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).