The take-it-or-leave-it option allows small penalties to overcome social dilemmas

Tatsuya Sasaki[a], Åke Brännström[a,b], Ulf Dieckmann[a], and Karl Sigmund[a,c,1]

[a]Evolution and Ecology Program, International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]Department of Mathematics and Mathematical Statistics, Umeå University, 90187 Umeå, Sweden; and [c]Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria

Edited by Kenneth Wachter, University of California, Berkeley, CA, and approved December 2, 2011 (received for review September 15, 2011)

Self-interest frequently causes individuals engaged in joint enterprises to choose actions that are counterproductive. Free-riders can invade a society of cooperators, causing a tragedy of the commons. Such social dilemmas can be overcome by positive or negative incentives. Even though an incentive-providing institution may protect a cooperative society from invasion by free-riders, it cannot always convert a society of free-riders to cooperation. In the latter case, both norms, cooperation and defection, are stable: To avoid a collapse to full defection, cooperators must be sufficiently numerous initially. A society of free-riders is then caught in a social trap, and the institution is unable to provide an escape, except at a high, possibly prohibitive cost. Here, we analyze the interplay of (*a*) incentives provided by institutions and (*b*) the effects of voluntary participation. We show that this combination fundamentally improves the efficiency of incentives. In particular, optional participation allows institutions punishing free-riders to overcome the social dilemma at a much lower cost, and to promote a globally stable regime of cooperation. This removes the social trap and implies that whenever a society of cooperators cannot be invaded by free-riders, it will necessarily become established in the long run, through social learning, irrespective of the initial number of cooperators. We also demonstrate that punishing provides a "lighter touch" than rewarding, guaranteeing full cooperation at considerably lower cost.

punishment | rewards | public goods | social contract | evolutionary games

In many species, cooperation has evolved through natural selection. In human societies, it can additionally be promoted through institutions. Institutions may be viewed as "tools that offer incentives to enable humans to overcome social dilemmas," to paraphrase Ostrom (1). The threat of punishment or the promise of reward can induce self-interested players to prefer actions that sustain the public good, and turn away from free-riding (2–13).

It is easy to understand the outcome of public good games in terms of the size of the incentive. If the incentive is too small, it has no effect and selfish players keep defecting by refraining from contributing to the public good (Fig. 1*a*). If, on the other hand, the incentive is sufficiently large, it compels all players to cooperate by contributing to the public good (Fig. 1*d*). It is the range of intermediate incentives that is of interest, and here, the effects of positive and negative incentives differ. Rewarding causes the stable coexistence of defectors and cooperators, with a larger proportion of cooperators when rewards are higher (Fig. 1*b*). Punishing, in contrast, leads to alternative stable states. As a result of the competition between cooperators and defectors, one or the other behavior will become established, but there can be no long-term coexistence (Fig. 1*c*). Whatever behavior prevails initially becomes fully established. Thus, each of the two behaviors may be viewed as a social norm: As long as the others stick to it, it does not pay to deviate. In particular, when cooperators are initially rare, the population will remain trapped in the asocial norm, with everyone defecting. Social learning cannot lead, in that case, to the more beneficial, prosocial norm of cooperating.

Here, we show that the option to abstain from the joint enterprise (14–17) offers an escape from the social trap. Indeed, when free-riding is the norm, players will turn away from unpromising joint ventures. This leads to the decline of exploiters and allows the reemergence of cooperators. If the incentives are too low, this is followed by the comeback of defectors, in a rock-paper-scissors type of cycle (18, 19) (Fig. 2*a*). However, even a modest degree of punishment breaks the rock-paper-scissors cycle and allows the fixation of the cooperative norm (Fig. 2 *e*–*g*). Thus, optional participation allows a permanent escape from the social trap. In contrast, we show that optional participation has little impact on rewarding systems (Fig. 2 *b*–*d*).

Methods

Specifically, we apply evolutionary game theory (20) to cultural evolution, based on (*a*) social learning (i.e., the preferential imitation of more successful strategies) and (*b*) occasional exploratory steps (modeled as small and rare random perturbations). Because the diversity of public good interactions and sanctioning mechanisms is huge, we first present a fully analytical investigation of a prototypical case (*SI Text*). We posit a large, well-mixed population of players. From time to time, a random sample of $n \geq 2$ players is faced with an opportunity to participate in a public good game, at a cost $g > 0$. We denote by m the number of players willing to participate ($0 \leq m \leq n$) and assume that $m \geq 2$ players are required for the game to take place. If it does, each of the m players decides whether or not to contribute a fixed amount $c > 0$, knowing that it will be multiplied by r (with $1 < r < n$) and distributed equally among all $m - 1$ other members of the group. If all group members invest into the common pool, each obtains a payoff $(r - 1)c - g$, which we assume to be positive. The social dilemma arises because players can improve their payoffs by not contributing. If all do so, each obtains the negative payoff $-g$. Thus, they would have done better to refrain from participation.

We now introduce the incentive. It is convenient to write the total incentive stipulated by an authority ("the institution") in the form mI, where I is the per capita incentive. If rewards are used, the total incentive will be shared among those players who cooperated. Hence, each cooperator obtains a reward mI/m_C, where m_C denotes the number of cooperators among the m players. If penalties are used, players who defect have their payoffs analogously reduced by mI/m_D, where m_D denotes the number of defectors among the m players. We will see that in the compulsory case, there exist two alternative stable norms for intermediate strength of punishment. In particular, a homogeneous population of defectors is unable to escape from the social trap (Fig. 1). In the optional case, cultural evolution leads to a stable homogeneous population of cooperators (Fig. 2 *e*–*g*), irrespective of the initial number of cooperators. Thus, voluntary participation overcomes the social trap plaguing the compulsory case. Remarkably, this is achieved at a fraction $1/n$ of the cost necessary in the compulsory case (*SI Text, S2*).

We base our analysis of the underlying evolutionary game on replicator dynamics (e.g., 20) for the three strategies C (cooperators), D (defectors), and

Author contributions: T.S., Å.B., U.D., and K.S. designed research; T.S. performed research; T.S. analyzed data; and T.S., Å.B., U.D., and K.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: karl.sigmund@univie.ac.at.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1115219109/-/DCSupplemental.

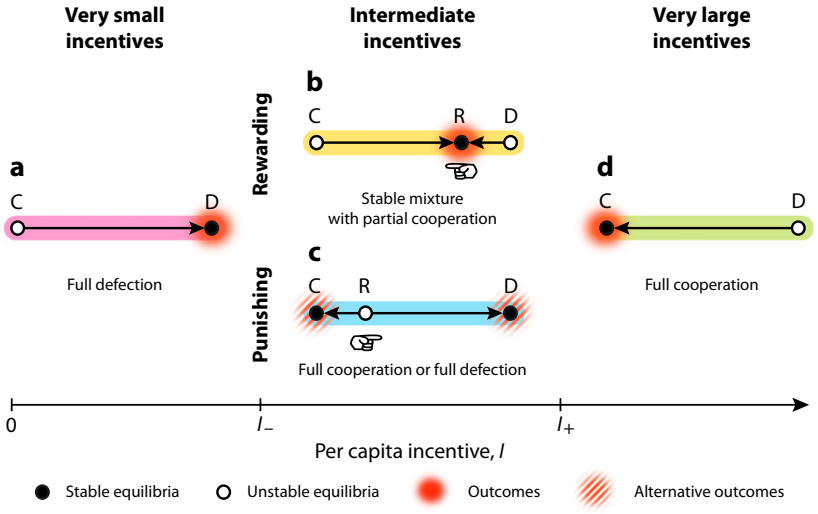

Fig. 1. Effects of institutional rewarding and punishing on the compulsory public good game for different per capita incentives I. For rewarding and punishing, full cooperation requires large incentives, even though the transition from full defection to full cooperation differs for the two types of incentive (*b* and *c*). (*a*) If I is smaller than $I_- = c/n$, the incentives have no effect on the outcome of the public good game and defection prevails. (*d*) If I is larger than $I_+ = c$, the incentives reverse the outcome and cooperation prevails. (*b* and *c*) For intermediate incentive I, rewarding leads to the stable coexistence of cooperation and defection, whereas punishing leads to alternative stable states. C and D correspond to the two homogeneous states in which the population consists exclusively of cooperators and defectors, respectively. With increasing incentive I, the equilibrium R moves toward C in the case of rewarding and toward D in the case of punishing.

N (nonparticipants), with frequencies x, y, and z. The state space Δ is the triangle of all (x, y, z) with $x, y, z \geq 0$ and $x + y + z = 1$. If $0 < g < (r - 1)c$, these three strategies form a rock-scissors-paper cycle in the absence of incentives, as shown in Fig. 2*a*: D beats C, N beats D, and C beats N. In the interior of the state space, all trajectories of the replicator dynamics originate from, and converge to, the state N of nonparticipation ($z = 1$) (21). Hence, cooperation can only emerge in brief bursts, sparked by random perturbations. The long-term payoff is that of nonparticipants (i.e., 0).

Results

If the game is compulsory [i.e., if all n players are obliged to participate ($z = 0$)], the outcome changes with increasing per capita incentive I (Fig. 1). For small I, defection dominates. The replicator dynamics have two equilibria: one stable (a homogeneous population of D-players) and one unstable (a homogeneous population of C-players). In the case of rewarding, as I crosses the threshold $I_- = c/n$, the equilibrium D becomes unstable, spawning a stable equilibrium R at a mixture of C- and D-players. As I increases further, the fraction of cooperators becomes larger and larger. Finally, when I reaches the threshold $I_+ = c$, the stable mixture merges with the formerly unstable equilibrium C, which becomes stable. In the case of punishing, as I crosses the threshold I_-, it is the unstable equilibrium C that becomes stable, spawning an unstable equilibrium R at a mixture of C- and D-players. R thus separates the regions of attraction of the equilibria C and D. With increasing I, the region of attraction of D becomes smaller and smaller, until I attains the value I_+. Here, the unstable equilibrium R merges with the formerly stable equilibrium D, which becomes unstable. For larger values of I, everyone cooperates. As shown in *SI Text, S2*, the values of I_+ and I_- are the same, irrespective of whether we consider rewarding or punishing.

We next investigate the interplay of (*a*) institutional incentives and (*b*) optional participation. Clearly, if the public good game is too expensive [i.e., if $g \geq (r - 1)c + I$ in the case of rewarding or $g \geq (r - 1)c$ in the case of punishing], players will opt for nonparticipation. We do not further consider this trivial case.

We first examine the case of punishing, for increasing per capita incentives I. For $I < I_-$, the effect of the incentive is negligible and all trajectories converge to N. As I crosses the threshold I_-, the equilibrium R appears on the CD-edge. At first, it is a saddle point. A trajectory leading from N to R separates the interior of the triangle Δ into two regions (Fig. 2*e*). One region is filled with trajectories issuing from N and converging to C, and the other is filled with trajectories issuing from and returning to N. If we assume that arbitrarily small random perturbations can, from time to time, affect the population (corresponding to occasional individual explorations of an alternative strategy), we see that the population will eventually end up at the stable equilibrium C. If I increases beyond a threshold K_-, an equilibrium Q enters the triangle Δ at R through a saddle-node bifurcation. With increasing I, the point Q moves along a straight line to N, whereas R keeps moving, along the CD-edge, to D (Fig. 2*f*). In *SI Text*, we show that Q is the unique equilibrium in the interior of the state space Δ (i.e., with all three strategies present) and that it is a saddle point. If I increases still further and crosses a threshold K_+, the equilibrium Q exits the triangle Δ through N. The point R becomes a source and remains so until it merges with D (for $I = I_+$) (Fig. 2*g*). Almost all trajectories in Δ either converge directly to C or to N. However, N is not stable. If the population is in the vicinity of N, arbitrarily small and rare random perturbations will eventually send it into the region of attraction of C. Hence, the population ultimately settles at the stable equilibrium C whenever $I > I_-$. This means that as soon as a homogeneous population of cooperators is immune against invasion by rare defectors, it becomes established in the long run.

In the case of rewarding, for $I < I_-$, the incentive has a negligible effect and all trajectories converge to N. As I crosses the threshold I_-, the equilibrium R appears on the CD-edge. Again, it is a saddle, but a trajectory now leads away from R to N (Fig. 2*b*). It separates a region where all trajectories lead from D to N from a region filled with trajectories issuing from and returning to N. As I increases and crosses a threshold J_-, a saddle-node bifurcation occurs at R, spawning an equilibrium Q into the triangle Δ (Fig. 2*c*). Again, one can show that this interior equilibrium is unique, and is a saddle point (*SI Text*). If I crosses a threshold J_+, the equilibrium Q exits the triangle Δ through N. All trajectories in the interior of the triangle Δ converge to R

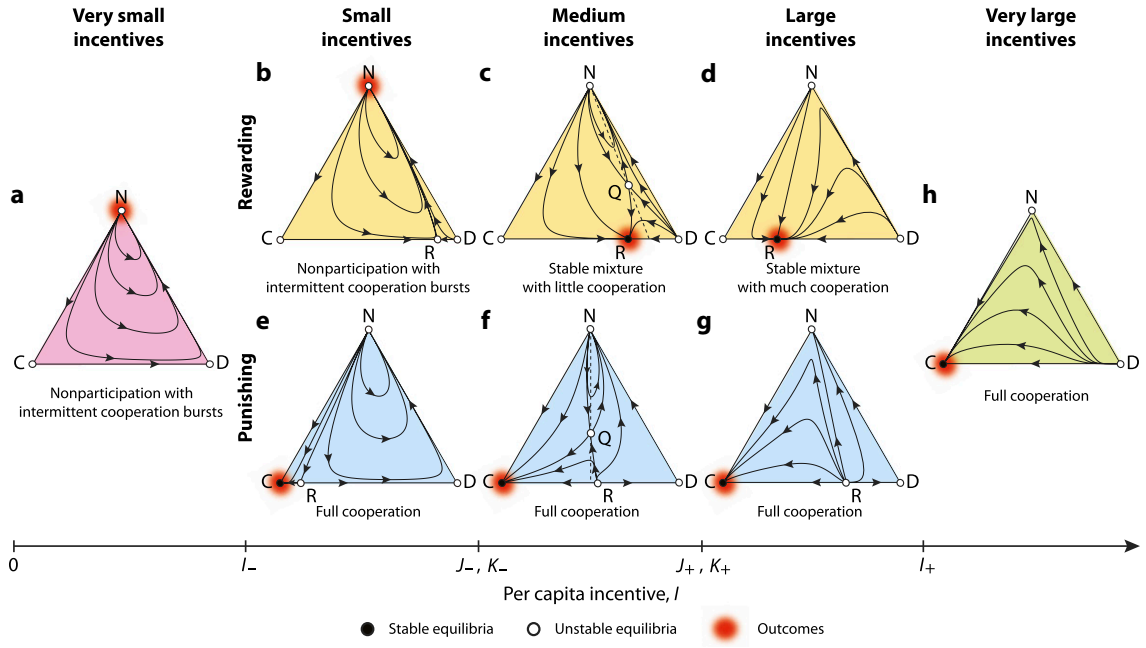

Fig. 2. Effects of institutional rewarding and punishing on the optional public good game for different per capita incentives I. Combining punishing with optional participation enables full cooperation for a small fraction of the cost needed in the compulsory case. The triangles Δ represent the state space $\Delta = \{(x, y, z)$: $x, y, z \geq 0$, $x + y + z = 1\}$, where x, y, and z are the frequencies of cooperators, defectors, and nonparticipants, respectively. The three vertices C, D, and N correspond to the three homogeneous states in which the population consists exclusively of cooperators ($x = 1$), defectors ($y = 1$), or nonparticipants ($z = 1$). (*a*) If I is smaller than $I_- = c/n$, the incentives have no effect on the outcome of the public good game. The interior of the triangle Δ is filled with trajectories issuing from and converging to the vertex N of nonparticipation in the joint enterprise. In that state, arbitrarily small random perturbations lead to short bursts of cooperation, immediately subverted by defection and followed by a return to nonparticipation. (*h*) If I is larger than $I_+ = c$, the incentives alter the outcome and cooperation prevails. All trajectories converge to C, the state of full cooperation. For the range of incentives in between *a* and *h*, the impacts of rewards and penalties differ. Rewarding: (*b*) For $I_- < I < J_-$, the equilibrium R on the CD-edge is a saddle point. All trajectories in the interior of the triangle Δ lead to N. (*c*) For $J_- < I < J_+$, an interior saddle point Q moves, with increasing I, along the dashed line from the CD-edge to N. Trajectories either converge to R, now a sink, or else to N. From there, an arbitrarily small random perturbation will send the state into the region of attraction of R. (*d*) For $J_+ < I < I_+$, the interior equilibrium Q has exited through N, and all trajectories converge to R, implying stable coexistence of defectors and cooperators. Punishing: (*e*) For $I_- < I < K_-$, the equilibrium R on the CD-edge is a saddle point. A trajectory from N to R separates a region where all trajectories lead to C from a region where all trajectories lead to N. An arbitrarily small random perturbation of N can lead to the region of attraction of C, and hence to the fixation of full cooperation. (*f*) For $K_- < I < K_+$, an interior saddle point Q moves, with increasing I, along the dashed line from the CD-edge to N. R is now a source. (*g*) For $K_+ < I < I_+$, the interior equilibrium Q has exited through N. In *f* and *g*, trajectories converge to C, either directly, or after a small random perturbation away from N. In summary, combining punishing with optional participation causes full cooperation from any initial condition for per capita incentives exceeding I_-, whereas combining rewarding with optional participation achieves this only for per capita incentives exceeding I_+. Parameters: $n = 5$, $r = 3$, $c = 1$, $g = 0.5$, and $I = 0$ (*a*); 0.25 (*b* and *e*); 0.35 (*c*); 0.55 (*f*); 0.7 (*d* and *g*); or (punishment) 1.2 (*h*).

(Fig. 2*d*). As I increases beyond I_+, the stable equilibrium R merges with C and all trajectories converge to C, just as in the case of punishment (Fig. 2*h*).

For enhancing a group's welfare, rewarding obviously works better than punishing (just as in the classic behaviorist analysis of reinforcements). However, the price of the rewarding has to be substantial. Punishing can achieve all-out cooperation (in the long run) for a much smaller price, namely, I_- (which is smaller the larger the group). From the viewpoint of institutionalizing a sanctioning mechanism, punishing thus has an advantage over rewarding: It achieves a higher average payoff at lower costs.

So far, we have treated g (the price an individual is willing to pay to participate in a joint enterprise) and I (the per capita size of the total incentive) as independent parameters. However, if individuals can freely decide whether or not to participate in the game, it makes sense to assume that they pay for the institution providing the incentives. For instance, I could be some fraction of the entrance cost g, or (equivalently) the total entrance cost could be viewed as the sum $g + aI$ of a part g kept by the authority and a part aI used for the incentive, with $a > 0$ (it is natural to assume that this part is proportional to the per capita incentive I). A rewarding system, if $a = 1$, simply redistributes the payoff without increasing group welfare, whereas a punishing system decreases it even if no one has to be punished. (We have to pay for the costly apparatus of law enforcement even if no one defaults.)

In the case of rewarding, optional participation increases the group welfare only marginally to 0 (Fig. 3*b*), for the small range $I_- < I < J_-$, where compulsory participation leads to negative average payoffs. In that range, combining rewarding with optional participation even reduces the cooperator frequency to 0 (Fig. 3*a*). For punishing, the situation is very different. The group welfare is highest when I is just barely larger than the minimum $I_- = c/n$ required to obtain full cooperation (Fig. 3*d*). The learning process, in that case, will take some time, and the population can undergo violent oscillations between the N-, C-, and D-states; however, in the end, the C-norm will prevail (Fig. 3*c*).

In *SI Text*, we test by extensive numerical investigations the robustness of our analytical results with respect to alternative model variants:

i) If we assume that part of the contribution to the public good returns to the contributing player, the dynamics become more complex but the evolutionary outcome remains unchanged (*SI Text, S3* and Figs. S1 and S2).

ii) Requiring participants to pay a fee for the sanctioning system also has little effect on the predicted outcome, as long as

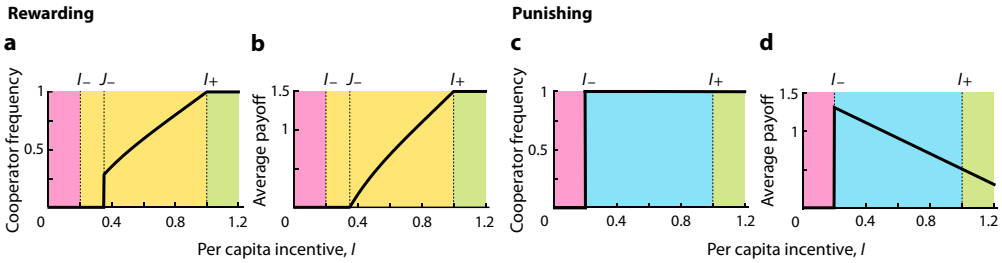

Fig. 3. "User-pays" variant. In this variant, players are obliged to pay an entrance fee $g + aI$. The panels show cooperator frequencies (*a* and *c*) and long-term average payoffs in the population (*b* and *d*), for rewarding (*a* and *b*) and punishing (*c* and *d*) and different per capita incentives I. Parameters: $n = 5$, $r = 3$, $c = 1$, $a = 1$, and $g = 0.5$.

this fee does not become unreasonably large (Fig. 3 and *SI Text, S5*).

iii) Moreover, when unused fees are returned, small negative per capita incentives suffice to maximize social welfare (*SI Text, S5*).

We can also model the sanctioning system in different ways. Results remain unchanged as long as reward, or punishment, decreases with the number of free-riders:

iv) This is the case, for instance, if only one defector is exemplarily punished, because the probability for being singled out decreases [in the old Navy, the slowest sailor was liable to get "prompted," (i.e., beaten)] (*SI Text, S4*).

v) It also holds whenever the institution needs to spend some resource (e.g., time) to punish a convicted free-rider. Indeed, this diminishes the resources to hunt for other free-riders. Such a "handling time" [to borrow an expression from predator–prey models (22)] will reduce the average punishment expected per defector, which is proportional to $mI/(a + bm_D)$, with $a, b > 0$ (*SI Text, S4*).

vi) Also, the capping of individual penalties leaves our qualitative findings unchanged (*SI Text, S4*). For these and related scenarios, optional participation leads to the establishment of full cooperation whenever the sanction is strong enough to deter free-riders from invading.

Surprisingly, in all cases we have considered, the cost of the negative incentive required to establish a norm of full cooperation is a small fraction of the cost needed in the case of compulsory participation.

Discussion

In his famous *Leviathan*, published in 1651, Hobbes stressed the necessity of an authority to curb the selfish motivations of individuals. He attributed its existence to a social contract intended to promote the commonwealth. Here, we assume that such a Leviathan-like authority exists, and is able to provide sanctions in the form of penalties and rewards. Indeed, most of our joint enterprises are protected by an elaborate apparatus of regulations, controls, and contract-enforcement devices to provide the necessary coercion. The theory of the social contract is a major topic in political philosophy, and a rich field of applications for game theory (e.g., 13).

The large majority of economic experiments and theoretical studies dealing with sanctions use peer-punishment, and thus make do without Leviathan, at least at first sight. Players can decide, independent of each other, whether to punish coplayers or not. This setting is of particular interest for investigating how prosocial coercion evolved, out of a world of anarchy (e.g., 1). Studies of peer-punishment attempt to address such a scenario (23–32). It seems clear, however, that in all economic experiments, Leviathan looms in the background. Players can pick their decisions, but usually only in a very narrow, regularized framework of alternatives. In modern human societies, anarchy is rare and players can almost always appeal to a higher authority.

There are many intermediate stages between pure peer-punishment and institutionalized punishment. Several authors have considered scenarios in which punishment is meted out only if two, or a majority, of players opt for it, or have allowed players to vote between treatments with or without peer-punishment (33–35). Thus, sanctions were supported by some social consensus, which can be mediated by communication ["cheap talk" (36)]. In other studies, players could contribute, before engaging in the public good game, to a punishment pool. This is like paying the wages of a police force before knowing whether, or against whom, it will be deployed (4, 37). Both theory and experiments have shown that delegating punishment is an efficient way to promote cooperation (38–40). Often, however, players of a public good game can engage in second-order free-riding by not paying toward the sanctions, which, in turn, raises the issue of second-order punishment. In our model, whoever wants to join the game has to pay an entrance fee. Second-order free-riding is no option, nor is asocial punishment targeted against cooperators (30). Leviathan sees to it.

The interplay of punishing, on the one hand, and optional participation, on the other hand, has already been investigated in several papers (21, 41–43). However, these studies mainly examined the problem of second-order free-riding. In contrast to these papers, we consider institutional punishment enforced by a higher authority. In our study, evolutionary game theory is applied to the implementation of an authority through social contract (by allowing individuals to voluntarily participate in a joint interaction). This establishes an interesting analogy with the suppression of competition occurring in several fields of evolutionary biology (e.g., "selfish genes") (44).

Voluntary submission under a sanctioning institution occurs in many real-life instances of cooperation. Practically all joint commercial and industrial enterprises are protected by enforceable contracts. Adherence is voluntary but commits the parties to mutually beneficial contributions. Punitive clauses ensure that noncompliance will be sanctioned. This principle also works, although at a less regulated level, in small-scale societies (1, 5, 38) and permits the sustainable use of common grazing or fishing grounds, or the construction and maintenance of irrigation systems. Medieval guilds delegated authority to chosen agents, and settlers hired sheriffs to deter villains. In day-to-day life, we may think of janitors, umpires, referees, or wardens who uphold rules in housing blocks, team games, private clubs, or public parks. All these examples rely on formal or informal agreements that can be freely joined but are then backed up by a higher authority. Thus, the situation we have addressed in our model is both fundamental and widespread.

ACKNOWLEDGMENTS. This study was enabled by financial support (TECT I-106 G11) from the Austrian Science Fund (to U.D.), through a grant for the research project The Adaptive Evolution of Mutualistic Interactions as part

of the multinational collaborative research project Mutualisms, Contracts, Space, and Dispersal (BIOCONTRACT) selected by the European Science Foundation as part of the European Collaborative Research (EUROCORES) Programme The Evolution of Cooperation and Trading (TECT). U.D. gratefully acknowledges additional support by the European Commission, the European Science Foundation, the Austrian Ministry of Science and Research, and the Vienna Science and Technology Fund. K.S. acknowledges support from TECT I-104 G1.

1. Ostrom E (2005) *Understanding Institutional Diversity* (Princeton Univ Press, Princeton).
2. Hardin G (1968) The tragedy of the commons. *Science* 162:1243–1248.
3. Olson E (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard Univ Press, Cambridge, MA).
4. Yamagishi T (1986) The provision of a sanctioning system as a public good. *J Pers Soc Psychol* 51:110–116.
5. Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge Univ Press, New York).
6. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57.
7. Camerer C (2003) *Behavioral Game Theory: Experiments in Strategic Interaction* (Russell Sage Foundation, New York).
8. Dickinson DL (2001) The carrot vs. the stick in work team motivation. *Exp Econ* 4: 107–124.
9. Henrich J, et al. (2006) Costly punishment across human societies. *Science* 312: 1767–1770.
10. Sigmund K (2007) Punish or perish? Retaliation and collaboration among humans. *Trends Ecol Evol* 22:593–600.
11. Skyrms B (2004) *The Stag Hunt and the Evolution of Social Structure* (Cambridge Univ Press, Cambridge, UK).
12. Sugden R (1998) *The Economics of Rights, Cooperation and Welfare* (Blackwell, Oxford).
13. Binmore KG (1994) *Playing Fair: Game Theory and the Social Contract* (MIT Press Cambridge, MA).
14. Orbell JM, Dawes RM (1993) Social welfare, cooperators' advantage, and the option of not playing the game. *Am Sociol Rev* 58:787–800.
15. Batali J, Kitcher P (1995) Evolution of altruism in optional and compulsory games. *J Theor Biol* 175:161–171.
16. Semmann D, Krambeck HJ, Milinski M (2003) Volunteering leads to rock-paper-scissors dynamics in a public goods game. *Nature* 425:390–393.
17. Sasaki T, Okada I, Unemi T (2007) Probabilistic participation in public goods games. *Proc Biol Sci* 274:2639–2642.
18. 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.
19. Hauert C, De Monte S, Hofbauer J, Sigmund K (2002) Replicator dynamics for optional public good games. *J Theor Biol* 218:187–194.
20. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, UK).
21. De Silva H, Hauert C, Traulsen A, Sigmund K (2009) Freedom, enforcement, and the social dilemma of strong altruism. *J Evol Econ* 20:203–217.
22. Holling CS (1959) Some characteristics of simple types of predation and parasitism. *Can Entomol* 91:385–398.
23. Boyd R, Richerson P (1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol Sociobiol* 13:171–195.
24. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* 415:137–140.
25. Fehr E, Rockenbach B (2003) Detrimental effects of sanctions on human altruism. *Nature* 422:137–140.
26. Gardner A, West SA (2004) Cooperation and punishment, especially in humans. *Am Nat* 164:753–764.
27. Gürerk O, Irlenbusch B, Rockenbach B (2006) The competitive advantage of sanctioning institutions. *Science* 312:108–111.
28. Egas M, Riedl A (2008) The economics of altruistic punishment and the maintenance of cooperation. *Proc Biol Sci* 275:871–878.
29. Dreber A, Rand DG, Fudenberg D, Nowak MA (2008) Winners don't punish. *Nature* 452:348–351.
30. Herrmann B, Thöni C, Gächter S (2008) Antisocial punishment across societies. *Science* 319:1362–1367.
31. Casari M (2005) On the design of peer punishment experiments. *Exp Econ* 8:107–115.
32. Nakamaru M, Dieckmann U (2009) Runaway selection for cooperation and strict-and-severe punishment. *J Theor Biol* 257:1–8.
33. Boyd R, Gintis H, Bowles S (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328:617–620.
34. Ertan A, Page T, Putterman L (2009) Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *Eur Econ Rev* 53:495–511.
35. Kosfeld M, Okada A, Riedl A (2009) Institution formation in public goods games. *Am Econ Rev* 99:1335–1355.
36. Bochet O, Page T, Putterman L (2006) Communication and punishment in voluntary contribution experiments. *J Econ Behav Organ* 60:11–26.
37. Sigmund K, De Silva H, Traulsen A, Hauert C (2010) Social learning promotes institutions for governing the commons. *Nature* 466:861–863.
38. Poteete A, Janssen M, Ostrom E (2010) *Working Together: Collective Action, the Commons, and Multiple Methods in Practice* (Princeton Univ Press, Princeton).
39. O'Gorman R, Henrich J, Van Vugt M (2009) Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proc Biol Sci* 276:323–329.
40. Baldassarri D, Grossman G (2011) Centralized sanctioning and legitimate authority promote cooperation in humans. *Proc Natl Acad Sci USA* 108:11023–11027.
41. Fowler JH (2005) Altruistic punishment and the origin of cooperation. *Proc Natl Acad Sci USA* 102:7047–7049.
42. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K (2007) Via freedom to coercion: The emergence of costly punishment. *Science* 316:1905–1907.
43. Mathew S, Boyd R (2009) When does optional participation allow the evolution of cooperation? *Proc Biol Sci* 276:1167–1174.
44. Frank SA (1995) Mutual policing and repression of competition in the evolution of cooperative groups. *Nature* 377:520–522.