LETTER

doi:10.1038/nature25763

Social norm complexity and past reputations in the evolution of cooperation

Fernando P. Santos[1,2], Francisco C. Santos[1,2] & Jorge M. Pacheco[2,3,4]

Indirect reciprocity is the most elaborate and cognitively demanding[1] of all known cooperation mechanisms[2], and is the most specifically human[1,3] because it involves reputation and status. By helping someone, individuals may increase their reputation, which may change the predisposition of others to help them in future. The revision of an individual's reputation depends on the social norms that establish what characterizes a good or bad action and thus provide a basis for morality[3]. Norms based on indirect reciprocity are often sufficiently complex that an individual's ability to follow subjective rules becomes important[4–6], even in models that disregard the past reputations of individuals, and reduce reputations to either 'good' or 'bad' and actions to binary decisions[7,8]. Here we include past reputations in such a model and identify the key pattern in the associated norms that promotes cooperation. Of the norms that comply with this pattern, the one that leads to maximal cooperation (greater than 90 per cent) with minimum complexity does not discriminate on the basis of past reputation; the relative performance of this norm is particularly evident when we consider a 'complexity cost' in the decision process. This combination of high cooperation and low complexity suggests that simple moral principles can elicit cooperation even in complex environments.

Under indirect reciprocity, an individual expects a return not from someone whom they have helped directly but from a third party. Helping (or not helping) the 'right' individuals can increase the chance of being helped by someone else at a later stage[9,10]. Ohtsuki and Iwasa[7,8,11] defined a binary world in which an individual's reputation can be either 'good' or 'bad'. Even in such a simple world, an arbitrarily large set of associated social norms can be used to classify decisions made in a donation game. In each instance of this donation game, involving a 'donor' and a 'recipient', the donor may either cooperate, helping the recipient at a cost c to themselves while conferring a benefit b to the recipient (with $b > c$), or defect (not providing help), in which case neither player incurs any costs or distributes any benefits. Everyone in the population uses the same social norm to assign public reputations to individuals. This reputation is attributed (errors aside; see Methods) and disseminated[12–14] by a bystander who witnesses a pairwise interaction. In this context, if all that matters for assigning a new reputation to the donor is their action towards the recipient[10], then we have a first-order norm. If the current reputation of the recipient matters as well as the action of the donor, then we obtain a second-order norm. A third-order norm additionally includes the current reputation of the donor.

Most norms studied so far reach up to third order (see ref. 15 for an exception) and therefore rely, at most, on the action of the donor and on the current reputations of both the donor and the recipient. For a norm of a given order, the information used by an observer to assign a new reputation is the same information that a donor may use to decide how to act towards a recipient. Consequently, studies of indirect reciprocity involving norms of increasing order typically use behavioural strategies (often designated action rules) and strategy spaces that also increase (exponentially with order). For this reason, a combination of a norm and a strategy that promotes cooperation in the space of nth-order norms does not necessarily perform equally well in a space of higher-order norms because the availability of more complex behaviours (together with those for lower-order norms) often has non-trivial effects on cooperation[16]. Furthermore, the performance of a complex social norm can be constrained by an individual's ability to follow complex subjective rules[4–6]. This raises two fundamental questions: (1) whether the moral principles that underlie successful strategies and norms in the space of third-order norms remain valid within a larger space, and if so which ones; and (2) how the cognitive skills associated with social norms and strategies impair individuals' performance. Using the donation game and binary reputations we answer these questions by investigating the cooperative capacity of social norms in a space that encompasses norms of up to fourth order and that span a wide range of cognitive complexities[4,17,18]. Increasing the number of possibilities to consider when assigning a good or a bad reputation to individuals enables us to identify the key pattern of social norms that provides the necessary conditions for promoting cooperation.

Fourth-order norms additionally incorporate (on top of the features of third-order norms) the previous reputation of the recipient, requiring individuals with increased memory capabilities and that are therefore able to enact more elaborate behaviours. We encode norms up to fourth order and corresponding strategies as 16- and 8-bit tuples, respectively; consequently, there are 2^{16} different norms and 2^8 different strategies that individuals may use when playing the donation game described above (see Methods for details). Furthermore, we define the complexity of a norm using the index κ, which describes the number of literals (that is, the logic variables and their complements) in the shortest logical expression that can define the norm (see Methods). This index has been used previously to describe an individual's ability to learn a concept[4,17]. Here, the simplest norm has $\kappa = 0$ and the most complex norm has $\kappa = 32$. In Fig. 1 we illustrate norms of different orders and complexities, providing intuitive representations of the raw information in Supplementary Table 4. Norms of the same order may have different complexities, as demonstrated for second-order norms in Fig. 1: different reputation tables (corresponding to different norms) translate to different numbers of literals in the corresponding minimal logical expressions. Moreover, similarly to norms, strategies also exhibit an intrinsic complexity (κ_s) that can influence their adoption. Equipped with these tools, we investigate which norms promote the emergence of cooperation. In Methods, we describe computer simulations of the evolutionary dynamics, in which individuals in a population, each starting with a random strategy, play the donation game with their peers. Throughout the game, the players change strategies via social learning[19], whereby strategies with higher fitness are adopted more frequently[20]. The simulations return the cooperation index η, a real number between 0 and 1 that

[1]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016 Porto Salvo, Portugal. [2]ATP-group, 2744-016 Porto Salvo, Portugal. [3]Centro de Biologia Molecular e Ambiental, Universidade do Minho, 4710-057 Braga, Portugal. [4]Departamento de Matemática e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal.

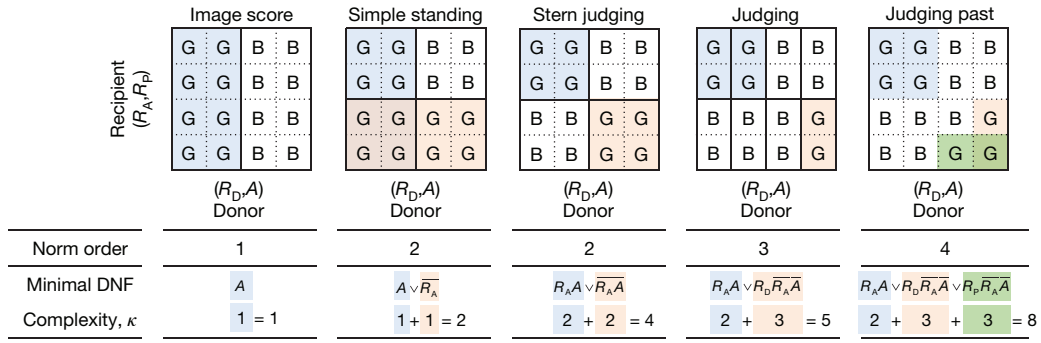


Figure 1 | Norm complexity. A norm is represented by a 'reputation table'. Each entry in each table indicates the new reputation of the donor (good, G; bad, B), assigned on the basis of their current reputation ($R_D \in \{G, B\}$), their action ($A \in \{C, D\}$, where C denotes cooperation and D defection), and the current ($R_A \in \{G, B\}$) and past ($R_P \in \{G, B\}$) reputations of the recipient. Rows are ordered, from top to bottom, as (G,G), (G,B), (B,B), (B,G) and columns are ordered, from left to right, as (G,C), (B,C), (B,D), (G,D). The complexity κ is determined by counting the number of literals of the shortest logical expression (the minimal disjunctive normal form (DNF), where A denotes $A = C$ and $\overline{A}$ denotes the complement ($\overline{A} = D$), and similarly $R_{A,D}$ and $\overline{R_{A,D}}$ denote G and B; see Methods) that can be used to prescribe a donor reputation of 'G'. Alternatively, κ can be determined by counting the number of blocks of 2^k 'G's[30] (where k is chosen to be as large as possible and blocks can overlap; see coloured squares and rectangles): each block of 2^k 'G's increases κ by $4 - k$ (starting from $\kappa = 0$). See Supplementary Information for further details.

describes the average number of interactions that lead to donations as a fraction of the total number of interactions observed in a population that evolves under a given social norm.

In Fig. 2 we compare η for the leading eight norms shown[7,8] to stabilize cooperation (in the sense discussed in Supplementary Information, section 1.4) under indirect reciprocity at third order, in the space of third-order (blue bars) and fourth-order (red bars) norms. The results show that when more elaborate strategies become possible (when up to fourth-order norms are considered) only a subset of the leading eight norms still fosters similar levels of cooperation as in the third order space. Overall, about 0.2% of the 2^{16} norms in fourth-order space lead to $\eta > 0.9$, compared to about 2% of the 2^8 norms in third-order space (Extended Data Fig. 1). Many 'new' fourth-order norms (that is, those that cannot be represented in lower-order spaces) foster high levels of cooperation. Of the leading two second-order norms[21,22] (stern judging and simple standing; see Supplementary Information for details), only stern judging remains highly cooperative in fourth-order space. This norm can be stated as: "help good people and refuse help otherwise, and we shall be nice to you; otherwise, you will be punished."[23]

Next, we investigate the role of norm complexity in promoting cooperation by plotting the cooperation level (η) of the norm that leads to maximum cooperation for a given complexity (κ). Figure 3 demonstrates that the highest values of η are attained by norms with complexities as low as $\kappa = 4$. The same happens even when individuals incur a complexity cost $c_c = \gamma\kappa_s$ when using a strategy of complexity κ_s (where γ is a real constant; see Extended Data Figs 2 and 3 and Supplementary Information for details; we also demonstrate that these results remain valid when the past reputation of the donor instead of the recipient is used in defining fourth-order norms).

Figure 3 demonstrates that for $\kappa > 4$ only fourth-order norms maximize η, despite the fact that the complexity of norms of the same order can vary substantially (see Fig. 1). Consequently, taking complexity into account opens up new questions regarding the features that make fourth-order norms successful, and the features of the third- and

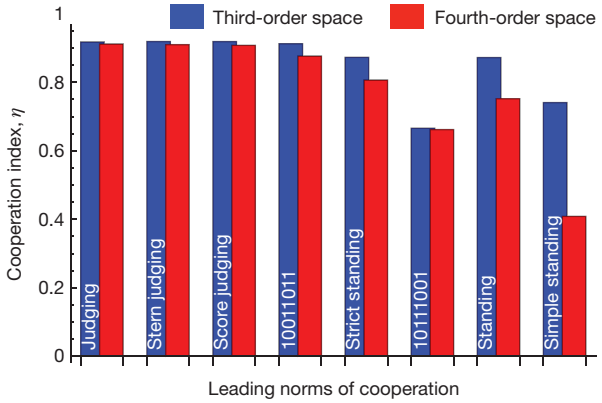


Figure 2 | Cooperation index of leading norms. When the space of the norms (and strategies) is extended from third-order (blue bars) to fourth-order (red bars), some of the leading eight norms of cooperation[8] (in third-order space)—and particularly simple standing (which, together with stern judging, make up the leading two norms in second-order space[21])—no longer promote cooperation. See Extended Data Fig. 1 for results involving all norms. The model parameters used (see Methods for definitions) are $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$, $c = 1$ and $\gamma = 0$. The results are qualitatively insensitive to the ratio b/c, to the population size, to any errors in assessment or assignments made by individuals and to different mutation schemes (see Methods and Extended Data Figs 4, 5). See Fig. 1 and Supplementary Table 4 for definitions and characterization of norms; unnamed norms are defined by their binary representation in third-order space (see Methods).

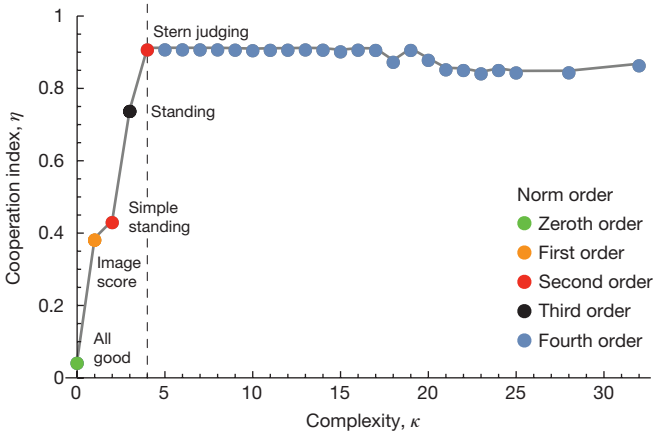


Figure 3 | Cooperation index versus norm complexity. Maximal levels of cooperation ($\eta > 0.9$) are attained under the simple norm stern judging ($\kappa = 4$). More complex norms ($\kappa > 4$) do not lead to higher levels of cooperation. Some well-known norms that maximize η for a given κ are identified. In Extended Data Fig. 2 we show the dependence of η on κ when a complexity cost is imposed on strategies and the past reputation of the donor is considered instead of that of the recipient. The model parameters used (see Methods for definitions) are $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$, $c = 1$ and $\gamma = 0$. See Extended Data Figs 4 and 5 for robustness analysis. See Fig. 1 and Supplementary Table 4 for definitions and characterization of norms.

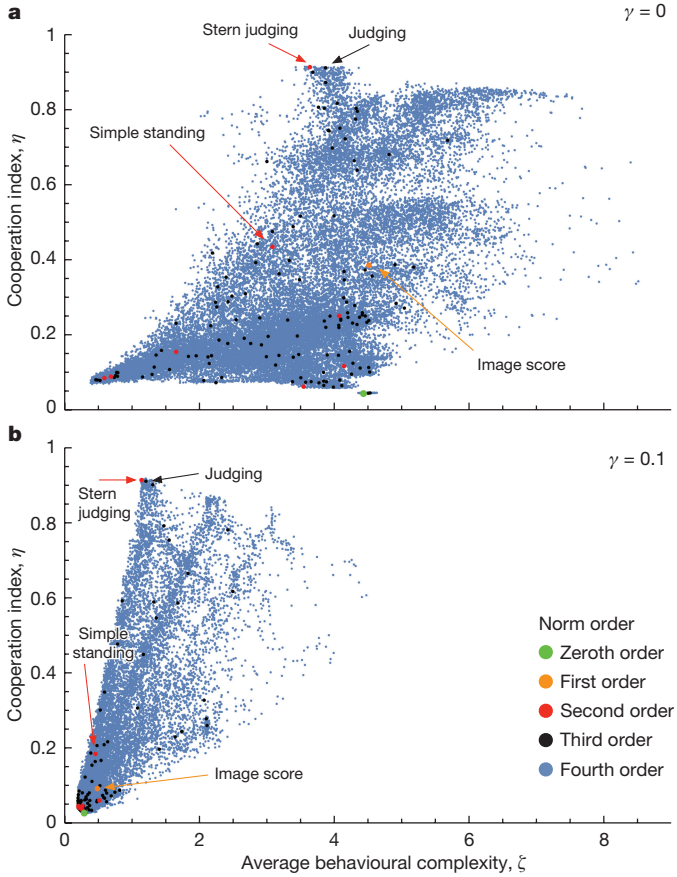


Figure 4 | Average behavioural complexity. Norms induce different levels of strategic complexity κ_s in a population. Because the simplest strategies ignore reputations, and are thus unable to secure cooperation, we expect high levels of cooperation to require some average behavioural complexity ζ. **a**, We find that stern judging and judging lead to maximum values of η at relatively low values of ζ. **b**, This finding is emphasized if we consider a strategic complexity cost $c_c = \gamma\kappa_s$ with $\gamma \neq 0$ (see Extended Data Fig. 2). The model parameters used, detailed in Methods, are $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$ and $c = 1$. See Fig. 1 for definitions of judging, stern judging, simple standing and image score.

second-order norms that ensure (or not) their capacity to sustain cooperation in the more complex fourth-order space.

To address these questions, we conducted an exhaustive search in the space of fourth-order norms and identified (for a specific set of model parameters) a recurrent pattern common to the fourth-order norms that promote cooperation (see Supplementary Tables 1 and 2). This pattern states that the bystander assigns a 'good' label to donors that either (i) cooperate with enduring good individuals or (ii) are already good and defect against enduring bad individuals, and assigns a 'bad' label to those who act otherwise in these contexts (that is, who defect against enduring good individuals or who are good but cooperate with enduring bad individuals). Here, enduring individuals are those who retain the same good or bad label in the present and in the past. The pattern can therefore be summarized by the following rule: "donors become good (bad) if they help (refuse to help) an enduring good individual; they maintain (lose) their good label if they refuse to help (help) an enduring bad individual."

This rule has immediate implications at lower orders. Only four of the leading eight norms[8] in third-order space comply with this fourth-order rule—those that promote the highest levels of cooperation (Fig. 2). Not surprisingly (see Fig. 3), stern judging is the only one of the leading two[21] norms in second-order space that complies (simple standing violates the rule by prescribing a good reputation whenever a player helps an enduring bad individual).

In Fig. 3 we show that stern judging leads to a maximal value of η ($\eta > 0.9$), while having a κ value less than that of any third- or fourth-order social norm that leads to comparable values of η (see also Extended Data Fig. 2). Furthermore, strategies that prevail under stern judging are remarkably simple. We demonstrate this by first computing the complexity κ_s of the prevalent strategies under each norm. Subsequently, we compute the (norm-dependent) fraction of time that each individual spends adopting each strategy and calculate the weighted average complexity of the strategies used, which we designate by the average behavioural complexity (ζ). In Fig. 4 we depict all norms in fourth-order space by plotting η as a function of ζ. Stern judging (a second-order norm), judging and score judging (third-order norms; see Supplementary Table 4) lead to high η using strategies with low ζ (Fig. 4a)—a feature that is maintained in the presence of a complexity cost $c_c = \gamma\kappa_s$ (Fig. 4b).

Our results show that cooperation under indirect reciprocity can emerge even when the cognitive capacity of individuals is limited. In this context, it becomes clear why stern judging proves to be so robust, remaining the most successful norm (in terms of the combination of high cooperation and low complexity) in all norm spaces studied even when considering populations of different sizes (from small-scale societies to large communities of individuals[22]). It is the norm of lowest order and complexity that is compatible with the pattern described here, requiring little cognitive skill both in assigning reputations and in inducing behaviours that lead to high levels of cooperation. It is therefore not surprising that the fingerprint of stern judging is present in the moral judgment of toddlers (as young as five months old[24]), who show a preference not only for individuals who helped others, but also for individuals who harmed those who hindered others[25].

The modelling approach used here can also be informative when designing pervasive reputation systems[26], in which optimality should be combined with simplicity. Game-theoretical models have been used to study reputation systems in the context of trading platforms, crowdsourcing markets and peer-to-peer systems[27–29]. It has been shown that very simple and intuitive social norms may suffice to promote cooperation[28] and that publicizing a detailed account of a seller's feedback history—as compared with only the most recent rating—does not improve cooperation in online trading platforms[27]. Both of these features—simplicity and the irrelevance of history—bear similarity to the results presented here, despite the fact that our model would need to be modified to be applicable to reputation systems in online platforms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 September 2017; accepted 15 January 2018.

1. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437,** 1291–1298 (2005).
2. Rand, D. G. & Nowak, M. A. Human cooperation. *Trends Cogn. Sci.* **17,** 413–425 (2013).
3. Alexander, R. D. *The Biology of Moral Systems* (Transaction Publishers, 1987).
4. Feldman, J. Minimization of Boolean complexity in human concept learning. *Nature* **407,** 630–633 (2000).
5. Chater, N. & Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7,** 19–22 (2003).
6. Feldman, J. The simplicity principle in perception and cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **7,** 330–340 (2016).
7. Ohtsuki, H. & Iwasa, Y. How should we define goodness?—reputation dynamics in indirect reciprocity. *J. Theor. Biol.* **231,** 107–120 (2004).
8. Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239,** 435–444 (2006).
9. Brandt, H. & Sigmund, K. The logic of reprobation: assessment and action rules for indirect reciprocation. *J. Theor. Biol.* **231,** 475–486 (2004).
10. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393,** 573–577 (1998).
11. Ohtsuki, H., Iwasa, Y. & Nowak, M. A. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457,** 79–82 (2009).
12. Dunbar, R. *Grooming, Gossip, and the Evolution of Language* (Harvard Univ. Press, 1998).

13. Sommerfeld, R. D., Krambeck, H.-J., Semmann, D. & Milinski, M. Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc. Natl Acad. Sci. USA* **104,** 17435–17440 (2007).
14. Skyrms, B. *Signals: Evolution, Learning and Information* (Oxford Univ. Press, 2010).
15. Kandori, M. Social norms and community enforcement. *Rev. Econ. Stud.* **59,** 63–80 (1992).
16. Stewart, A. J., Parsons, T. L. & Plotkin, J. B. Evolutionary consequences of behavioral diversity. *Proc. Natl Acad. Sci. USA* **113,** E7003–E7009 (2016).
17. Wegener, I. & Teubner, B. *The Complexity of Boolean Functions* Vol. 1 (B. G. Teubner, 1987).
18. McCluskey, E. J. Minimization of Boolean functions. *Bell Labs Tech. J.* **35,** 1417–1444 (1956).
19. Rendell, L. *et al.* Why copy others? Insights from the social learning strategies tournament. *Science* **328,** 208–213 (2010).
20. Sigmund, K. *The Calculus of Selfishness* (Princeton Univ. Press, 2010).
21. Ohtsuki, H. & Iwasa, Y. Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J. Theor. Biol.* **244,** 518–531 (2007).
22. Santos, F. P., Santos, F. C. & Pacheco, J. M. Social norms of cooperation in small-scale societies. *PLOS Comput. Biol.* **12,** e1004709 (2016).
23. Pacheco, J. M., Santos, F. C. & Chalub, F. A. C. Stern-judging: a simple, successful norm which promotes cooperation under indirect reciprocity. *PLOS Comput. Biol.* **2,** e178 (2006).
24. Hamlin, J. K. Moral judgment and action in preverbal infants and toddlers evidence for an innate moral core. *Curr. Dir. Psychol. Sci.* **22,** 186–193 (2013).
25. Hamlin, J. K., Wynn, K., Bloom, P. & Mahajan, N. How infants and toddlers react to antisocial others. *Proc. Natl Acad. Sci. USA* **108,** 19931–19936 (2011).
26. Resnick, P., Kuwabara, K., Zeckhauser, R. & Friedman, E. Reputation systems. *Commun. ACM* **43,** 45–48 (2000).
27. Dellarocas, C. Reputation mechanism design in online trading environments with pure moral hazard. *Inform. Syst. Res.* **16,** 209–230 (2005).
28. Ho, C.-J., Zhang, Y., Vaughan, J. & Van Der Schaar, M. *Towards Social Norm Design for Crowdsourcing Markets.* Report No. WS-12-08 (AAAI, 2012).
29. Zhang, Y. & van der Schaar, M. Peer-to-peer multimedia sharing based on social norms. *Signal. Process. Image Commun.* **27,** 383–400 (2012).
30. Karnaugh, M. The map method for synthesis of combinational logic circuits. *Trans. AIEE Part I* **72,** 593–599 (1953).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia (FCT) through grants SFRH/BD/94736/2013, PTDC/EEI-SII/5081/2014, PTDC/MAT/STA/3358/2014, UID/BIA/04050/2013 and UID/CEC/50021/2013. We are grateful to A. P. Francisco and M. Janota for comments.

Author Contributions F.P.S., F.C.S. and J.M.P. conceived the project. F.P.S. performed the mathematical and numerical analysis. F.P.S., F.C.S. and J.M.P. analysed the results and wrote the paper. All authors contributed to all other aspects of the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.M.P. (jmpacheco@math.uminho.pt).

Reviewer Information *Nature* thanks C. Efferson, E. Fehr, G. Szabó and A. Tavoni for their contribution to the peer review of this work.

METHODS

Here we summarize the model and mathematical methods; further details are provided in Supplementary Information.

Actions conditional on reputations. The action of the donor in each interaction depends on the current reputation of the donor (R_D) and the recipient (R_A), together with the past reputation of the recipient (R_P). Assuming binary reputations (1 = 'good' = G or 0 = 'bad' = B), the strategy used by each player is an 8-bit string that prescribes an action (1 = 'cooperate' = C or 0 = 'defect' = D) on the basis of the aforementioned reputations. We extend previously used notation[7,8,21] to denote each strategy by a tuple $P = (p_0, p_1, p_2, p_3, p_4, p_5, p_6, p_7)$, in which $p_i \in \{0, 1\}$ denotes the action of the donor for each of the possible combinations of reputations R in the order R_P, R_D, R_A (that is, with R_P (R_A) being the most (least) significant bit when defining a position within a strategy), and with $R_i = 1$ considered before $R_i = 0$ (that is, for example, p_0 corresponds to $R_P = R_D = R_A = G = 1$ and p_7 to $R_P = R_D = R_A = B = 0$); this yields 2^8 different strategies. We consider execution errors (ε) that represent the inability of individuals to act in the way that their strategy dictates[31]. It is common practice to consider errors in the form of 'failed intended cooperation'[21,32] due, for instance, to an individual's lack of resources, time or energy available to donate in their role as donor[33]. Our results remain valid even if the execution errors additionally induce defectors to involuntarily cooperate.

Social norms. We consider that the new reputation of an acting individual follows a norm that can be written as a tuple $d = (d_0, d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{14}, d_{15})$, in which $d_i \in \{0, 1\}$ denotes the new reputation assigned to the donor for each of the possible combinations of action A and reputations R in the order R_P, R_D, R_A, A (that is, with R_P (A) being the most (least) significant bit when defining a position within a norm). For convenience, we use R_P, R_D and R_A both as the names of a reputation layer in a norm (see Extended Data Fig. 3 and Supplementary Table 3) and as a Boolean variable that can assume the values $1 = G = R$ and $0 = B = \overline{R}$. Similarly, $1 = C = A$ and $0 = D = \overline{A}$. As stated in the main text (see Fig. 1), there are 2^{16} social norms up to fourth order. We consider assignment errors[8] α that occur when the observer fails to assign the correct reputation. We assume that, once the reputation of an individual is assigned, it is widely disseminated throughout the population (for example through gossiping[11–14]), so that everyone shares the same opinion regarding the reputation of others. However, we include errors at the level of individuals, when retrieving the public reputation of others, which occur with a probability χ: whenever these errors occur, an individual may perform the wrong action as a donor or assign the wrong (public) reputation as a bystander.

Complexity. Social norms and individual strategies can both be regarded as Boolean functions that determine: (1) when an individual has a good reputation (G; social norms), or (2) when the appropriate action is to cooperate (C; strategies or action rules). These functions take the Boolean inputs A (action of the donor is C), R_A (current reputation of the recipient is G), R_P (past reputation of the recipient or donor is G) and R_D (current reputation of the donor is G). For instance, the well-known second-order discriminator strategy whereby an individual cooperates with only those players who have a G reputation is given by $P = (1, 0, 1, 0, 1, 0, 1, 0)$, or by the Boolean function R_A. The fourth-order discriminator strategy, whereby an individual cooperates only if an opponent has a G reputation both in the present and in the past, can be written as $P = (1, 0, 1, 0, 0, 0, 0, 0)$ or $R_A \wedge R_P$. In the context of social norms, the 'image score' norm $d = (1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)$ corresponds to R_A, and the 'stern judging' norm $d = (1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1)$ can be written as $(R_A \wedge A) \vee (\overline{R_A} \wedge \overline{A})$. The complexity of a norm or strategy (κ or κ_s) is the length of the shortest Boolean formula (here in disjunctive normal form (DNF); that is, a sum of products) that is logically equivalent to the corresponding Boolean function[4,17]. This quantity is also known as the Boolean complexity[4,17]. To calculate the Boolean complexity of a norm or strategy, we generate and simplify the corresponding DNF and count the number of literals that it includes. We apply a standard algorithm to minimize Boolean functions (the Quine–McCluskey algorithm[18]), using the version implemented in Mathematica (Wolfram) through the function *BooleanMinimize*. This algorithm generates a DNF with a minimum number of literals but that is logically equivalent to the original (full) DNF—the minimal DNF (see Fig. 1). Here we focus on the minimal DNF representation of a logic expression. However, other representations could be devised in which, in some cases, there is departure from a minimal DNF and the number of literals is reduced slightly—such as by applying De Morgan's laws and/or the distributive law of Boolean algebra[18]. In fact, reaching a minimal Boolean function is a computational challenge[34], and for this reason it is often calculated as an approximation[4,35,36]. By adopting a complexity measure based on the number of literals of a minimal DNF form, we provide an upper bound on the Boolean complexity of each social norm, while ensuring computational tractability and an easy generalization to norms of higher order. In Supplementary Information we define (and provide an example of) the three-step process that we use to compute κ for any norm (and κ_s for any strategy).

In Fig. 1 we also provide an alternative visual method to determine κ. It relies on counting the number of different blocks of 'G's of size 2^k, a method that is associated with so-called Karnaugh maps[30] (a graphical method for simplifying logic circuits): a size-2^3 block contributes 1 to the complexity; a size-2^2 block contributes 2; a size-2^1 block contributes 3; and a size-2^0 block contributes 4. In general, a 2^k-size G block contributes $4 - k$ to κ. Some rules apply when defining G blocks[37]: they must contain only G values, being formed by joining adjacent cells (diagonal links do not count); torus boundary conditions apply; and they must be the largest possible size. Importantly, the choice of row and column order in defining the reputation table in Fig. 1 is not arbitrary: the entries in two adjacent rows or columns must differ only by one bit.

It is also worth pointing out that Fig. 1 provides visual cues that show the symmetries of a reputation table that are associated with a norm of a given order; for example, for norms of order one, all of the entries of the left and right eight-entry blocks are identical. In all cases in Fig. 1, blocks of entries are delimited by solid lines: norms of second order have four blocks that each contain four identical entries; norms of third order have eight blocks that each contain two identical entries; and norms of fourth order have no such blocks in which multiple identical entries can be identified.

Evolutionary dynamics. In the computer simulations, evolution proceeds in discrete steps. At the beginning of one simulation (or run), each individual adopts one of the 2^8 (256) possible strategies, chosen using a uniform probability distribution (UPD). Individual reputations, both present and past, are also assigned using a UPD. Each simulation is executed for a large number g of generations. In each generation, Z individuals selected using a UPD revise their strategy. After selecting one of the Z individuals (say, individual X), strategy revision can happen through mutation or imitation. Mutation[38] happens with probability μ: a new strategy is adopted randomly (UPD) out of the 256 possible. This approach allows us to study the evolutionary robustness of strategies against the invasion of others[39–42] (see Supplementary Information). Alternatively, we consider a bit-wise (or local) mutation (see Extended Data Fig. 5), which leads to similar results. Imitation happens with probability $1 - \mu$: a new individual (say, individual Y, the role model) is selected randomly, and individual X is given the opportunity to update their strategy. The fitness of both individuals (F_X and F_Y) is calculated as the average payoff earned in $g = 2Z$ games played against individuals in the population selected randomly using a UPD. This number of games is adequate to obtain a clear assessment of the average payoff, given the number of strategies, and to account for the dynamic reputation assignment described below. After each game is played, a reputation update occurs according to the social norm and subject to the assessment (α) and private (χ) errors described above. Individual X adopts the strategy of individual Y with probability $(1 + e^{F_X + F_Y})^{-1}$—the so-called Fermi update or pairwise comparison rule[43].

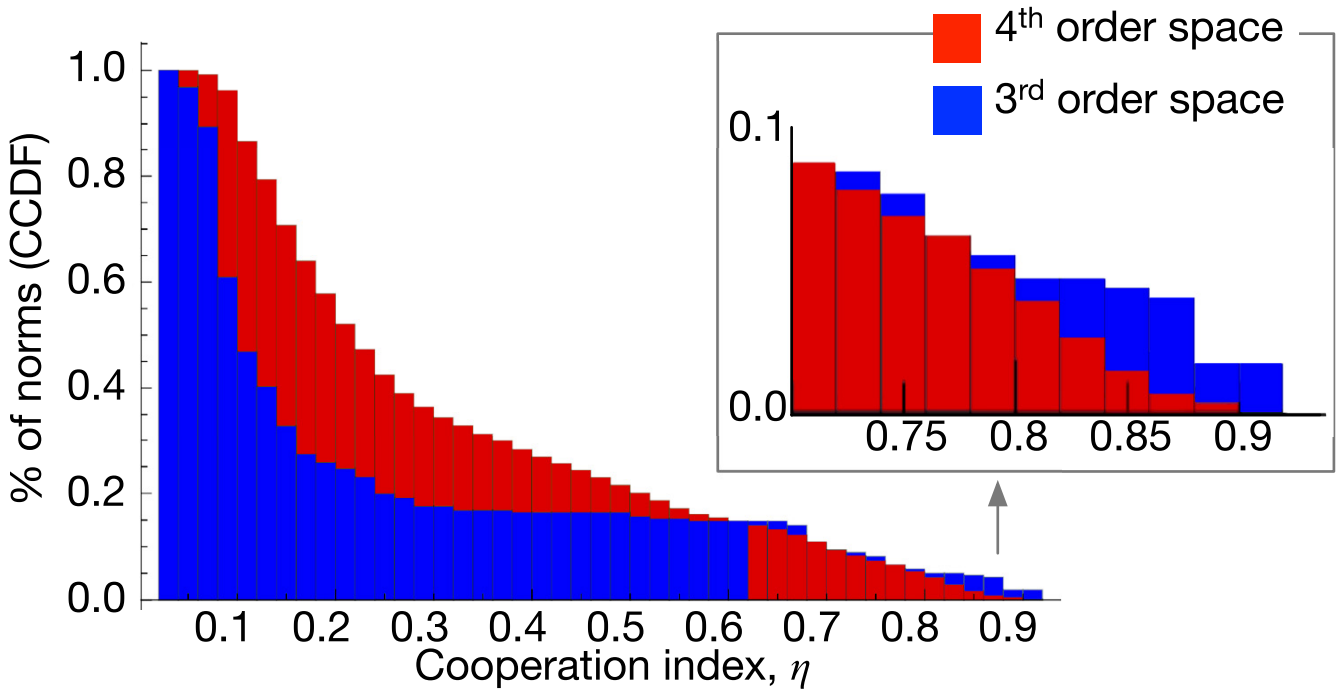
Cooperation index. The cooperation index η for a given social norm is computed as the fraction of cooperative acts that take place out of the total number of acts during the simulation time. Thus, η reflects both the dependence of strategy adoption on the relative frequency of strategies present in the population (frequency-dependent selection) and the evolution of reputations given the fixed social norm in the population. More details on the computer simulations are provided in Supplementary Information and in Extended Data Fig. 6. The full set of parameters explored is summarized in Supplementary Table 2.

Code availability. A comprehensive description of the standard algorithms that we implemented to compute the evolutionary dynamics of strategies is provided in Supplementary Information and Extended Data Fig. 6. Code that exemplifies the calculation of Boolean complexity is available at https://doi.org/10.5281/zenodo.1041379.

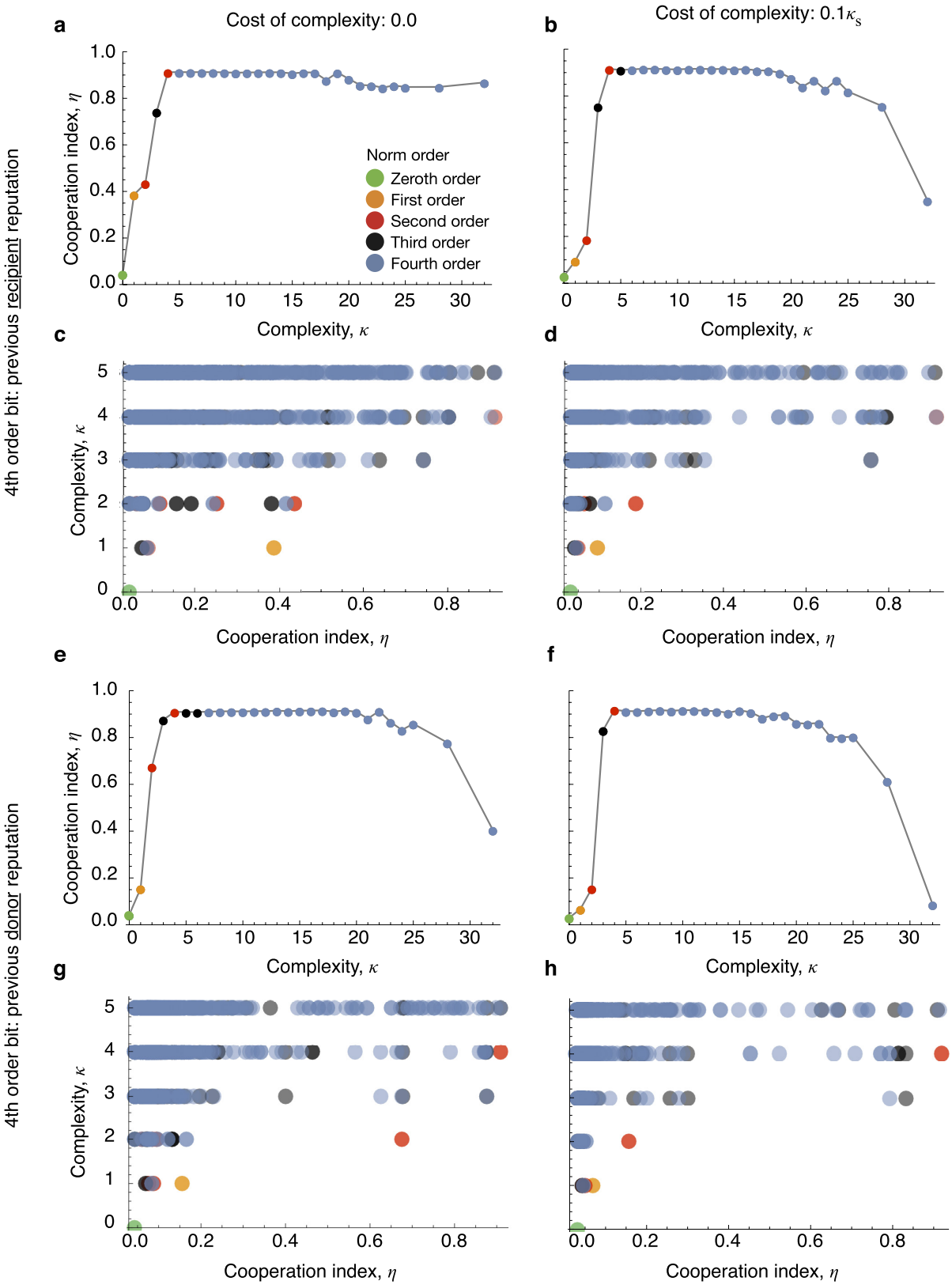
Data availability. The raw data generated, which were used to create Figs 2–4 and which support our conclusions, is available with the online version of the paper as Source Data.

31. Fishman, M. A. Indirect reciprocity among imperfect individuals. *J. Theor. Biol.* **225,** 285–292 (2003).
32. Roberts, G. Evolution of direct and indirect reciprocity. *Proc. R. Soc. Lond. B* **275,** 173–179 (2008).
33. Sherratt, T. N. & Roberts, G. The importance of phenotypic defectors in stabilizing reciprocal altruism. *Behav. Ecol.* **12,** 313–317 (2001).
34. Umans, C. The minimum equivalent DNF problem and shortest implicants. *J. Comput. Syst. Sci.* **63,** 597–611 (2001).
35. Vigo, R. A note on the complexity of Boolean concepts. *J. Math. Psychol.* **50,** 501–510 (2006).
36. Feldman, J. The simplicity principle in human concept learning. *Curr. Dir. Psychol. Sci.* **12,** 227–232 (2003).

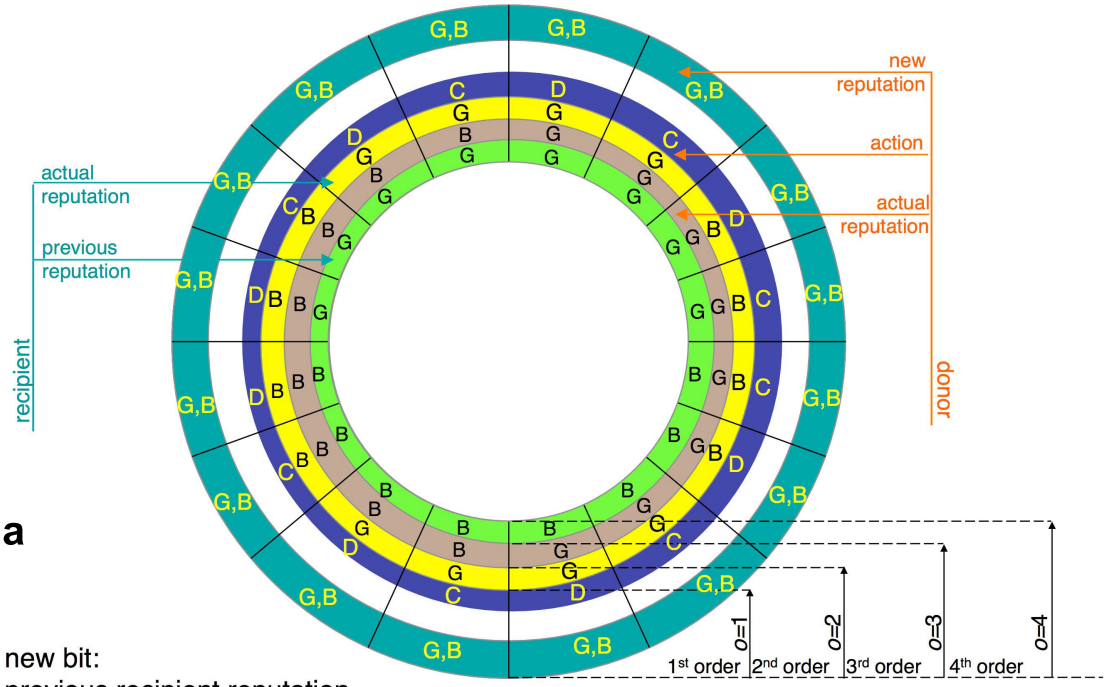
37. Null, L. & Lobur, J. *The Essentials of Computer Organization and Architecture* Ch. 3 (Jones & Bartlett Publishers, 2014).
38. Santos, F. P., Pacheco, J. M. & Santos, F. C. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* **6,** 37517 (2016).
39. Stewart, A. J. & Plotkin, J. B. From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proc. Natl Acad. Sci. USA* **110,** 15348–15353 (2013).
40. Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proc. Natl Acad. Sci. USA* **111,** 17558–17563 (2014).
41. Pinheiro, F. L., Vasconcelos, V. V., Santos, F. C. & Pacheco, J. M. Evolution of all-or-none strategies in repeated public goods dilemmas. *PLOS Comput. Biol.* **10,** e1003945 (2014).
42. Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K. & Nowak, M. A. Memory-*n* strategies of direct reciprocity. *Proc. Natl Acad. Sci. USA* **114,** 4715–4720 (2017).
43. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys. Rev. E* **74,** 011909 (2006).

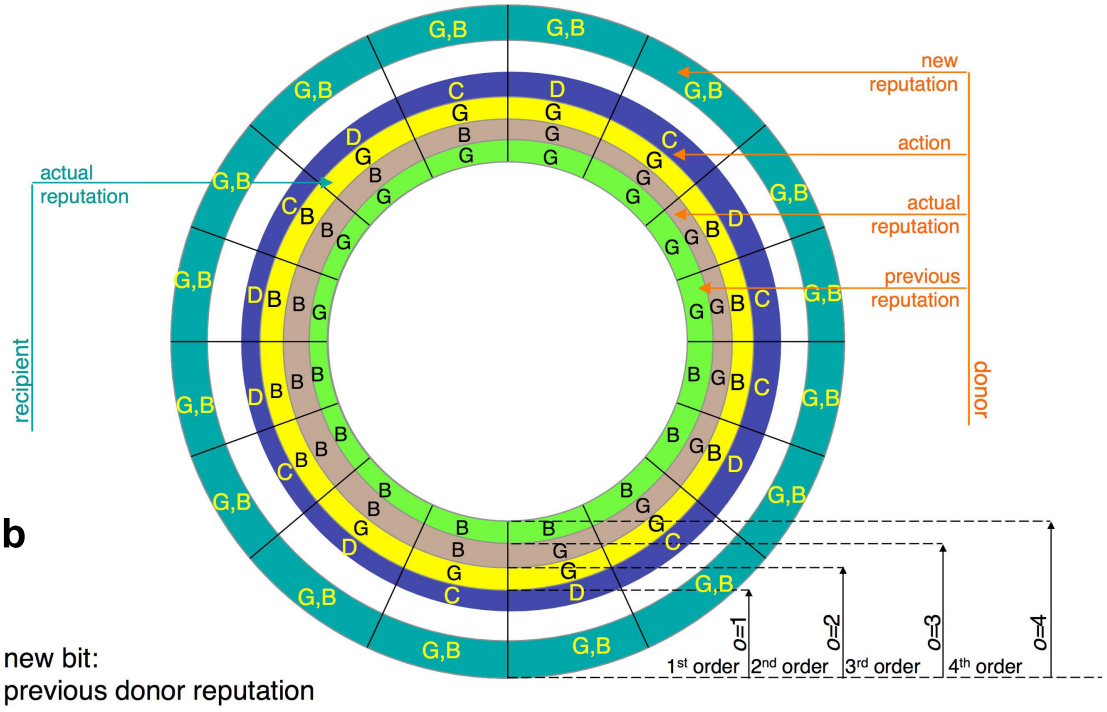


Extended Data Figure 1 | Cooperation index of third- and fourth-order norms. In the space of fourth-order norms (red bars), only a small fraction of norms (about 0.2% of 2^{16}) foster high levels of cooperation ($\eta > 0.9$), as conveyed by the complementary cumulative distribution function (CCDF; see inset for a close-up of the tail). In the space of third-order norms (blue bars), about 2% of norms (of a total of 2^8) promote high levels of cooperation ($\eta > 0.9$). Other parameters: $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$, $c = 1$, $\gamma = 0$.



Extended Data Figure 2 | The most cooperative norms. **a**, **c**, **e**, **g**, Data from simulations in which individuals pay a complexity cost c_c proportional to the complexity κ_s of the strategy that they employ ($c_c = c\kappa_s/10 = \gamma\kappa_s$). **b**, **d**, **f**, **h**, Data when no complexity cost is involved. Irrespective of whether the previous reputation of the recipient (**a**–**d**) or the donor (**e**–**h**) is used as the fourth consideration (as the fourth-order bit; see Extended Data Fig. 3), or whether there is a complexity cost involved, the highest levels of cooperation are already achieved for $\kappa = 4$. Moreover, when we plot norm performance (in terms of the cooperation index), separating norms according to their complexity κ (for $\kappa \leq 5$; **c**, **d**, **g** and **h**) it becomes apparent that fourth-order norms are generally outperformed by lower order norms. Furthermore, paying a complexity cost is most detrimental to the more sophisticated fourth-order norms, which no longer promote cooperation under indirect reciprocity. Other parameters: $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$, $c = 1$.

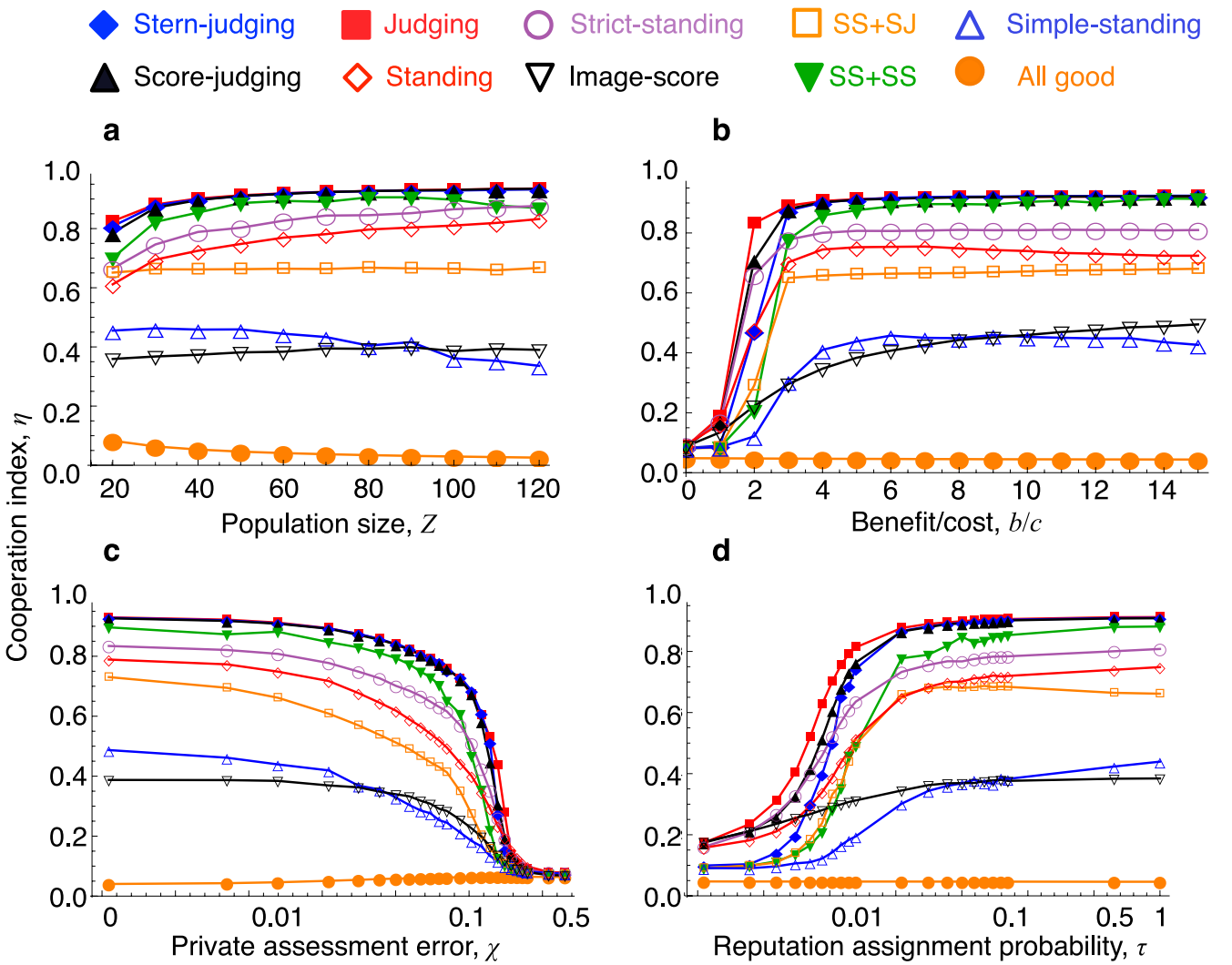




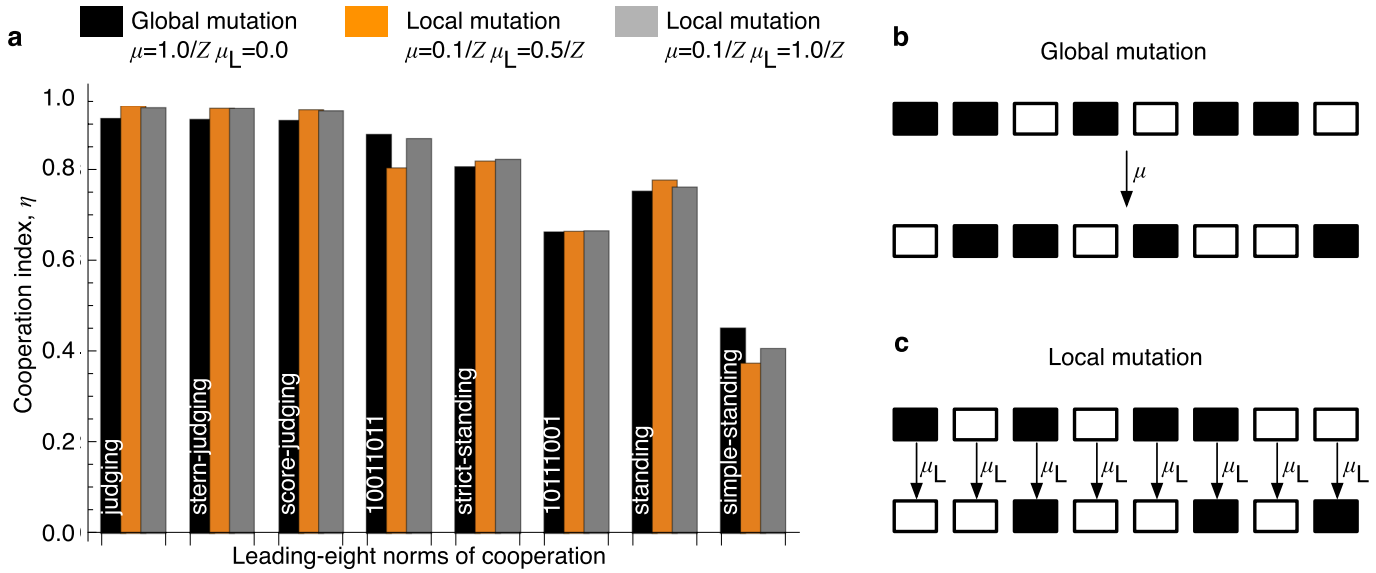
c

order (o)	1	2	3	4	…	o
number of reputation layers included (l)	0	1	2	3	…	o-1
number of strategies (s)	2^1	2^2	2^4	2^8	…	$2^{2^{o-1}}$
number of norms	2^2	2^4	2^8	2^{16}	…	2^{2^o}

Extended Data Figure 3 | Alternative ways of defining a social norm. **a**, **b**, We consider norms that attribute a new reputation (outer ring) on the basis of (i) the action of the donor (first-order bit; blue ring); (ii) the current (actual) reputation of the receiver (second-order bit; yellow ring); (iii) the current (actual) reputation of the donor (third-order bit; pink ring); and (iv) the previous reputation of either the recipient (**a**) or the donor (**b**) (fourth-order bit; green ring). In **a** and **b**, there are 2^{16} norms in total. **c**, Number of bits (layers, l) used for each norm order, and the corresponding number of possible strategies (s) and norms. Because actions are taken using the same information used by a norm to attribute a new reputation, we consider 2^8 different strategies for norms up to fourth-order.



Extended Data Figure 4 | Robustness of results to parameter variations. A full list and detailed description of all model parameters is provided in Supplementary Information. **a**–**d**, Norm performance (in terms of the complexity index) as a function of population size Z (**a**), the benefit-to-cost ratio b/c in the donation game in which individuals interact (**b**), the private assessment error probability χ (**c**) and the reputation assignment probability τ (**d**). Here we use the previous reputation of the recipient as the fourth-order bit (as in the main text) and investigate, within the space of fourth-order norms, the performance of the (second- and third-order) leading eight norms together with the (first-order) image score and (zeroth-order) all good norms. The norms 'ss' and 'sj' denote simple standing and stern judging; 'ss + sj' has the first 8 bits equal to the third-order representation of simple standing and the last 8 equal to the third-order representation of stern judging; and 'sj + ss' is defined similarly; see Supplementary Table 4 for details of these norms. Other parameters: $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $\mu = 1/Z$, $b = 5$, $c = 1$, $\gamma = 0$.



Extended Data Figure 5 | Global versus local mutation schemes. **b**, We consider a mutation scheme in which a new strategy is adopted with probability μ (drawn from a UPD) when a mutation occurs[39–42]. **c**, Alternatively, we consider a local mutation (in each strategy), whereby with probability μ_L (drawn from a UPD) one bit changes. **a**, For the leading eight norms[8], we find that the same conclusions are attained regardless of the mutation scheme considered. Other parameters: $Z = 50$, $\varepsilon = \alpha = \chi = 0.01$, $b = 5$, $c = 1$, $\gamma = 0$.

Runs: number of runs;
Gens: number of generations;
Z population size;
P: vector of all individual strategies;
P_k: strategy of individual k;
R: vector of all individual public reputations;
R^k: public reputation of individual k;
R_P^k: previous public reputation of individual k;
$\mathcal{U}\{a,b\}$: uniform distribution over integers between a and b;
$Rand \sim \mathcal{U}(0,1)$: random value sampled following the standard uniform distribution;
F_a : fitness of individual a;
γ: behavioural complexity cost;
$\kappa(P_k)$: boolean complexity of strategy P_k;
$d(A, R_A, R_D, R_P)$: new reputation given social norm d, action A, recipient actual reputation R_A, donor actual reputation R_D and previous reputation R_P;
$\Pi(x,y)$: payoff to individual x after an interaction with y where both x and y may donate following their strategies (P_x and P_y). A potential update of reputations (R^x and R^y) occurs with probability τ. This step takes into account the execution, assignment and private assessment errors.

Cooperate $\equiv 1$; **Defect** $\equiv 0$;

Good $\equiv 1$; **Bad** $\equiv 0$;

for $r \leftarrow 1$ **to** $Runs$ **do**
  for $k \leftarrow 1$ **to** Z **do**
    $P_k \leftarrow X \sim \mathcal{U}\{0, 255\}$
    $R_P^k \leftarrow X \sim \mathcal{U}\{0, 1\}$
    $R^k \leftarrow X \sim \mathcal{U}\{0, 1\}$
  end
  for $t \leftarrow 1$ **to** $Gens$ **do**
    $a \leftarrow X \sim \mathcal{U}\{1, Z\}$
    if $Rand < \mu$ **then** $P_a \leftarrow X \sim \mathcal{U}\{0, 255\}$;
    else
    $b \leftarrow X \sim \mathcal{U}\{1, Z\}$, $b \neq a$;
    $F_a \leftarrow 0$;
    $F_b \leftarrow 0$;
    for $i \leftarrow 1$ **to** $2Z$ **do**
      $c \leftarrow X \sim \mathcal{U}\{1, Z\}$, $c \neq a$;
      $F_a \leftarrow F_a + \Pi(a, c) - \kappa(P_a)\gamma$;
      /* update R_P^a, R_P^c, R^a, R^c */
      $c \leftarrow X \sim \mathcal{U}\{1, Z\}$, $c \neq b$;
      $F_b \leftarrow F_b + \Pi(b, c) - \kappa(P_b)\gamma$;
      /* update R_P^b, R_P^c, R^b, R^c */
    end
    $F_a \leftarrow \frac{F_a}{2Z}$
    $F_b \leftarrow \frac{F_b}{2Z}$
    if $Rand < (1 + e^{F_a - F_b})^{-1}$ **then** $P_a \leftarrow P_b$;
    if $t > 0.2Gens$ **then**
     /* keep track of the average number of cooperations */;
    ;
  end
end

Extended Data Figure 6 | Pseudo code for the Monte Carlo simulations used to calculate the cooperation index under each norm. Given the large number of norms considered, we used Runs = 15 and Gens = 1.5×10^4 in Figs 2–4 and Extended Data Figs 1 and 2, and Runs = 50 and Gens = 10^5 in Extended Data Figs 4 and 5.

doi:10.1038/nature25763

1. Supplementary Discussion

1.1. Alternative ways of defining 4th order norms

As already stated in the main text, norms of 4th order incorporate, on top of all information contained in norms of 3rd order, either information on the previous reputation of the recipient or the previous reputation of the donor. The differences are detailed in Extended Data Figure 3, where in panel **a** we show how an additional layer of information associated with the previous reputation of the recipient is organized, while in panel **b** we show the layout associated with encoding the previous reputation of the donor.

These two possibilities entail the same amount of information processing to the observer assigning a reputation to the donor. However, they imply different amounts of information management by the donor, as knowledge of the previous reputation of the recipient by the donor may be harder to retain – and more prone to be affected by errors – than knowledge of her/his own previous reputation.

In this work, social norms and strategies are represented as bit strings (see Methods). Thus, while acquiring a similar form, the two definitions of social norm differ in the meaning associated with the position of the bit representing past reputation information. Despite these differences, the two formulations of 4th order social norms lead, overall, to results that are qualitatively similar. In the following, and using as a reference the discussion carried out in the main text in connection with formulation **a** in Extended Data Figure 3, we summarize the main differences found regarding formulation **b**.

In Extended Data Figure 2 we compare directly the results of formulations **a** (top 4 panels) and **b** (bottom 4 panels) of Extended Data Figure 3 using the format adopted in Figure 3 for panels **a**, **b**, **e** and **f** (for convenience, panel **a** of Extended Data Figure 2 reproduces the results already contained in Figure 3).

Comparison of panels **a** and **e** shows that, similar to formulation **a**, the highest values of cooperation are attained for $\kappa \geq 4$ in formulation **b**. Panels **b** and **f**, in turn, allow the comparison of the results obtained in both formulations whenever individuals incur a complexity cost c_c by employing a strategy of complexity κ_s, with $c_c=\gamma\kappa_s$. In both formulations, adding a complexity cost hampers cooperation whenever populations operate under norms of high complexity κ. These results, in turn, strongly suggest that norms with high κ require, in general, strategies

with sizeable complexity to achieve the highest values of η. Interestingly, one also observes that in formulation **b** the cooperation levels decrease for high values of κ, even in the absence of any behavioral complexity cost (γ=0).

Panels **c**, **d**, **g** and **h** in Extended Data Figure 2 provide an alternative view of norm performance (for both formulations and in the presence and absence of a complexity cost): We plot the distribution of cooperation levels of social norms with a given complexity κ (for all norms with κ<6), as a function of η. The results clearly highlight the large number of 4th order social norms that are outperformed by lower order social norms in all cases.

1.2. Robustness of cooperation under well-known norms

In Extended Data Figure 4 we test the robustness of our results with respect to changes of different model parameters: population size, benefit/cost ratio, private assessment error and reputation assignment probability. Most of the results we discussed were computed for populations of size Z=50 which, as Extended Data Figure 4a shows, reflect the trend observed for most of the size interval spanned (from 20 to 120), with the exception of the norms simple-standing and image-score, whose η-values reverse order for $Z \geq 90$. Notwithstanding, the overall impact on the cooperation levels is small. In particular, the most cooperative, low-κ social norms (stern-judging, judging and score-judging) maintain high levels of cooperation for all population sizes. Similar conclusions are obtained if one considers different *b/c* ratios, as shown in Extended Data Figure 4b. We further study the robustness of cooperation under leading norms to different private assessment errors (χ, Extended Data Figure 4c) and reputation assignment probability (τ, Extended Data Figure 4d). The results are qualitatively similar as long as $\chi<0.1$ and $\tau>0.01$. This is particularly impressive given that **IR** may strongly depend on how faithful dissemination of information is. This point has been explicitly simulated in Ref. [1] by studying information diffusion in a graph.

1.3 Numerical analysis of norms bits

In Table 1 we provide numerical data that summarizes the most common bits that occur in the social norms that promote the highest levels of cooperation, and which provide evidence for the pattern (discussed in the main text) identified in those norms that successfully promote cooperation.

Supplementary Table 1 | Common bits in the most cooperative norms

bits	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R_P	R_P	R_P	R_P	R_P	R_P	R_P	R_P	R_P	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$
R_D	R_D	R_D	R_D	R_D	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	R_D	R_D	R_D	R_D	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$
R_A	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$
A	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$
η>0.91 13 norms	**1.00**	**0.00**	0.38	0.62	**1.00**	**0.00**	0.15	0.31	0.62	0.38	**0.00**	**1.00**	0.31	0.31	0.31	0.08
η>0.9 66 norms	**1.00**	**0.00**	0.49	0.54	**1.00**	**0.00**	0.43	0.34	0.55	0.51	**0.00**	**1.00**	0.38	0.31	0.42	0.25
η>0.85 359 norms	**1.00**	**0.00**	0.45	0.79	0.99	0.17	0.45	0.41	0.48	0.78	0.12	0.94	0.45	0.41	0.47	0.35
η>0.8 1413 norms	**1.00**	**0.00**	0.37	0.86	0.71	0.68	0.54	0.49	0.37	0.86	0.39	0.74	0.53	0.49	0.50	0.43
η>0.5 6602 norms	**1.00**	0.01	0.46	0.68	0.56	0.58	0.52	0.52	0.47	0.69	0.46	0.70	0.52	0.52	0.56	0.50

The first row of the table enumerates the different bits that form one social norm; The next 4 rows provide information of the combination of bits that define each norm. A new reputation of a donor depends on the present reputation of the donor (R_D) and recipient (R_A), together with the past reputation of the recipient (R_P) and the action by the donor (A). The following rows contain numerical values representing the fraction of norms – among those satisfying a given threshold η specified on the left column – that have value G=1 in each bit position. Other parameters: Z=50, ε=α=χ=0.01, μ=1/Z, β=1, b=5, c=1, γ=0. Here we consider the previous reputation of the recipient. For convenience, we use R_P, R_D, and R_A both as the name of a reputation layer in a social norm (Extended Data Figure 3) and as a Boolean variable that can assume value 1 = G = R or 0 = B = $\overline{R}$. Alongside, A can assume value 1=C= A or 0=D= $\overline{A}$.

First, we note that all cooperative norms (η>0.8) agree in what concerns bits in positions 0 and 1 (respectively, columns 2 and 3 in Table 1): anyone that is Good and cooperates with a Good opponent (both in the present and past, thereby called *enduring* Good) should maintain the Good reputation; anyone that defects in this scenario should have the reputation updated to Bad.

Regarding the norms that lead to η>0.9 we find that, in addition, 4 bits are remarkably constant: In these 66 (distinct) norms, bits 4 and 11 are always 1 and bits 5 and 10 are always 0. This means that the social norms promoting more than 90% of cooperation all agree that

1) those that are Bad, and cooperate with someone who is an *enduring* Good and

2) those that are Good and defect against someone who is an enduring Bad

should have a Good reputation.

Furthermore, all of these norms attribute a Bad reputation to whoever

1) is already Bad and defected against someone who is an enduring Good

or

2) is Good and cooperated with someone who is an enduring Bad.

All together, these features lead to the following pattern:

Become G (B) if helped (refused to help) an enduring G; maintain (lose) G reputation if refused to help (helped) an enduring B.

It is worth pointing out that this is a necessary (though not sufficient) pattern to achieve cooperation levels higher than 0.9, for the particular set of parameters tested. A more comprehensive study should be carried out to unravel those patterns providing sufficient conditions guaranteeing high levels of cooperation. Moreover, here we only count the "truly" distinct norms since, through mirror symmetry (the Boolean value of Good and Bad can be swapped[2]) there are pairs of equivalent norms promoting the same levels of cooperation. To remove the noise effect introduced by those norms we only take into account the ones leading to a majority of individuals with reputation Good.

1.4. Simulation details

Several analytical and numerical methods may be employed to assess the performance of a social norm. An Evolutionary Stable Strategy (ESS) analysis[2-4], elegantly offers information about the maintenance of cooperative strategies. Additionally, evolutionary dynamics in finite population — *e.g.*, in the limit of rare mutations[5,6] — provides an overall description of the most likely configurations of the population (or the prevailing strategies), which does not necessarily correlate with ESSs. This powerful approach also provides an easy means to study the evolutionary robustness of strategies against the invasion of any other[7-10], for arbitrary intensities of selection. The limit of rare mutations, however, fails to account for possible co-existence scenarios[11], and the performance of social norms under arbitrary mutation rates[12] — although the recent development of hierarchical methods[11] does provide a possible solution to this shortcoming. Nonetheless, to have a complete assessment of the performance of each social norm, here we resorted to computer simulations. In Extended Data Figure 6 we provide the pseudo-code employed in the (standard Monte Carlo) numerical computation of the cooperation levels under each social norm. In Table 2 we provide a detailed description of the full parameter space considered:

Supplementary Table 2 | Model Parameters and parameter space analyzed

Parameter	Symbol	Range analyzed	Figure
population size	Z	{20, 30, …,120}	Extended Data Fig 4
execution error	ε	{0.01}	-
assignment error	α	{0.01}	-
private error	χ	{0, 0.001, 0,002,…,0.01,0.02, …, 0.5}	Extended Data Fig 4
global mutation	μ	{1/Z, 0.1/Z}	Extended Data Fig 5
benefit/cost (donation game)	b/c	{0, 1, 2, …, 15}	Extended Data Fig 4
behavioral complexity cost	γ	{0, 0.1}	Extended Data Fig 3
local mutation	μ	{0, 0.5/Z, 1/Z}	Extended Data Fig 5
probability reputation assignment	τ	{0, 0.001, 0,002,…,0.01,0.02, …, 1}	Extended Data Fig 4

2. Calculating social norm complexity: an explicit example

Let us summarize the procedure of calculating the Boolean complexity of a social norm by means of an example. In the following, we choose the social norm *Judging*. Calculating its Boolean complexity involves three steps:

Step 1. Translate the social norm to the corresponding DNF, converting each bit of the norm to the corresponding *minterm*:

Table 3 represents a truth table with 4 input variables. The inputs are $\mathbf{R_P}$ (previous reputation, where R_P means Good and $\overline{R_P}$ means Bad – a notation that we follow throughout this section), $\mathbf{R_D}$ (actual reputation of the Donor), $\mathbf{R_A}$ (actual reputation of the Recipient) and **A** (action of the Donor, where A means Cooperate and $\overline{A}$ means Defect). The last row of this table corresponds to a Boolean function, in this case representing the social norm *Judging*[2]. That function receives the previous inputs and produces *True* (or 1) if the next reputation of the Donor is Good, and *False* (or 0) if the next reputation is Bad. This way, we can write *Judging* as a disjunction of *minterms* (i.e., products of inputs that have value one in exactly one position of the previous table). *Judging* prescribes Good in 6 different situations, so its Boolean function will be composed by 6 *minterms*:

$R_P R_D R_A A \vee R_P R_D \overline{R_A}\,\overline{A} \vee R_P \overline{R_D} R_A A \vee \overline{R_P} R_D R_A A \vee \overline{R_P} R_D \overline{R_A}\,\overline{A} \vee \overline{R_P}\,\overline{R_D} R_A A$. We could also use the *minterm* notation: $\Sigma m(0,3,4,8,11,12)$, where the bits leading to reputation 1 (Table 3) are enumerated after Σm. In the next step, we simplify this Boolean function.

Supplementary Table 3 | Truth table of a social norm

bits	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
R_P	R_P	R_P	R_P	R_P	R_P	R_P	R_P	R_P	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$	$\overline{R_P}$
R_D	R_D	R_D	R_D	R_D	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	R_D	R_D	R_D	R_D	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$	$\overline{R_D}$
R_A	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$	R_A	R_A	$\overline{R_A}$	$\overline{R_A}$
A	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$	A	$\overline{A}$
Judging	1	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0

Here we provide the example of computing the DNF form for the social norm *Judging*, identified by the 6 Good (represented by 1) entries in the last row. We use the same format of Table 1 for the first 5 rows.

Step 2. Apply a DNF minimization algorithm (QM algorithm):

The *Quine-McCluskey* (QM) algorithm[13] constitutes a computationally friendly algorithm to minimize a Boolean function. First, this algorithm proceeds by finding the redundant literals in the different products. In the example above, we note that the products $R_P R_D R_A A$ and $R_P \overline{R_D} R_A A$ only differ in R_D and thus the terms can be combined into $R_P R_A A$ (the consensus theorem). After applying a similar procedure iteratively, one can compute the terms that can no longer be combined with other terms, which are called prime implicants (i.e., terms that are not redundant). If every *minterm* is covered by a prime implicant, the method returns the disjunction of the prime implicants as the minimized DNF, with a minimum number of terms. Additional procedures (such as the *Petrick's method*) can be used to generate a minimal DNF from the obtained prime implicants. In the example of *Judging*, QM would return the minimal DNF $R_A A \vee \overline{R_D}\,\overline{R_A}\,\overline{A}$.

Step 3. Count the number of literals:

Once a minimal DNF is obtained, we simply count the number of literals. The minimal DNF $R_A A \vee \overline{R_D}\,\overline{R_A}\,\overline{A}$ is composed by 5 literals, which translates into a Boolean complexity κ of 5.

Another simple example is simple-standing, whose minimal DNF is $A \vee \overline{R_A}$, which translates into a Boolean complexity of 2. In Table 4 we provide the *minterm* notation, minimal DNF and Boolean complexity of most of the well-known social norms found to date.

Supplementary Table 4 | The Boolean function of some of the most well-known social norms

Decimal	Name	*minterm* notation	Minimal DNF	κ	Order
0	All-Bad	$\Sigma m()$	False	0	0
65535	All-Good	$\Sigma m(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)$	True	0	0
34952	Shunning	$\Sigma m(0,4,8,12)$	$R_A A$	2	2
39064	**Judging**	$\Sigma m(0,3,4,8,11,12)$	$R_A A \vee R_D \overline{R_A}\,\overline{A}$	5	3
39321	**Stern-Judging**	$\Sigma m(0,3,4,7,8,11,12,15)$	$R_A A \vee \overline{R_A}\,\overline{A}$	4	2
39578	**Score-Judging**	$\Sigma m(0,3,4,6,8,11,12,14)$	$R_A A \vee R_D \overline{R_A}\,\overline{A} \vee \overline{R_D} A$	7	3
39835	**SJ+SS**	$\Sigma m(0,3,4,6,7,8,11,12,14,15)$	$R_A A \vee \overline{R_A}\,\overline{A} \vee \overline{R_A}\,\overline{R_D}$	6	3
43690	Image-Score	$\Sigma m(0,2,4,6,8,10,12,14)$	A	1	1
47288	**Strict-Standing**	$\Sigma m(0,2,3,4,8,10,11,12)$	$R_A A \vee \overline{R_A} R_D$	4	3
47545	**SS+SJ**	$\Sigma m(0,2,3,4,7,8,10,11,12,15)$	$R_A A \vee \overline{R_A}\,\overline{A} \vee \overline{R_A} R_D$	6	3
47802	**Standing**	$\Sigma m(0,2,3,4,6,8,10,11,12,14)$	$A \vee \overline{R_A} R_D$	3	3
48059	**Simple-Standing**	$\Sigma m(0,2,3,4,6,7,8,10,11,12,14,15)$	$A \vee \overline{R_A}$	2	2

From the Boolean representation of a social norm, we can compute the Boolean complexity (κ) of a norm as the number of literals in its minimal DNF form. Norms in **boldface** represent the leading-eight norms of cooperation identified by Ohtsuki and Iwasa[2,3]. Interestingly, we find another 3rd order leading-eight norm (a variant of Stern-Judging, named above as Score-Judging) that is able to foster high levels of cooperation, combined with a low average behavioural complexity ζ. This norm is shown in Figure 4 to perform almost as well as Stern-Judging and Judging (see black circle in the vicinity of these norms), yet exhibiting a higher Boolean complexity (κ=7).

References

1 Ohtsuki, H., Iwasa, Y. & Nowak, M. A. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. *Nature* **457**, 79-82 (2009).

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

3 Ohtsuki, H. & Iwasa, Y. The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J. Theor. Biol.* **239**, 435-444 (2006).

4 Hofbauer, J. & Sigmund, K. *Evolutionary games and population dynamics*. (Cambridge University Press, 1998).

5 Santos, F. P., Santos, F. C. & Pacheco, J. M. Social Norms of Cooperation in Small-Scale Societies. *PLoS Comput. Biol.* **12**, e1004709 (2016).

6 Fudenberg, D. & Imhof, L. Imitation Processes with Small Mutations. *J. Econ. Theory* **131**, 251-262 (2005).

7 Stewart, A. J. & Plotkin, J. B. From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proc. Natl. Acad. Sci. USA* **110**, 15348-15353 (2013).

8 Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proc. Natl. Acad. Sci. USA* **111**, 17558-17563 (2014).

9 Pinheiro, F. L., Vasconcelos, V. V., Santos, F. C. & Pacheco, J. M. Evolution of All-or-None Strategies in Repeated Public Goods Dilemmas. *PLoS Comput. Biol.* **10**, e1003945 (2014).

10 Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K. & Nowak, M. A. Memory-n strategies of direct reciprocity. *Proc. Natl. Acad. Sci. USA* **114**, 4715-4720 (2017).

11 Vasconcelos, V. V., Santos, F. P., Santos, F. C. & Pacheco, J. M. Stochastic Dynamics through Hierarchically Embedded Markov Chains. *Phys Rev Lett* **118**, 058301 (2017).

12 Santos, F. P., Pacheco, J. M. & Santos, F. C. Evolution of cooperation under indirect reciprocity and arbitrary exploration rates. *Sci. Rep.* **6** (2016).

13 McCluskey, E. J. Minimization of Boolean functions. *Bell Labs Technical Journal* **35**, 1417-1444 (1956).