuals can adopt three out of the 16 strategies considered in this work. Their analytic study leads to the conclusion that, in the presence of errors, stable coexistence between conditional and unconditional cooperators is possible, ***stern-judging*** constituting one of the leading norms promoting cooperative behaviour.

In Fig. S3 we show results for the ratio between the average payoff reached in each tribe and the maximum average payoff attainable in that tribe, given the tribe size and the benefit (keeping cost=1). This quantity is plotted as a function of the benefit from cooperation, *b.* The results in Fig. S3 show that ***stern-judging*** performs better than any of the other norms. Both *standing* and *simple-standing* lead to very similar performance, which reinforces the idea that second order norms are enough to promote cooperation under indirect reciprocity. Finally, image-scoring performs poorly compared to any of the other norms, a feature which is also related to the fact that the present analysis was carried out in the presence of errors.

The marginal advantage of ***stern-judging***, obtained via the present analysis, may not be enough to justify its ubiquity and insensitivity with respect to the mechanisms of selection between tribes as well as to the intensity of selection between tribes. We believe that, besides its excellent overall performance, ***stern-judging*** is more robust to invasion by other strategies, which gives it an evolutionary advantage with respect to other successful norms which promote cooperation.

References

1. Traulsen A, Nowak MA, Pacheco JM ((in press)) Stochastic dynamics of invasion and fixation. Physical Review E
2. Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. J Theor Biol 241: 233-240.

3. Crowley PH (2000) Hawks, doves, and mixed-symmetry games. J Theor Biol 204: 543-563.

4. Ohtsuki H, Iwasa Y (submitted) Global analysis of evolutionary dynamics and exhaustive search for social norms that maintain cooperation and reputation.

5. Ohtsuki H, Iwasa Y (2004) How should we define goodness?--reputation dynamics in indirect reciprocity. J Theor Biol 231: 107-120.

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

donor's reputation	**Recipient's reputation**	**donor's action**
GOOD	*GOOD*	**Y** / **N**
GOOD	*BAD*	**Y** / **N**
BAD	*GOOD*	**Y** / **N**
BAD	*BAD*	**Y** / **N**

Table S1. Bit-encoding of individual strategies. Each individual has a strategy encoded as a four-bit string. For each combination pair of donor and recipient reputations, the strategy prescribes individual's action. There are a total of 2^4=16 strategies, identified in Table S2.

strategy name	*GG*	*GB*	*BG*	*BB*
ALLD	**N**	**N**	**N**	**N**
1	**N**	**N**	**N**	**Y**
AND	**N**	**N**	**Y**	**N**
SELF	**N**	**N**	**Y**	**Y**
4	**N**	**Y**	**N**	**N**
5	**N**	**Y**	**N**	**Y**
6	**N**	**Y**	**Y**	**N**
7	**N**	**Y**	**Y**	**Y**
8	**Y**	**N**	**N**	**N**
9	**Y**	**N**	**N**	**Y**
CO	**Y**	**N**	**Y**	**N**
OR	**Y**	**N**	**Y**	**Y**
12	**Y**	**Y**	**N**	**N**
13	**Y**	**Y**	**N**	**Y**
14	**Y**	**Y**	**Y**	**N**
ALLC	**Y**	**Y**	**Y**	**Y**

Table S2. Different individual strategies in indirect reciprocity game. We identify the different strategies and how they determine the action of a donor (**N**=no, do not provide help, **Y**=yes, provide help), given the reputation pair donor/recipient. Whereas some of these strategies have assumed well-known designations in the literature, others remain named by their numeric order. This convention has been adopted in Fig. S1.

b	$\beta = 10^5$	$\beta = 10^4$	$\beta = 10^3$	$\beta = 10^2$	$\beta = 10^1$	$\beta = 10^0$
2	1001 1001	1□01 1001	1□01 100□	1□01 100□	1□01 100□	□□□□ □□□□
≥ 3	1001 1001	1001 1001	1001 1001	1001 1001	1001 1001	1001 1001

Table S3. Emergence of stern-judging for different intensities of selection. We carried out the bit-fixation analysis described in main text for the evolution of social norms under the pairwise comparison rule, for different values of the intensity of selection β . Intensity of selection decreases from left to right. Whereas for strong selection all norm bits fixate for b≥2, fixation becomes more difficult for b=2 as β decreases. Yet, in no case did we obtain fixation of a digit incompatible with stern-judging.

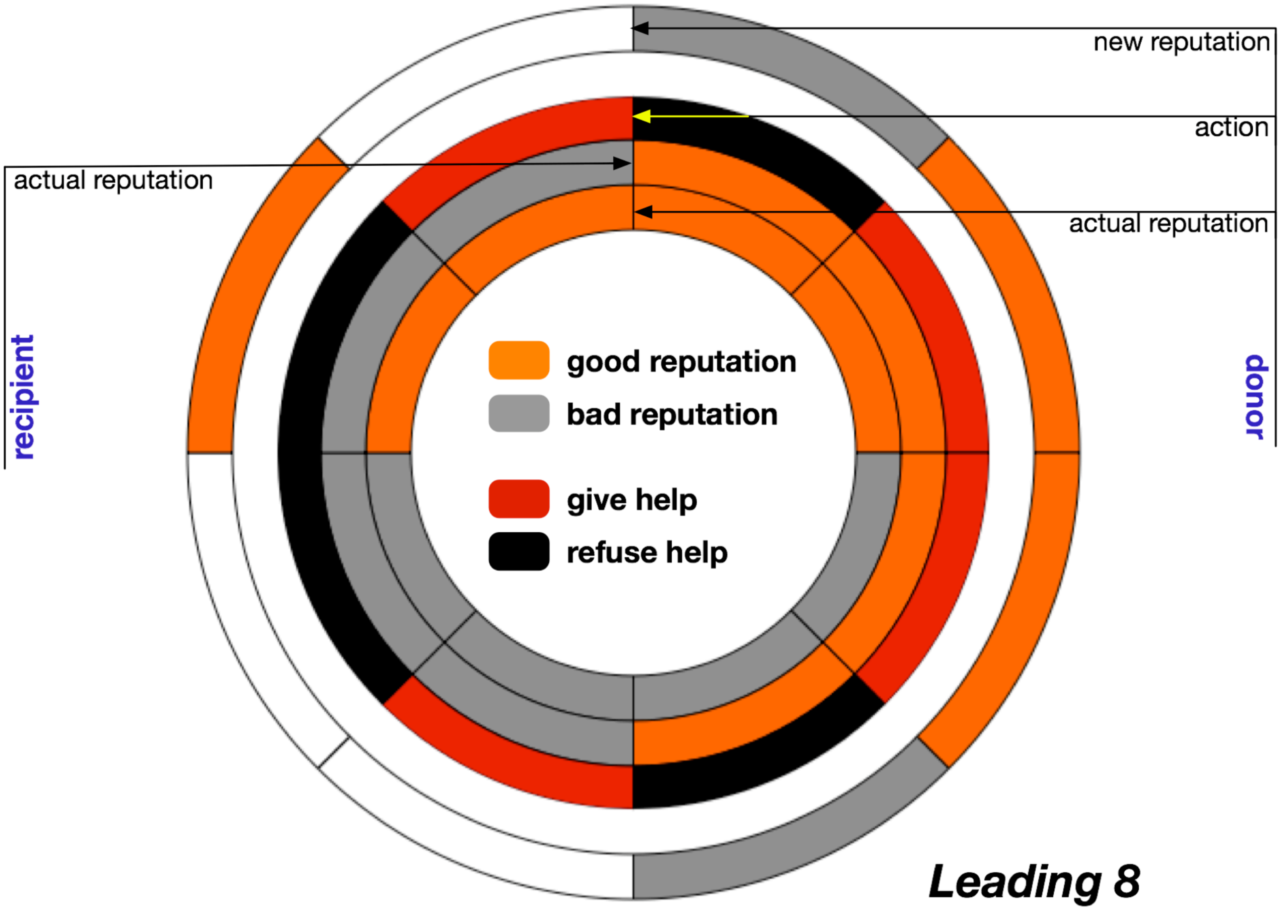

Figure S1. The Leading Eight Norms of Ohtsuki and Iwasa. We use the same notation as in Fig. 2, leaving in white those "slices" in the final norm which can be associated with either *GOOD* (orange) or *BAD* (grey) reputations. Note that, in this convention, second order norms exhibit a mirror symmetry with respect to the equatorial plane (disregarding the innermost layer). As a result, only two second order norms can incorporate the leading-eight – ***stern-judging*** and *simple-standing*, as recently coined by Ohtsuki and Iwasa – see Fig. S2 for details.

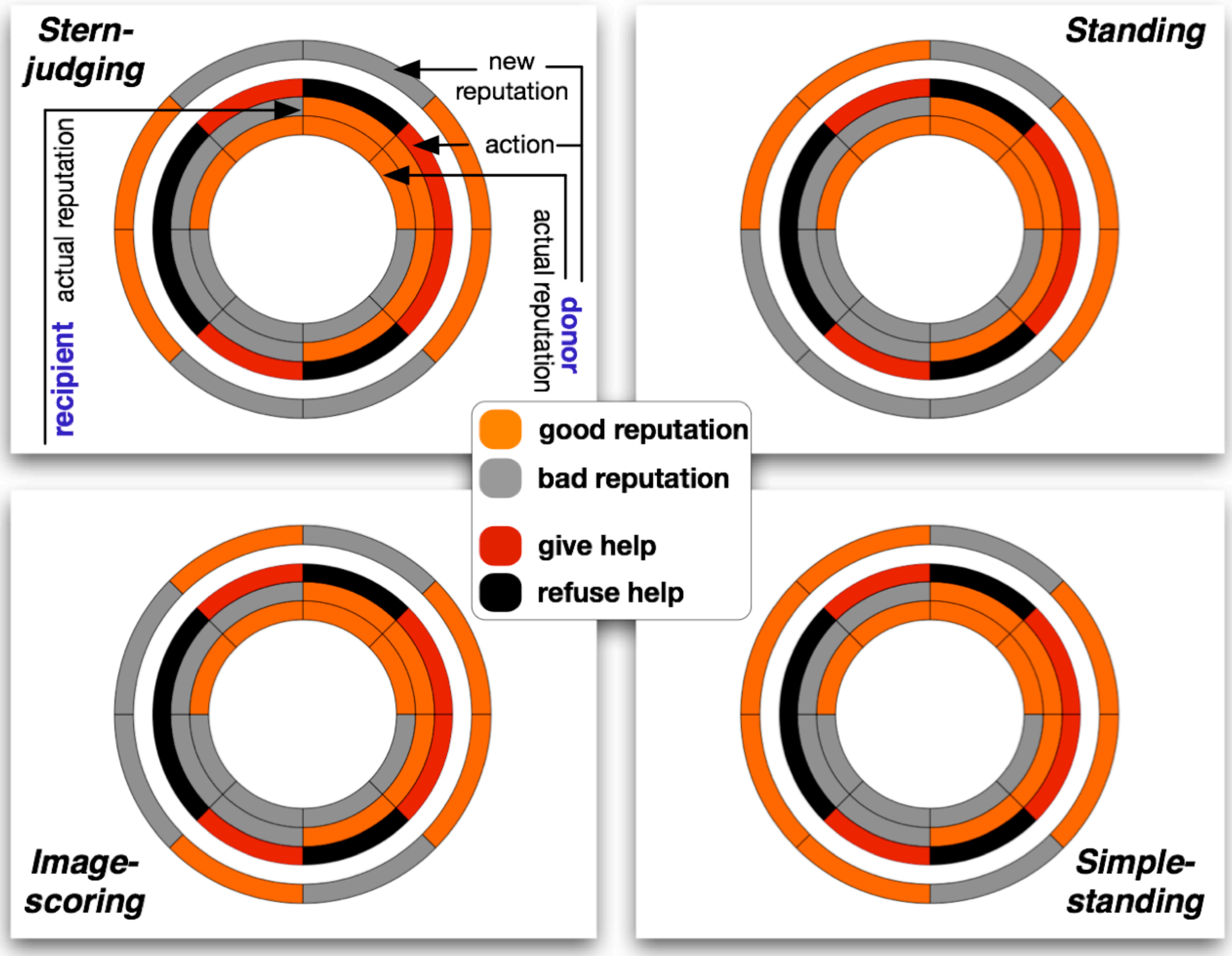

Figure S2. Four norms which promote cooperation. We depict the four norms, the performance of which we analysed. Both ***stern-judging***, *simple-standing* and *image-scoring* are symmetric with respect with the equatorial plane, and as such are second order norms. As for *standing*, it clearly breaks this symmetry, constituting a third order norm. In this representation, it is also clear that ***stern-judging***, exhibits the simplest symmetry of all norms.

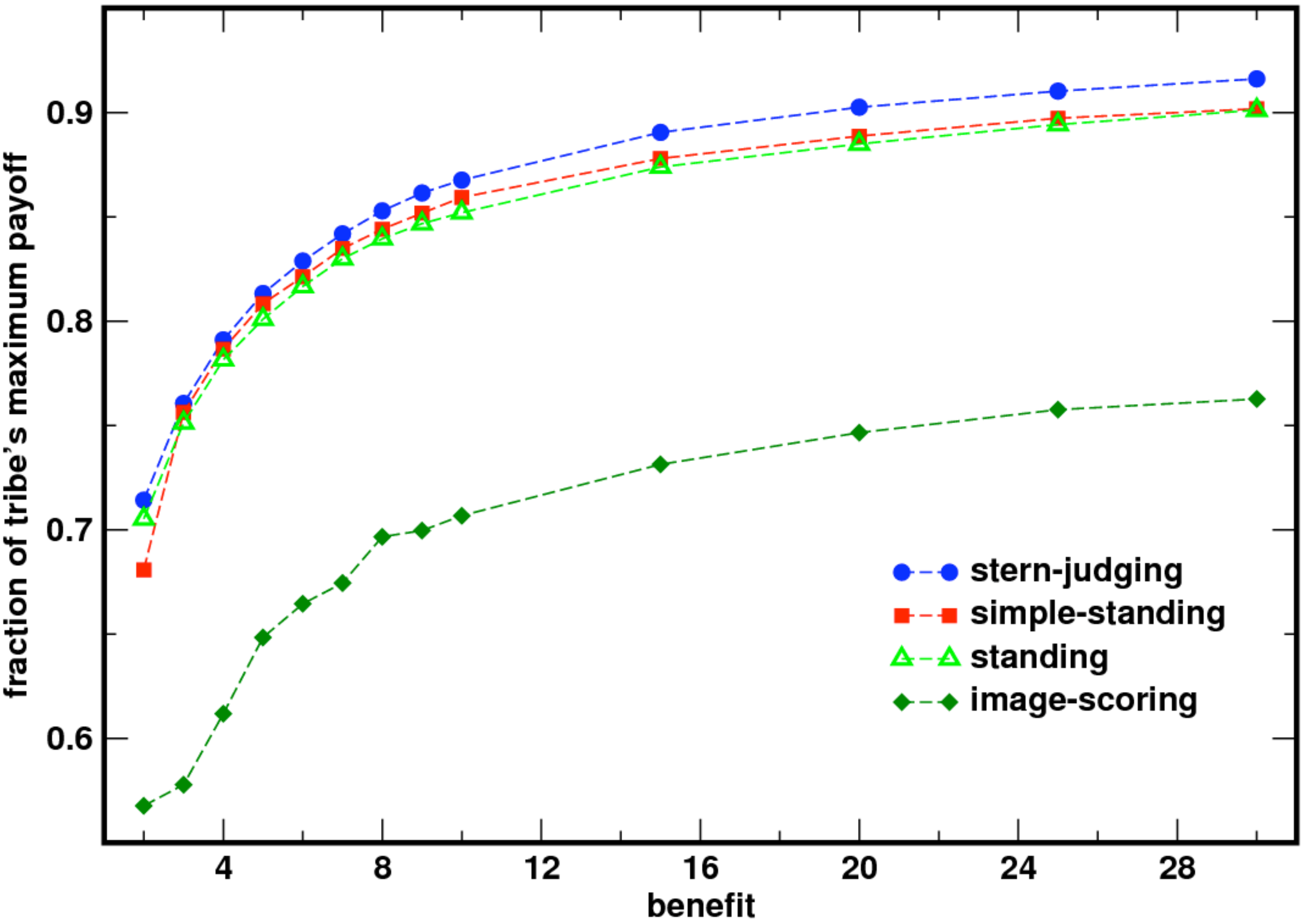

Figure S3. Individual performance of norms. We plot the ratio between the average payoff attained by each tribe under the influence of a single, fixed norm, and the maximum value possible, given the population size (64), the benefit from cooperation (*b*) and the cost of cooperation (*c=1*). Overall, ***stern-judging*** performs better than the other three norms of cooperation considered (cf. Fig. S2). For small values of *b,* the advantage is smaller than for larger values, but it is never superseded by any other norm. It is remarkable that *standing*, a third order norm, performs almost as well as *simple-standing*, a simpler, second-order norm. Finally, in all cases *image-score* is unable to match the performance of the other three norms. We ran 500 simulations for each tribe with 64 inhabitants, and used the last 1000 generations from a total of 10000 in each simulation to compute the average values depicted. We have included errors of execution as well as mutation of strategies.