# Author Queries

*Journal:* Journal of the Royal Society Interface

*Manuscript:* rsif20190127

As the publishing schedule is strict, please note that this might be the only stage at which you are able to thoroughly review your paper.

Please pay special attention to author names, affiliations and contact details, and figures, tables and their captions.

The corresponding author must provide an ORCID ID if they haven't done so already. If you or your co-authors have an ORCID ID please supply this with your corrections. More information about ORCID can be found at http://orcid.org/.

No changes can be made after publication.

**SQ1** Please confirm that this paper is intended to be Open Access. The charge for Open Access should be paid before publication. If you have not yet received an email requesting payment please let us know when returning your corrections.

**SQ2** Your supplementary material will be published online alongside your article and on rs.figshare.com exactly as the file(s) are provided. Therefore, please could you either confirm that your supplementary material is correct, or – if you have any changes to make to these files – email these along with your proof corrections to the journal inbox. Your ESM files are listed here for your convenience:

evil_interface_ESM_revision3_fin.pdf

**Q1** Mandatory end section has been added to your paper; please check if that is correct.

**Q2** Please provide publisher location in reference [56].

ARTICLE IN PRESS

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

**Cite this article:** García J, Traulsen A. 2019 Evolution of coordinated punishment to enforce cooperation from an unbiased strategy space. *J. R. Soc. Interface* 20190127. http://dx.doi.org/10.1098/rsif.2019.0127

Received: 27 February 2019
Accepted: 25 June 2019

**Subject Category:**
Life Sciences–Mathematics interface

**Subject Areas:**
biomathematics, evolution, computational biology

**Keywords:**
collective action, punishment, evolution, cooperation

**Author for correspondence:**
Arne Traulsen
e-mail: traulsen@evolbio.mpg.de

Electronic supplementary material is available online at rs.figshare.com.

THE ROYAL SOCIETY PUBLISHING

# Evolution of coordinated punishment to enforce cooperation from an unbiased strategy space

Julián García[1] and Arne Traulsen[2]

[1]Faculty of Information Technology, Monash University, Melbourne, Australia
[2]Department of Evolutionary Theory, Max Planck Institute for Evolutionary Biology, 24306 Plön, Germany

AT, 0000-0002-0669-5267

The emergence and maintenance of punishment to protect the commons remains an open puzzle in social and biological sciences. Even in societies where pro-social punishing is common, some individuals seek to cheat the system if they see a chance to do so—and public goods are often maintained in spite of cheaters who do not contribute. We present a model accounting for all possible strategies in a public goods game with punishment. While most models of punishment restrict the set of possible behaviours, excluding seemingly paradoxical anti-social strategies from the start, we show that these strategies can play an important role in explaining large-scale cooperation as observed in human societies. We find that coordinated punishment can emerge from individual interactions, but the stability of the associated institutions is limited due to anti-social and opportunistic behaviour. In particular, coordinated anti-social punishment can undermine cooperation if individuals cannot condition their behaviour on the existence of institutions that punish. Only when we allow for observability and conditional behaviours, anti-social strategies do no longer threat cooperation. This is due to a stable coexistence of a minority supporting pro-social institutions and those who only cooperate if such institutions are in place. This minority of supporters is enough to guarantee substantial cooperation under a wide range of conditions. Our findings resonate with the empirical observation that public goods are resilient to opportunistic cheaters in large groups of unrelated individuals. They also highlight the importance of letting evolution, and not modellers, decide which strategies matter.

## 1. Introduction

Most modern societies have put in place institutions that support and promote collective action. Understanding the origin of these institutions is an important challenge across biological and social sciences [1]. The outstanding capacity of humans to engage in large-scale cooperation often relies on these institutionalized enforcement mechanisms [2]. Centralized institutions for cooperation also have experimental and empirical support [3,4], but explaining how these institutions arise from individual incentives is an open problem [5]. Here, we propose that these institutions play a role in enabling cooperation, not only by implementing punishment against free-riders [6–8], but also by means of their visibility, which enables agents to condition their actions on whether these institutions are present or not.

Punishment provides a possible solution to the problem of collective action [9–14]. The vast majority of theoretical and experimental work focuses on pro-social peer punishment, exerted by peers and individually directed towards those that do not cooperate [15–20]. Anti-social peer punishment is instead directed towards those that do contribute to the public good [21]. Experiments and models have shown that anti-social punishment can diminish the effectiveness of punishment in promoting cooperation [22–25]. In many instances,

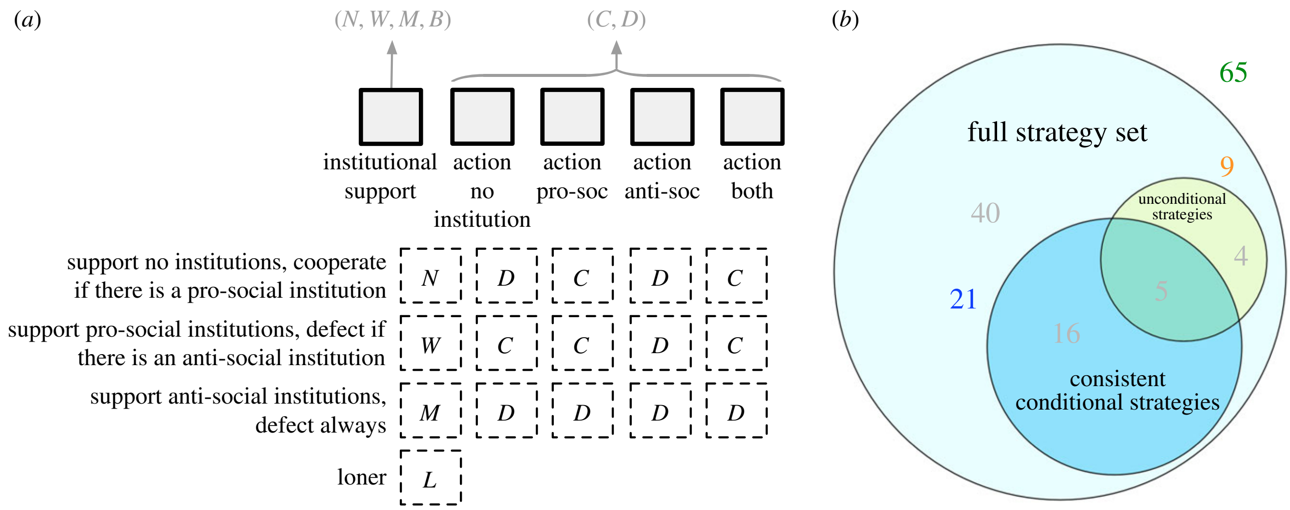


**Figure 1.** Description of the strategy sets. (*a*) The 64 strategies taking part in the game are characterized by their contribution to the punishment pool (none (*N*), pro-social (*W*), anti-social (*M*) or both (*B*)) and their action for each institutional set-up. In addition, in an optional game, we have the loner strategy which neither punishes nor gets punished. (*b*) The 65 strategies can be reduced to 21 by assuming consistent actions—players do not support pools that punish them. Alternatively, one can focus on nine unconditional strategies, where actions are not affected by the punishment set-up. Iterated removal of dominated strategies results in five strategies that are also consistent. (Online version in colour.)

however, punishment is not individual, but a coordinated action of many individuals [26]. An extreme form of such coordination is the kind of pool punishment that emerged as the typical way of punishment in modern human societies: individuals commit an investment into a pool to pay for the punishment of those that do not comply with a social norm [6,27–29]. Here, we show that prosocial punishment can withstand the presence of cheaters and anti-social behaviour, but this outcome only emerges when considering *all* possible strategies in a public goods game. This result highlights the importance of avoiding artificial restrictions in the strategy set of evolutionary models.

We study the evolutionary dynamics of coordinated anti-social punishment and ask whether the associated coordinated punishment can emerge, potentially undermining cooperation. Our model allows for the evolutionary competition between anti-social and pro-social punishment. We consider two main scenarios: (i) Observable institutions allow individuals to condition their actions on the existence of punishment institutions. (ii) If institutions cannot be observed, individuals are unable to condition their behaviour. When punishment is not observable, anti-social punishment triggers a collapse of the public good. If institutions are observable, cooperation can be established and stabilized by pro-social punishment, even in the face of anti-social behaviour.

## 2. The model

### 2.1. A game of cooperation with institutional punishment

Our model follows Sigmund *et al.* [6] in the basic set-up of an optional public goods game between $n$ players with three stages. (i) The first stage is institutional commitment, in which players may commit funds to an institution that will later punish free-riders or cooperators. (ii) The second stage is the public goods game, in which individuals may decide contribute or not to a public good. (iii) The third stage is punishment, in which players are fined in accordance to the institutions in place.

In the institutional commitment stage (i), participants choose what kind of institution they want to support. They can support pro-social punishment directed to defectors, or anti-social punishment of cooperators, both, or none. Funding an institution costs a fixed amount $\gamma$ if the punishment takes place. An institution is established—and therefore costly—only if there are at least $k$ players contributing to it.

(ii) During the public goods stage, players use the information about the institutions in place, choosing whether they contribute an amount $c > 0$. Contributions are multiplied by a factor $r > 1$ and distributed among the $n - 1$ other players [30].

(iii) During the punishment stage, agents are fined according to the institutions in place, and the amount of players supporting the corresponding institutions. Non-contributors are punished by an amount $\beta$ times the number of supporters of the pro-social institution. Contributors are punished by an amount $\beta$ times the number of supporters of the anti-social pool.

Since the game is optional, we also let agents opt out of the game altogether. Those that do not take part in the game obtain a loner pay-off $\sigma > 0$, regardless of the decisions of others [31]. An optional game, therefore, includes the 'Loner' strategy, whereas a non-optional game precludes it.

When agents can decide what to do depending on the existence of punishment institutions, a large strategy set emerges, as follows. In the first stage of the game, an individual decides their institutional support; with options for supporting no institution (*N*), only a prosocial institution (*W*), only an antisocial institution (*M*) or both institutions (*B*)—thus institutional support entails four possibilities. In addition, agents also decide whether to contribute or not to the public good, contingent on the institutional arrangement in place; i.e. cooperate or defect given there is no institution, cooperate or defect when there is a prosocial institution only, what to do when there is only an antisocial institution, and what to do when both institutions are in place—thus, we have $2^4$ possibilities. This yields in total $4 \times 2^4 = 64$ strategies. If we further make the game optional and include the loner option, we obtain a total of 65 strategies (figure 1*a*).

In the analysis that follows we will focus on different subsets of this large strategy set. These subsets will also imply different assumptions in the game. In particular, we study a set of nine non-conditional strategies, in which cooperation or defection does not depend on institutional arrangements—this is equivalent to institutions that cannot be observed. We also study the set of all 65 strategies, whose analysis will be shown equivalent to that arising from the set of 21 consistent strategies. A consistent strategy is such that a player will not contribute to an institution that would punish her actions (figure 1*b*).

## 2.2. Evolutionary dynamics

The game described above determines the pay-off of each player in the population. We calculate the average pay-off across all possible configurations of groups given the current numbers of each type in the population, such that all players using the same strategy have the same pay-off $\pi$. This pay-off determines how many player will adopt the corresponding strategy, as successful strategies spread in a finite population in proportion to their relative fitness. We consider a Moran process, where a single individual chooses a new strategy in each time step with probability proportional to fitness $f$. We assume an exponential pay-off to fitness mapping [32,33], such that fitness is given by $f = \exp[+\omega\pi]$, where $\omega$ is the intensity of selection, see electronic supplementary material for details. In addition, there is a small probability $\mu$ that an individual switches to a new random type. In our simulations, we focus on the case of population size $N = 50$, mutation rate $\mu = 0.001$, and intensity of selection $\omega = 10$.

# 3. Results

We study which strategies are favoured by an evolutionary process. Under small mutation rates, the dynamics of the evolutionary process is confined to edges between two strategies [34,35]. Therefore, it is instructive to first compare pay-offs between any two strategies. A full overview of the $(1/2) \times 65 \times 64 = 2080$ strategy pairs is possible, but hard to grasp. We thus start by reducing the size of this large strategy set, making further assumptions on the nature of possible strategies.

## 3.1. Non-observable institutions

In the simplest case, individuals cannot condition their actions on the existence of a punishment institution [6]. This is equivalent to punishment institutions that cannot be observed. This case leads to nine strategies, as follows. Individuals have four options to support institutions times two possible actions in the public goods game. In addition, individuals can choose to abstain from the game. Out of these nine strategies, four are dominated by others. However, instead of neglecting these strategies from the beginning, we include them in our computational model and let evolution decide whether they play any role. Pro-social institutions can promote temporary cooperation, even when fines are exclusively directed towards defectors and not used to stabilize punishment [6]. But this kind of model assumes that anti-social institutions are excluded. When allowing for anti-social institutions, cooperation can not only be undermined by defectors not supporting any institution but also by defectors that in addition set up an anti-social punishment institution and stretch their relative advantage (figure 2). As a consequence, more players tend to abstain from the public goods game and—more importantly—fewer players cooperate. Thus, anti-social institution supporters temporarily invade. This dynamics triggers a sizeable reduction of cooperation as shown in figure 2.

Figure 2 also shows two unstable fixed points; between the prosocial and the antisocial institution (*WCCCC* and *MDDDD*), and between the prosocial institution and defectors (*WCCCC* and *NDDDD*).

The existence of these unstable fixed points can be illustrated from the competition between the two associated strategies.

In the case of *WCCCC* and *NDDDD*, the associated pay-offs in a population with $j$ cooperating players are

$$\begin{aligned}\pi_{WCCCC} &= -\gamma - c + \sum_{i=0}^{n-1} \frac{\binom{j-1}{i}\binom{N-j}{n-i-1}}{\binom{N-1}{n-1}} \frac{cri}{n-1} \\ &= -\gamma - c + cr\frac{j-1}{N-1}\end{aligned} \tag{3.1}$$

and

$$\begin{aligned}\pi_{NDDDD} &= +\sum_{i=0}^{n-1} \frac{\binom{j}{i}\binom{N-j-1}{n-i-1}}{\binom{N-1}{n-1}} \left(\frac{cri}{n-1} - \beta i\right) \\ &= (cr - \beta(n-1))\frac{j}{N-1}\end{aligned} \tag{3.2}$$

For $j = 1$, we have $\pi_{WCCCC} < \pi_{NDDDD}$ when the costs of cooperation and supporting the institution outweigh the fine imposed on the defectors and the additional benefit they get from the public goods (due to the setting where a cooperating player does not benefit from her own contribution), $-\gamma - c < (cr - \beta(n-1)/N-1)$. This condition will always be fulfilled for large $N$. In this case, defectors cannot be invaded by cooperating supporters of a prosocial institution. For $j = n-1$, the condition for $\pi_{WCCCC} > \pi_{NDDDD}$ reduces to $+\gamma + c < \beta(n-1)$, i.e. the costs of cooperation and supporting the institution must be smaller than the fine imposed on the defectors. Thus, cooperating supporters of a prosocial institution cannot be invaded by defectors. Since neither of the two strategies can invade the others (and since the pay-offs are linear in $j$), this results in a bi-stability.

A similar arguments holds for the pair, *MDDDD* and *WCCCC*. A more comprehensive (numerical) analysis that includes all pairs of these nine strategies is presented in the electronic supplementary material.

The paths via such bi-stabilities are not prevalent in the computer simulations that will typically follow the paths highlighted in the figure. Four different cycles are prevalent: the first one, $L \rightarrow NCCCC \rightarrow NDDDD \rightarrow L$, has already been described in Hauert *et al.* [31], the second one, $L \rightarrow WCCCC \rightarrow NCCCC \rightarrow NDDDD \rightarrow L$ additionally emerges in the institutional punishment model of Sigmund *et al.* [6]. while the remaining two, $L \rightarrow NCCCC \rightarrow MDDDD \rightarrow L$, and $L \rightarrow NCCCC \rightarrow MDDDD \rightarrow NDDDD \rightarrow L$, emerge only in the presence of anti-social punishment institutions.

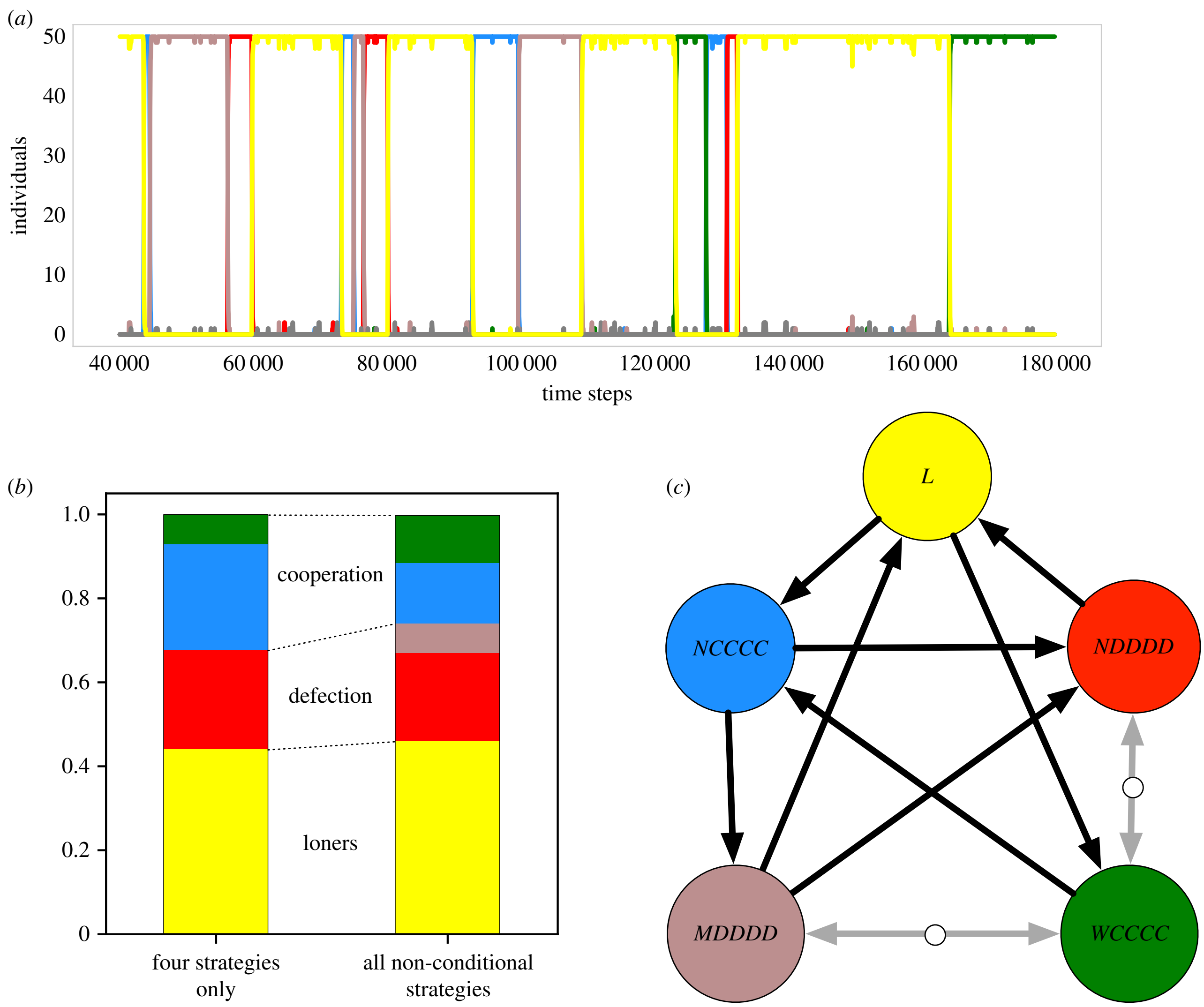


**Figure 2.** Evolutionary dynamics for the nine unconditional strategies. (*a*) Typical simulation run over 140 000 time steps shows different cycles in which strategies replace each other, the dominated strategies that punish themselves are only present in low abundance (colour code for the strategies as in the other panels). Due to low mutation rate, the abundances are typically close to 0 or 100%. (*b*) The stationary distribution obtained from computer simulations—four strategies from [6] only, without antisocial punishment, versus all unconditional strategies. Averages are taken over 300 independent repetitions, each running for $5 \times 10^6$ generations, averaging the second half of each replicate. The inclusion of the anti-social institution (*MDDDD*), which from the outside is paradoxical and should play no role in evolution, reduces the level of cooperation. (*c*) Pairwise invasion diagram for the five strategies that are not dominated. Circles represent the strategies, arrows indicate the direction of selection. Bold arrows represent the paths that are prevalent in computer simulations. The dynamics can follow several intertwined cycles, e.g. $L \rightarrow NCCCC \rightarrow MDDDD \rightarrow NDDDD \rightarrow L$, see main text (we use our default set of game parameters $n = 5$, $\sigma = 1$, $c = 1$, $r = 3$, $\gamma = 0.7$, $\beta = 1.5$, population size $N = 50$, mutation rate $\mu = 0.001$, intensity of selection $\omega = 10$). (Online version in colour.)

One may argue that an anti-social institution should never arise because cooperators can be invaded by defectors through an easier path without any anti-social punishment. However, this argument also applies to the transition from loners to cooperators, which could occur with a pro-social pool, but also without any punishment. The crucial difference is that while the pro-social pools can rise to high abundance [6], the anti-social behaviour only plays an important role in facilitating the emergence of other strategies without becoming prevalent itself. The complexity of implementing an anti-social and a pro-social institution per se is the same. Their asymmetry arises only from evolutionary competition.

Notably, figure 2*b* shows that artificially taking out seemingly unimportant strategies has an effect in the predicted level of cooperation. For our default set of parameters, we observe a slight increase in Loners, from 44% to 46%. Likewise, defection increases by about 4% and the overall level of cooperation decreases by 6% through the introduction of the antisocial punishment institution.

## 3.2. Observable institutions

In many cases, information on punishment pools may be available before players need to make a decision on their contribution, such that players can condition their actions on the existence of institutions [26,28,29]. For example, criminals are arguably less likely to offend if they know an institution is in place to punish them [36]—although see also [37]. First, we focus on consistent conditional strategies (figure 1) which do not punish themselves: If individuals support a pro-social institution, they cooperate if that institution is in place. If individuals support an anti-social institution, they defect if that institution is in place. Thus, they do not support both institutions at the same time. Moreover, if they cooperate (defect) under a single institution, they also cooperate (defect) when both institutions exist. This set contains four strategies that support the pro-social institution, $W \circ C \circ C$ (where the entries $\circ$ are either $C$ or $D$) and four strategies that support the anti-social institution, $M \circ \circ DD$. In addition, we have 12

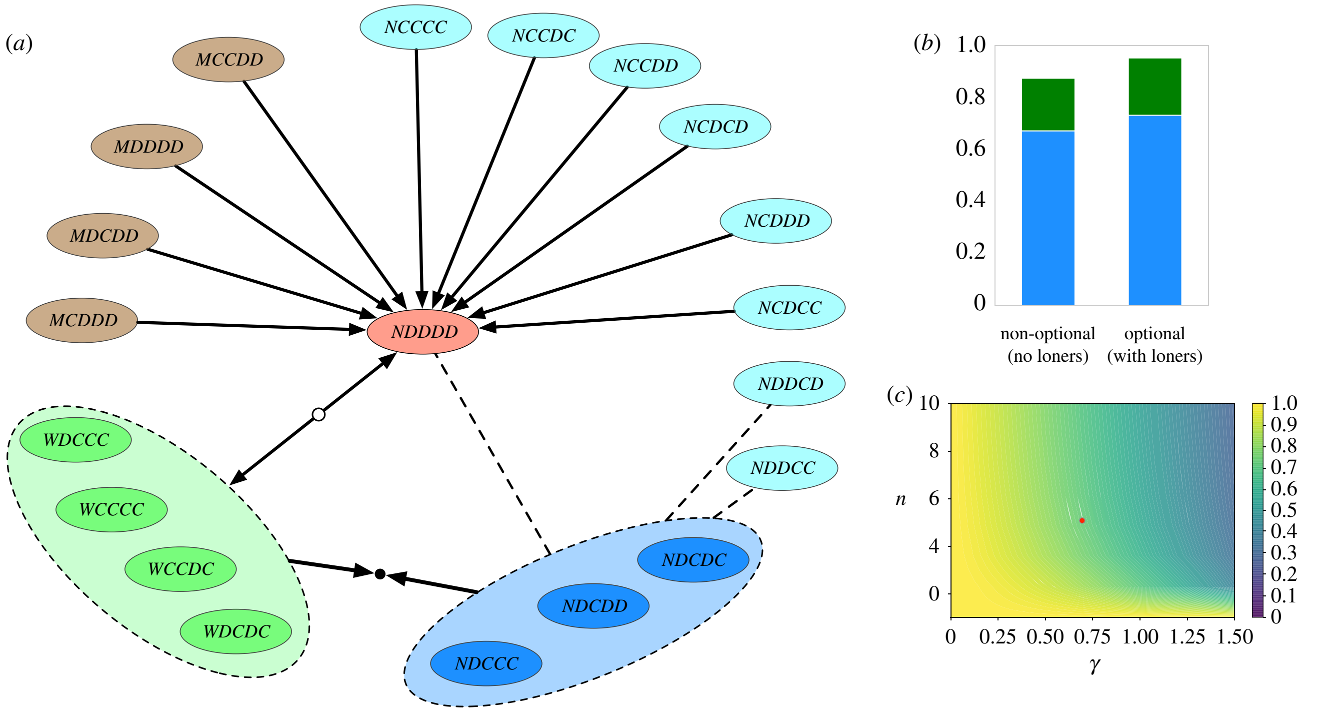


**Figure 3.** Evolutionary dynamics for the 20 consistent conditional strategies. (*a*) Evolutionary dynamics between pairs of strategies. We only depict invasions that are prevalent in the dynamics under strong selection. Given sufficient time, stable coexistences between two kinds of players emerge: those that support pro-social punishment and those not supporting any institution, but cooperating in the presence of pro-social punishment. For simplicity, the diagram focuses on a non-optional game, adding the Loner strategy creates a fast path from defection, via Loners into either side of the stable mixture. (*b*) There are 12 different stable coexistences with similar abundance based on the four strategies supporting pro-social punishment and the three opportunistic strategies that cooperate in the presence, but defect in the absence of coordinated pro-social punishment. We group these strategies in the cases of non-optional and optional public goods game. Averages taken as described in figure 2. (*c*) The probability that a pro-social pool is implemented in the coexistence decreases with increasing costs $\gamma$ and increasing group size $n$. The dot indicates our default set of parameters (see main text or figure 2). (Online version in colour.)

strategies that do not support any institution, $N\circ\circ\circ\circ$, where the four strategies $N\circ CCD$ and $N\circ DDC$ are excluded. Finally, we have the option to abstain from the public goods game, $L$.

Electronic supplementary material, figure S1 summarizes the dynamics between the associated 210 $(=(1/2)\times 21\times 20)$ pairs of consistent conditional strategies. In this set of strategies, no strategy is strictly dominated. The evolutionary dynamics is governed by stable coexistences between a minority of players that support the pro-social institution, $W\circ C\circ C$ (called $I$ below), and opportunists that cooperate only when the pro-social institution is in place [28], $NDC\circ\circ$ (called $O$ below). As long as a single supporter of the institution can induce its existence, $k=1$, their pay-off is $\pi_I = r\,c - c - \gamma$. The probability that a focal opportunist is in a group that contains at least one supporter of the pro-social institution is $1-x^{n-1}$, where $x$ is the fraction of opportunists, who obtain a pay-off $rc-c$ in that case. If opportunists are alone, no one cooperates and their pay-off is zero, such that their average pay-off becomes $\pi_O = (rc-c)(1-x^{n-1})$. The condition $\pi_I=\pi_O$ leads to a unique stable equilibrium,

$$x^* = \left(\frac{\gamma}{rc-c}\right)^{1/(n-1)}. \tag{3.3}$$

The probability that an institution is implemented in a group is then $1-(x^*)^n$ – a small fraction of supporters of the institution can induce high levels of cooperation (figure 3).

The resulting stable coexistences are remarkably resilient to evolutionary invasions. We can see this by doing a pairwise analysis, and considering strategies $W\circ C\circ C$ and $NDC\circ\circ$, where the entry in $\circ$ is irrelevant (and the last entry in $W\circ C\circ C$ follows from our restriction to consistent strategies). This coexistence cannot be invaded by any other single mutant:

— Players who cooperate in the absence of an institution, $NCC\circ\circ$, would be exploited by the $NDC\circ\circ$ resident.
— Those that defect in the presence of an institution, $NDD\circ\circ$, would suffer from punishment.
— Any player supporting the anti-social institution would obtain a lower payoff then the players in the stable coexistence: There, both players obtain $-\gamma + rc - c$. A single supporter of the anti-social institution would at most get $-\gamma + r\,c - \beta$. This assumes that in the presence of both institutions, all the other players cooperate and that at least one individual contributed to a pro-social institution. As $\beta > c$, this pay-off is always smaller than the pay-off of the two resident types. Therefore, supporters of an anti-social institution cannot invade.

Thus, the coexistence between $W\circ C\circ C$ and $NDC\circ\circ$ is stable against single mutants. As we have made no assumption on $\circ$, this holds for all such strategies. There is no other pairwise stable coexistence in the system.

Although this theoretical analysis assumes that populations are very large, our simulation results for $N=50$ perfectly match the prediction figure 3*b*. This is due to the fact that selection is strong in the simulations. As a general rule of thumb, we expect this prediction to hold whenever the product of intensity of selection and population size is large [38]. This relationship between infinite and finite populations has been studied in detail elsewhere, see [39].

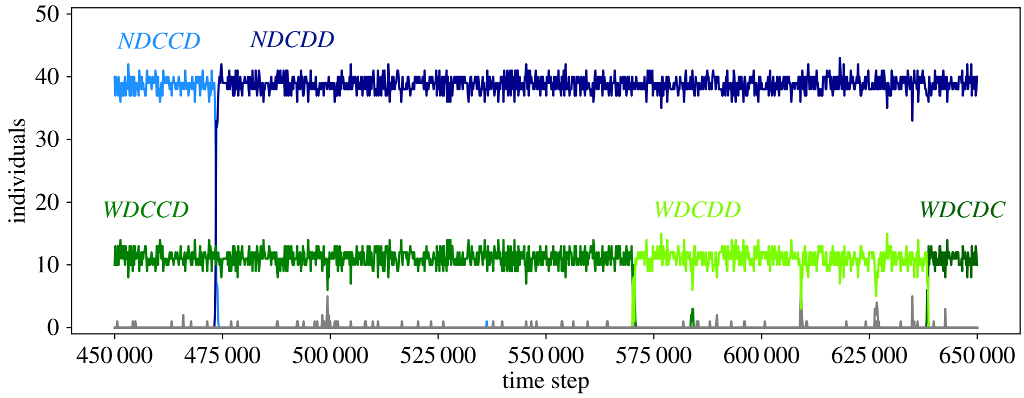


**Figure 4.** Time-series showing a snapshot of the evolution in the complete strategy space of 65 strategies. Different instances of stable coexistences in the form *WDC* ∘ *C* and *NDC* ∘∘ are following each other, where each strategy can be replaced independently of the other. For instance, here *NDCCD* is replaced by *NDCDD* before replacements within the institutional supporters take place (from *WDCCD* via *WDCDD* to *WDCDC*). For the present parameter set (see main text), 23% of institutional supporters induce cooperation in 73% of all games (window over $2 \times 10^5$ time steps, $k = 1$). (Online version in colour.)

Note that with conditional strategies, defection can be left via neutral paths towards strategies that do not support any institution and do not cooperate in the absence of institutions—but potentially in their presence. This implies that the game no longer needs to be optional for cooperation to evolve [28], a potential issue with previous models [40]. Figure 3 shows the strategies in this set of 20 strategies as well the typical evolutionary dynamics between pairs of strategies.

Our model assumes that if it is costly to make institutions visible, the cost is part of the funds paid in order to establish the institution. However, it is also possible to assume that this cost is paid by strategies using conditional information. This extension is discussed in the electronic supplementary material.

## 3.3. Including all conditional strategies

So far, we have given arguments that allow to focus on specific subsets of strategies. To verify the robustness of our findings, we also consider the full strategy set of 65 strategies and implement the same computational evolutionary model as before. Because this strategy space includes many new possibilities, a plethora of different combinations could evolve. However, the same coexistence between individuals supporting pro-social institutions and opportunists not supporting any institution is found again, with the same twelve stable coexistences as in the subset of the 20 consistent strategies considered above. Thus, the evolutionary outcome is independent of the choice of the strategy subset in our model once the key opportunistic strategies are considered. Figure 4 shows a typical time series of the dynamics with a complete strategy space. The system spends the vast majority of time in stable coexistences, where the strategies are occasionally replaced by others which display the same behaviour in this situation.

## 3.4. Higher thresholds for punishment implementation

When a certain number of supporters is required to implement an institution, all players can find themselves in groups with no punishment in place. This implies that the behaviour in such situations is under selection. Qualitatively, the results remain identical to the case where punishment can be implemented by a single supporter, $k = 1$, but the position of the fixed point and the level of cooperation can change. The structure of the game resembles a threshold public goods game [41,42] see electronic supplementary material. When at least two supporters of an institution are needed, $k = 2$, the equilibrium fraction of players supporting the pro-social set-up increases from ≈ 23% ($k = 1$) to ≈ 43% ($k = 2$). At the same time, the probability that the pool is actually implemented decreases slightly from ≈ 73% ($k = 1$) to ≈ 72%. A pairwise analysis of strategies reveals that now other coexistences are possible as well, for example between *WCCDD* and *NDCDD* (see figure 1 for an explanation of the strategy notation). In the electronic supplementary material, we show that for $k = 2$, all these additional coexistences are unstable with respect to the invasion of a third strategy. Only the coexistences discussed above for the case of $k = 1$ remain stable against all invasions.

The equilibrium that sustains cooperation resembles that arising in a volunteer's dilemma [43], where the volunteering threshold $k$, represents the number of contributors required to establish a punishment institution. We also note that the possibility of these type of coexistences has been discussed in the context of nonlinear public good games [44]. Here, we also show that this type of coexistence is particularly resilient and can arise with many flavours in the context of pool punishment, i.e. different combinations of conditional strategies that can resist invasions arising from a large strategy space. Our simulations also show that this equilibrium is stable when the population is finite and includes demographic noise.

# 4. Discussion

We find that pro-social, but not anti-social punishment emerge based on individual level selection with full symmetry between the two kinds of punishment institutions. Evolutionary dynamics introduces a symmetry breaking between the two kind of strategies selected that favours pro-social states. The prevalent outcome is a stable coexistence between cooperators supporting a pro-social pool and those willing to cooperate when such institutions are in place, but not cooperating otherwise. For peer punishment, such stable coexistences between strategies do not appear, because they require an honest signal prior to the game [45,46].

Typical models in evolutionary game theory restrict themselves to a small number of strategies that seem to be interesting from the outset, which greatly facilitates the analysis [47]. While such a restriction can be highly insightful, the conclusions from the model can in some cases strongly depend on the strategy set [23,24,48–50]. Our computational model implements the entire possible set of strategies in the context of coordinated, institutional punishment. One may be tempted to exclude strategies that seem illogical in the context of the model, but this can be misguided. Selecting a behaviour necessarily implies that other behaviours are driven out. Therefore, the absence of behaviour should be the result of evolutionary competition and not a result of the modeller's subjective choice. As shown in this paper, strategies can have a strong impact on the evolutionary outcome even if they do not rise to high average abundances and only temporarily pave the way for other strategies. We believe this robustness test is important in simple models, which may otherwise have biased conclusions.

In our model, the maintenance of pro-social punishment relies on a small minority which supports them, but their robustness stems from the fact that they are constantly challenged by the presence of players that stop cooperating in the absence of pro-social institutions. This combination is empirically prevalent, with resilient public goods often being supported by a minority of contributors [51]. The key to the maintenance of the public good is the observability of institutions: Only public knowledge of the presence of punishment institutions allows the conditional strategies that ultimately prevent the rise of anti-social behaviour, either in the form of defection or in the form of coordinated anti-social punishment institutions that never rise to high abundance, but can undermine the public good.

An interesting problem arising here is the possible effect of institutional asymmetries. For example, antisocial institutions may entail higher costs than prosocial ones, or implement asymmetric fines in which contributors and non-contributors are punished differently. In particular, anti-social institutions may offer an evolutionary advantage if they level intrinsic asymmetries between players.

Our model is primarily concerned with the emergence and establishment of primitive institutions, thus we do not explicitly model implementation details. Instead, we assume that once implemented, institutions for punishment will work as expected. In reality, additional issues may also arise, e.g. due to corruption [8,52,53], group heterogeneity [54]. Asymmetries in particular, have been shown to be important in infinite populations [55]. The tools necessary to study finite evolutionary dynamics arising from asymmetric games are not fully developed yet [56].

Most modern societies put law enforcement into the hands of institutions and do not allow their citizens to punish others directly. Our model suggests that our instinct against taking the law into our own hands is justified: The value of the signal conferred by the presence of pro-social punishment institutions may be crucial in promoting the kind of cooperation observed in humans.

Data accessibility. Our simulation code is available at https://osf.io/6tjmb/.

Authors' contributions. J.G. and A.T. developed the model together and wrote the paper together. J.G. performed numerical simulations, A.T. performed analytical approximations.

Competing interests. We declare we have no competing interests. Q1

Funding. We gratefully acknowledge partial funding by the Group of Eight Australia and DAAD (grant no. G8DAADGA14).

Acknowledgements. We thank the department of Evolutionary Theory, Jorge Peña and XYZ for fruitful discussions and three anonymous reviewers for their constructive criticism.

## References

1. Bowles S. 2004 *Microeconomics – behavior, institutions, and evolution*. Princeton, NJ: Princeton University Press.
2. Ostrom E. 1999 *Governing the commons*. Cambridge, UK: Cambridge University Press.
3. Baldassari D, Grossman G. 2011 Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc. Natl Acad. Sci. USA* **108**, 11 023–11 027. (doi:10.1073/pnas.1105456108)
4. Zhang B, Li C, De Silva H, Bednarik P, Sigmund K. 2013 The evolution of sanctioning institutions: an experimental approach to the social contract. *Exp. Econ.* **17**, 285–303. (doi:10.1007/s10683-013-9367-7)
5. Powers ST, Ekárt A, Lewis PR. 2018 Modelling enduring institutions: the complementarity of evolutionary and agent-based approaches. *Cogn. Syst. Res.* **52**, 67–81. (doi:10.1016/j.cogsys.2018.04.012)
6. Sigmund K, De Silva H, Traulsen A, Hauert C. 2010 Social learning promotes institutions for governing the commons.. *Nature* **466**, 861–863. (doi:10.1038/nature09203)
7. Sigmund K, Hauert C, Traulsen A, De Silva H. 2011 Social control and the social contract: the emergence of sanctioning systems for collective action. *Dyn. Games Appl.* **1**, 149–171. (doi:10.1007/s13235-010-0001-4)
8. Abdallah S, Sayed R, Rahwan I, LeVeck BL, Cebrian M, Rutherford A, Fowler JH. 2014 Corruption drives the emergence of civil society. *J. R. Soc. Interface* **11**, 20131044. (doi:10.1098/rsif.2013.1044)
9. Ostrom E. 1990 *Governing the commons: the evolution of institutions for collective action*. Cambridge, UK: Cambridge University Press.
10. Dreber A, Rand DG, Fudenberg D, Nowak MA. 2008 Winners don't punish. *Nature* **452**, 348–351. (doi:10.1038/nature06723)
11. Gächter S, Renner E, Sefton M. 2008 The long-run benefits of punishment. *Science* **322**, 1510. (doi:10.1126/science.1164744)
12. Pacheco JM, Santos FC, Souza MO, Skyrms B. 2009 Evolutionary dynamics of collective action in N-person stag hunt dilemmas. *Proc. R. Soc. B* **276**, 315–321. (doi:10.1098/rspb.2008.1126)
13. Raihani NJ, Bshary R. 2011 The evolution of punishment in n-player public goods games: a volunteer's dilemma. *Evolution* **65**, 2725–2728. (doi:10.1111/j.1558-5646.2011.01383.x)
14. Gavrilets S, Fortunato L. 2014 A solution to the collective action problem in between-group conflict with within-group inequality. *Nat. Commun.* **5**, 3256. (doi:10.1038/ncomms4526)
15. Fehr E, Gächter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)
16. Henrich J *et al.* 2006 Costly punishment across human societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)
17. Gürerk Ö, Irlenbusch B, Rockenbach B. 2006 The competitive advantage of sanctioning institutions. *Science* **312**, 108–111. (doi:10.1126/science.1123633)
18. Rockenbach B, Milinski M. 2006 The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718–723. (doi:10.1038/nature05229)
19. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via freedom to coercion: the emergence of costly punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)
20. De Silva H, Hauert C, Traulsen A, Sigmund K. 2010 Freedom, enforcement, and the social dilemma of strong altruism. *J. Evol. Econ.* **20**, 203–217. (doi:10.1007/s00191-009-0162-8)
21. Hilbe C, Traulsen A. 2012 Emergence of responsible sanctions without second order free riders,

antisocial punishment or spite. *Nat. Sci. Rep.* **2**, 458. (doi:10.1038/srep00458)

22. Herrmann B, Thöni C, Gächter S. 2008 Antisocial punishment across societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)
23. Rand DG, Nowak MA. 2011 The evolution of antisocial punishment in optional public goods games. *Nat. Commun.* **2**, 434. (doi:10.1038/ncomms1442)
24. García J, Traulsen A. 2012 Leaving the loners alone: evolution of cooperation in the presence of antisocial punishment. *J. Theor. Biol.* **307**, 168–173. (doi:10.1016/j.jtbi.2012.05.011)
25. Gächter S, Schulz JF. 2016 Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499. (doi:10.1038/nature17160)
26. Boyd R, Gintis H, Bowles S. 2010 Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* **328**, 617–620. (doi:10.1126/science.1183665)
27. Traulsen A, Röhl T, Milinski M. 2012 An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proc. R. Soc. B* **279**, 3716–3721. (doi:10.1098/rspb.2012.0937)
28. Schoenmakers S, Hilbe C, Blasius B, Traulsen A. 2014 Sanctions as honest signals – the evolution of pool punishment by public sanctioning institutions. *J. Theor. Biol.* **356**, 36–46. (doi:10.1016/j.jtbi.2014.04.019)
29. Hilbe C, Traulsen A, Röhl T, Milinski M. 2014 Democratic decisions establish stable authorities that overcome the paradox of second-order punishment. *Proc. Natl Acad. Sci. USA* **111**, 752–756. (doi:10.1073/pnas.1315273111)
30. Yamagishi T. 1986 The provision of a sanctioning system as a public good. *J. Pers. Soc. Psychol.* **51**, 110–116. (doi:10.1037/0022-3514.51.1.110)
31. Hauert C, De Monte S, Hofbauer J, Sigmund K. 2002 Volunteering as Red Queen mechanism for cooperation in public goods games. *Science* **296**, 1129–1132. (doi:10.1126/science.1070582)
32. Nowak MA, Sasaki A, Taylor C, Fudenberg D. 2004 Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650. (doi:10.1038/nature02414)
33. Traulsen A, Shoresh N, Nowak MA. 2008 Analytical results for individual and group selection of any intensity. *Bull. Math. Biol.* **70**, 1410–1424. (doi:10.1007/s11538-008-9305-6)
34. Fudenberg D, Imhof LA. 2008 Monotone imitation dynamics in large populations. *J. Econ. Theory* **140**, 229–245. (doi:10.1016/j.jet.2007.08.002)
35. Wu B, Gokhale CS, Wang L, Traulsen A. 2012 How small are small mutation rates? *J. Math. Biol.* **64**, 803–827. (doi:10.1007/s00285-011-0430-8)
36. Lin MJ. 2009 More police, less crime: evidence from US state data. *Int. Rev. Law Econ.* **29**, 73–80. (doi:10.1016/j.irle.2008.12.003)
37. Gambetta D. 2011 *Codes of the underworld: how criminals communicate*. Princeton, NJ: Princeton University Press.
38. Traulsen A, Santos FC, Pacheco JM. 2009 Evolutionary games in self-organizing populations. In *Adaptive Networks: Theory, Models and Applications* (eds T Gross, H Sayama). Berlin, Germany: Springer.
39. Traulsen A, Claussen JC, Hauert C. 2005 Coevolutionary dynamics: from finite to infinite populations. *Phys. Rev. Lett.* **95**, 238701. (doi:10.1103/PhysRevLett.95.238701)
40. Boyd R, Mathew S. 2007 A Narrow Road to Cooperation. *Science* **316**, 1858–1859. (doi:10.1126/science.1144339)
41. Palfrey TR, Rosenthal H. 1984 Participation and the provision of discrete public goods: a strategic analysis. *J. Public Econ.* **24**, 171–193. (doi:10.1016/0047-2727(84)90023-9)
42. Nöldeke G, Peña J. 2018 The Olson conjecture for discrete public goods. Available at SSRN.
43. Diekmann A. 1985 Volunteer's Dilemma. *J. Conflict Resolut.* **29**, 605–610. (doi:10.1177/0022002785029004003)
44. Archetti M, Scheuring I, Hoffman M, Frederickson ME, Pierce NE, Yu DW. 2011 Economic game theory for mutualism and cooperation. *Ecol. Lett.* **14**, 1300–1312. (doi:10.1111/j.1461-0248.2011.01697.x)
45. Raihani NJ, Bshary R. 2015 The reputation of punishers. *Trends Ecol. Evol.* **30**, 98–103. (doi:10.1016/j.tree.2014.12.003)
46. Jordan JJ, Hoffman M, Bloom P, Rand DG. 2016 Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476. (doi:10.1038/nature16981)
47. McNamara JM. 2013 Towards a richer evolutionary game theory. *J. R. Soc. Interface* **10**, 20130544. (doi:10.1098/rsif.2013.0544)
48. Lindgren K. 1991 Evolutionary phenomena in simple dynamics. In *Artificial Life II. SFI Studies in the Science of Complexity, vol. X* (eds CG Langton, C Taylor, JD Farmer, S Rasmussen). pp. 295–312. Redwood City: Addison-Wesley.
49. Burtsev M, Turchin P. 2006 Evolution of cooperative strategies from first principles. *Nature* **440**, 1041–1044. (doi:10.1038/nature04470)
50. dos Santos M. 2015 The evolution of anti-social rewarding and its countermeasures in public goods games. *Proc. R. Soc. B* **282**, 20141994. (doi:10.1098/rspb.2014.1994)
51. Schonmann RH, Boyd R. 2016 A simple rule for the evolution of contingent cooperation in large groups. *Phil. Trans. R. Soc. B* **371**, 20150099. (doi:10.1098/rstb.2015.0099)
52. Verma P, Sengupta S. 2015 Bribe and punishment: an evolutionary game-theoretic analysis of bribery. *PLoS ONE* **7**, e0133441. (doi:10.1371/journal.pone.0133441)
53. Lee JH, Iwasa Y, Dieckmann U, Sigmund K. 2019 Social evolution leads to persistent corruption. *Proc. Natl Acad. Sci. USA* **116**, 13 276–13 281. (doi:10.1073/pnas.1900078116)
54. Diekmann A, Przepiorka W. 2015 Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. *Sci. Rep.* **5**, 10321. (doi:10.1038/srep10321)
55. He JZ, Wang RW, Li YT. 2014 Evolutionary stability in the asymmetric volunteer's dilemma. *PLoS ONE* **9**, e103931. (doi:10.1371/journal.pone.0103931)
56. Broom M, Rychtář J. 2013 *Game-theoretical models in biology*. Chapman and Hall/CRC.