The Competitive Advantage of Sanctioning Institutions

Özgür Gürerk,[1] Bernd Irlenbusch,[2] Bettina Rockenbach[1]*

Understanding the fundamental patterns and determinants of human cooperation and the maintenance of social order in human societies is a challenge across disciplines. The existing empirical evidence for the higher levels of cooperation when altruistic punishment is present versus when it is absent systematically ignores the institutional competition inherent in human societies. Whether punishment would be deliberately adopted and would similarly enhance cooperation when directly competing with nonpunishment institutions is highly controversial in light of recent findings on the detrimental effects of punishment. We show experimentally that a sanctioning institution is the undisputed winner in a competition with a sanction-free institution. Despite initial aversion, the entire population migrates successively to the sanctioning institution and strongly cooperates, whereas the sanction-free society becomes fully depopulated. The findings demonstrate the competitive advantage of sanctioning institutions and exemplify the emergence and manifestation of social order driven by institutional selection.

The uniqueness of human cooperation necessitates investigations that reach beyond the explanations of cooperative behavior of nonhuman animals (*1*–*5*). Profound empirical evidence shows that the possibility of sanctioning norm violators stabilizes human cooperation at a high level, whereas cooperation typically collapses in the absence of sanctioning possibilities (*6*–*11*). Would a sanctioning institution deliberately be adopted when individuals can choose between a sanctioning and a sanction-free institution? The considerable payoff losses in the process toward stable cooperation—for both the punishers and the punished individuals—as well as natural resentments against punishment caused, for example, by its detrimental effects (*12*) might guide individuals' choice toward the sanction-free institution.

The argument that higher cooperation levels in sanctioning institutions "automatically" lead to their prevalence—because rational individuals choose the institution with the higher payoff (*13*)—is often brought forward as an affirmative argument for the competitive advantage of sanctioning institutions. The force of this argument can be questioned, however, because it displaces rather than solves the evolutionary puzzle of human cooperation. The reason for this is that stable cooperation requires a positive share of individuals who carry personal costs for cooperation and punishment to the benefit of the entire group (*14*–*16*). These individuals have a clear payoff disadvantage compared to cooperators who free-ride on the punishment acts. Recent research shows that a positive share of strong reciprocators—cooperating individuals who are willing to reward fair behavior and to punish unfair behavior even when they cannot gain materially from doing so—can be evolutionarily stable (*17*, *18*). But what happens if the population is perfectly mobile and is permanently invaded by outsiders from a noncooperative environment who are attracted by high payoffs from cooperation? Is the fraction of strong reciprocators who choose the sanctioning institution sufficiently large to keep up the cooperative culture? These arguments cast serious doubt on the prevalence of sanctioning institutions.

However, several affirmative arguments for the competitive advantage of sanctioning institutions also come to mind, e.g., the large number of institutional frameworks that facilitate the sanctioning of norm violators in human societies (*19*–*21*) and the recent finding that humans derive satisfaction from punishing defectors (*22*). Additionally, theories of cultural and institutional selection (*23*–*26*) that are grounded on the exceptional human ability of social learning support the competitive advantage of sanctioning institutions. They suggest that individuals preferentially migrate to groups with higher payoffs and imitate the decisions prevalent in these groups. Hence, group members punish, because it is common to do so. When cooperation is sufficiently widespread, the payoff-disadvantage from punishing is relatively small, and only a weak tendency for conformist behavior suffices to stabilize the punishment of noncooperators.

We inquire into the competitive advantage of sanctioning institutions in a laboratory experiment in which we implement permanent competition between a sanctioning and a sanction-free institution through endogenous choice. It allows one to study the evolution of the different institutions over time as well as the changes in behavior in the same individual when participating in different social settings.

In our experiment, 84 participants anonymously interact in a social dilemma situation in 30 repetitions. Each repetition consists of three stages: An institution choice stage (S0), a voluntary contribution stage (S1), and a sanctioning stage (S2). In stage S0, the participants simultaneously and independently choose between a sanctioning institution (SI) and a sanction-free institution (SFI) in which neither positive sanctioning (rewards) nor negative sanctioning (punishment) is possible. In stage S1, each participant interacts in a public goods game with all other participants who have chosen the same institution in S0; each player is endowed with 20 money units (MUs) and may contribute between 0 and 20 MUs to a public good. Each group member equally profits from the public good, independent of his or her own contribu-

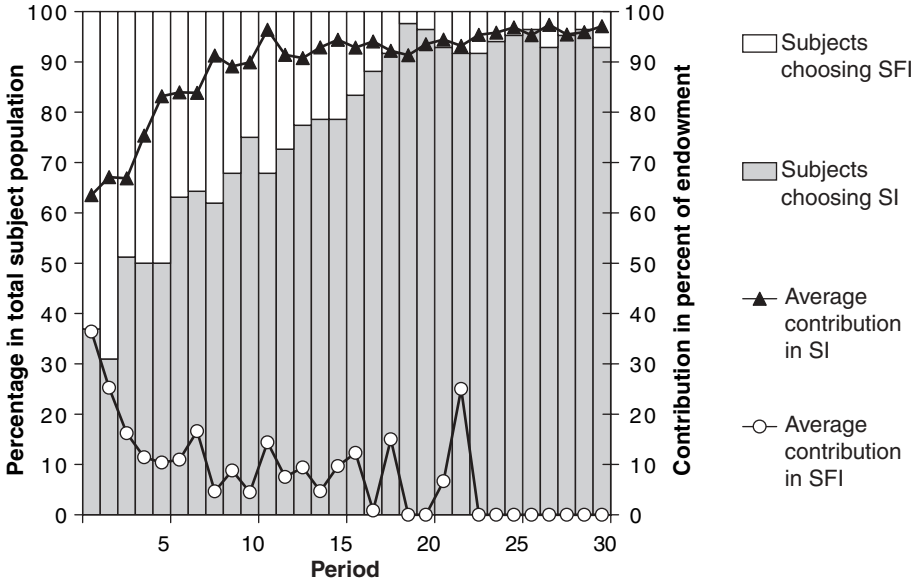

Fig. 1. Subjects' choice of institution and their contributions. The average contributions in both institutions over the 30 periods of the interaction are measured as the percentage of endowment contributed to the public good.

[1]University of Erfurt, Nordhäuser Strasse 63, 99089 Erfurt, Germany. [2]London School of Economics, Houghton Street, London WC2A 2AE, UK.

*To whom correspondence should be addressed. E-mail: bettina.rockenbach@uni-erfurt.de

tion. The MUs not contributed to the public good are transferred to the participant's private account. The diametrically opposed individual and collective interests constitute the social dilemma in public good provision: It is always in the material self-interest of any subject to free-ride on the contributions of others and to keep all MUs for the private account, whereas the collective interest demands full contribution of all group members. After the players have simultaneously made their contribution decisions, they are informed about the contributions of each member in their institution. In stage S2 each player in SI may positively or negatively sanction other members of SI by assigning between 0 and 20 tokens to other members. Each token used as a negative sanction costs the punished member 3 MUs and the punishing member 1 MU. Each token used as a positive sanction yields the receiving member 1 MU and costs the member who uses it 1 MU. At the end of the period each participant receives detailed (but anonymous) information about each of the other participants from both institutions (*27*).

The initial choice of institution provides a clear picture: Only about one-third of the participants (mean = 36.9%; SE = 4.0%) prefer SI to SFI in the first period. The revealed institution preference correlates with different types of behavior (*28*, *29*). Participants who initially join SI contribute on average 12.7 MUs (SE = 0.79) in the first period, while on average only 7.3 MUs (SE = 0.54) are contributed in SFI (Wilcoxon signed rank matched pairs test, $z = -2.366$, $P = 0.016$, two-tailed). Almost half the subjects (mean = 48.4%; SE = 8.5%) who opt for SI in the first period are "high contributors" in that they contribute at least 15 MUs. Almost three-fourths (mean = 73.3%; SE = 17.0%) of these high contributors exert punishment tokens to discipline low contributors and thus try to enforce and establish a norm of high cooperation. These subjects amount to 13.1% (SE = 4.0%) of the total subject population and can clearly be classified as "strong reciprocators," i.e., subjects with a predisposition to make high contributions and to punish norm violators. In contrast, 16.1% (SE = 5.2%) of the subjects in SI contribute 5 MUs or less ("free-riders") in the first period. The situation is completely different in SFI, where in the first period almost half of the subjects are free-riders (mean = 43.4%; SE = 3.4%), whereas high contributors are rare (mean = 11.3%; SE = 4.3%). A subject who chooses SFI in the first period with a contribution of more than 15 MUs and uses negative sanctions immediately after having switched to SI may also be classified as a strong reciprocator. We observed two subjects with this behavior in our subject population (2.4%), so that 15.5% (SE = 5.6%) is a lower bound for the proportion of strong reciprocators in the subject population. Initially, the significantly higher contributions in SI do not result in higher payoffs in SI: Average payoffs in the first period of SI (mean = 38.1 MUs; SE = 2.05) are significantly lower than in SFI (mean = 44.4; SE = 0.32) (Wilcoxon signed rank matched pairs test, $z = -2.047$, $P = 0.047$, two-tailed). Due to frequent punishment activities, free-riders earn significantly less in SI (mean = 30.2; SE = 4.51) than in SFI (mean = 49.7 MUs; SE = 0.86) in the first period (Wilcoxon signed rank matched pairs test, $z = -2.366$, $P = 0.016$, two-tailed).

Although subjects are initially reluctant to join SI, it becomes predominant over time; eventually, nearly all participants (mean = 92.9%; SE = 3.4%) choose SI and cooperate fully (Fig. 1) (*30*). Simultaneously, contributions in SFI decrease to zero. In period 10 the contributions in SI are on average 89.9% (SE = 10.3%) of the endowment and from there on they steadily increase. In the last period the difference between the two institutions is almost as extreme as it can be with average contributions of 19.4 MUs (SE = 0.714) in SI and 0 MUs (SE = 0.0) in SFI. Averaged over all periods, subjects in SI contribute 18.3 MUs (91.4% of the endowment; SE = 5.0%), whereas subjects in SFI contribute only 2.9 MUs (14.4% of the endowment; SE = 3.0%) (Wilcoxon signed rank matched pairs test, $z = -2.366$, $P = 0.016$, two-tailed).

What causes this dramatic change of mind? Pure imitation of the successful behavior would lead to an increase of free-riders in SFI because they earn the highest average payoffs in the first period. This is actually observed in period two. Consequently, the payoffs of free-riders in SFI decrease and over the periods, participants in SFI experience the typically observed collapse of cooperation in repeated social dilemma interactions (Fig. 1). A comparison of the payoffs of the two predominant behavioral patterns—free-riding in SFI and high contributions in SI (Fig. 2)—shows that from period five onward a high contributor in SI achieves a higher payoff than a free-rider in SFI (Wilcoxon signed rank matched pairs test, $z = -2.366$, $P = 0.016$, two-tailed). It therefore pays for a monetary payoff maximizing participant to switch from free-riding in SFI to contributing in SI. This triggers an amplifying effect; namely, the greater the number of cooperators in SI, the higher their payoffs. Indeed, from period 10 onward, 86.1% (SE = 13.1%) of all members of SI contribute fully (20 MUs) and 86.0% (SE = 8.6%) in SFI contribute almost nothing (2 MUs or less). The finding that players apparently choose institutions according to payoffs indicates that stochastic

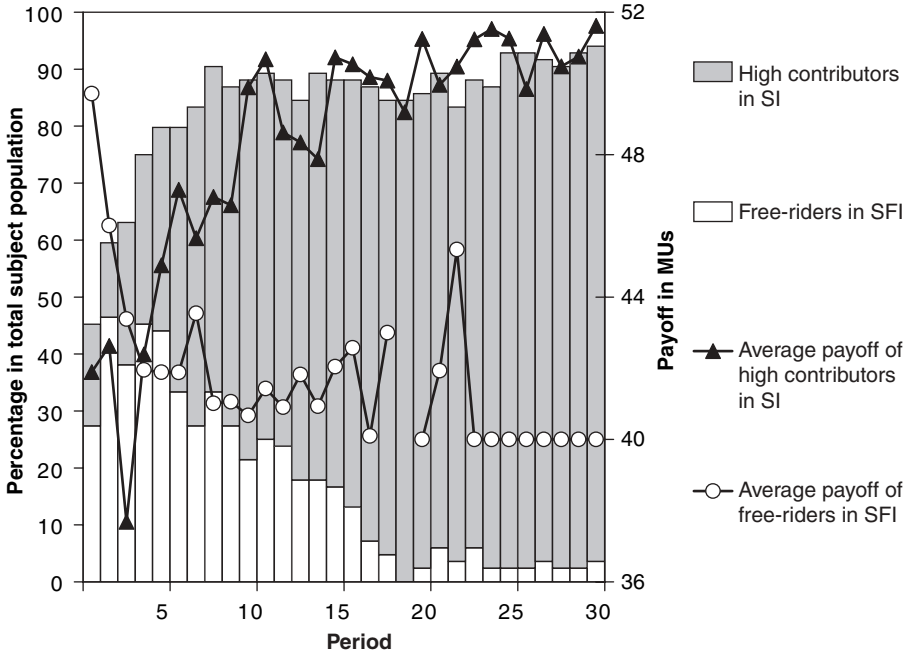

Fig. 2. Payoffs of the two predominant behavioral patterns, "free-riders" (contributions between 0 and 5 MUs) in the sanction-free institution (SFI) and "high contributors" (contributions between 15 and 20 MUs) in the sanctioning institution (SI). The highest attainable payoff (under full contributions of all subjects and no punishment) is 52 MUs and the payoff from complete free-riding and no punishment is 40 MUs.

Table 1. Results of a Tobit regression, independent variable: Contribution ($t + 1$) – Contribution (t). Tobit regression for subjects who opted for SI in period t and ($t + 1$) with a robust estimation for the standard errors using the independent observations as clusters. The values in parentheses denote the robust standard errors.

Independent variable	Coefficient	z value
Negative sanctions in t	0.444 (0.085)	5.24*
Positive sanctions in t	−0.148 (0.102)	−1.45
Constant	0.000 (0.053)	0.00

*Denotes significance at the 1% level.

forces play only a minor role in determining switching behavior (*31*).

A closer look at individual behavior immediately before and after migration from one institution to the other confirms the bipolar pattern of behavior induced by the two institutions. Indeed, 80.3% (SE = 5.0%) of subjects increase their contribution when migrating from SFI to SI in two consecutive periods. Moreover, 27.1% (SE = 5.3%) of subjects even "convert" from being a complete free-rider (contributing 0 MUs) to a full cooperator (contributing 20 MUs) when switching from SFI to SI. The migration behavior in the opposite direction, i.e., from SI to SFI, is similarly extreme. Roughly 70% (mean = 70.9%; SE = 4.9%) of subjects reduce their contribution when switching from SI to SFI and about 20% (mean = 17.0%; SE = 4.7%) switch from full cooperation to free-riding.

Individual payoff maximization cannot explain why new members in SI follow the second norm established by the strong reciprocators who joined SI in early periods, i.e., the norm to punish low contributors. The most successful behavior would be to contribute in SI (and hence avoid being punished), but refrain from the costly punishment of others. Because punishment of defectors constitutes a second-order public good (in which defection cannot be sanctioned in our setting), individual payoff maximization would rule out punishment. However, only a minority of subjects follow this payoff-maximizing behavior. The overwhelming majority of 62.9% (SE = 8.5%) of the subjects immediately conforms to and adopts the prevailing norm of punishment in SI, i.e., they always use punishment immediately after they switch to SI. This results in a quite stable proportion of ~40% (mean = 42.1%; SE = 5.9%) of subjects who both contribute highly and punish during the last 20 periods (Fig. 3). Figure 3 also shows that the payoff difference between high contributors who punish and those who do not constantly diminishes over time because punishment becomes ever more unnecessary. Additionally, because the absolute number of punishers increases, the individual burden from effectively punishing free-riders becomes smaller over time (*32*). Toward the end, subjects who both contribute highly and punish exhibit a payoff disadvantage of less than 2%; hence, the "selection pressure" against strong reciprocators becomes quite weak (*33*). This leads to a continuous increase in efficiency gains in SI up to 95.8% (SE = 4.6%) in the final period, whereas efficiency gains in SFI converge to zero (mean = 0; SE = 0.0).

Although the use of both positive and negative sanctions per individual decreases over time, the ratio in which they are used is rather stable; on average, 1.66 negative sanction points (SE = 0.60) are allocated per positive sanction point. A Tobit regression of the combined effect of positive and negative sanctions exhibits a clear positive impact of punishment on subsequent contributions, whereas positive sanctions have a slightly negative but rather insignificant effect (Table 1). It seems that positive sanctions are not perceived as an unambiguous encouragement to increase the contribution; perhaps they are taken as an indication that the contribution has been higher than expected by others and hence may be lowered. These observations reflect the asymmetry between negative and positive sanctions. Positive sanctions are addressed to those who already abide by the social norm and, to preserve the approval of cooperation, a continuous application of the instrument is required. Negative sanctioning, by contrast, is an instrument for disapproving of norm-violating behavior and need only be exerted if the norm is not followed. If an individual abides by the norm, punishment is not necessary. The threat of punishment alone is able to support cooperation.

Our results show that the sanctioning institution is the undisputed winner in a "voting-with-one's-feet" competition with a sanction-free institution. The results provide profound empirical evidence for the existence and importance of strong reciprocators, as well as a form of conformist behavior, as described in models of cultural selection. The initial establishment of the "norm to cooperate and punish free-riders" is mainly driven by the steadfastness of the strong reciprocators to punish noncooperative subjects, despite severe individual losses (*34*). Although strong reciprocators are a minority, they manage to establish and enforce a cooperative culture that attracts even previously noncooperative individuals and thus resolves the social dilemma. The predominant tendency to punish norm violators after a migration from the noncooperative environment of the sanctioning-free institution to the sanctioning institution provides support for the assumption that humans adapt to the common behavior although it deviates from the payoff-maximizing behavior. This tendency for conformism raises sanctioning activities at a high level such that cooperation can be stabilized.

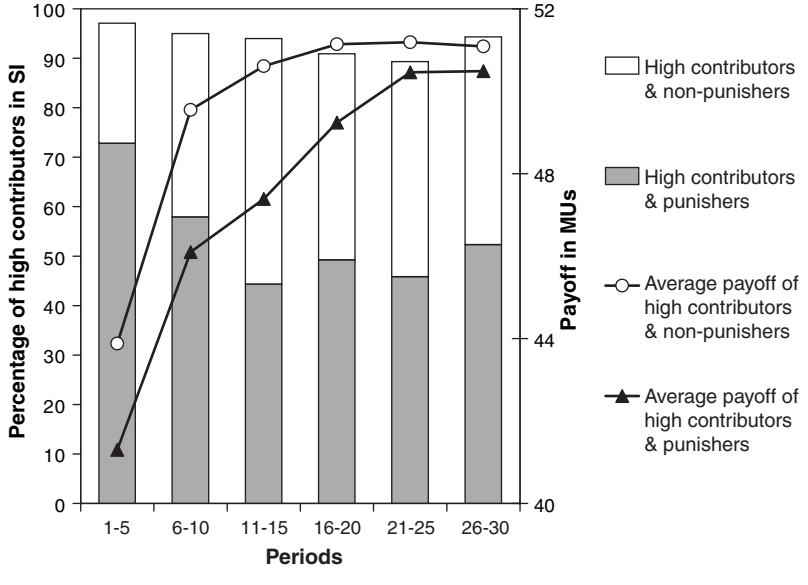

Fig. 3. Payoffs and percentages of punishers and nonpunishers among the "high contributors" (contributions between 15 and 20 MUs) in the sanctioning institution (SI). The highest attainable payoff (under full contributions of all subjects and no punishment) is 52 MUs and the payoff from complete free-riding and no punishment is 40 MUs.

References and Notes

1. J. R. Stevens, M. D. Hauser, *Trends Cogn. Sci.* **8**, 60 (2004).
2. E. Fehr, U. Fischbacher, *Nature* **425**, 785 (2003).
3. J. Henrich *et al.*, *Am. Econ. Rev.* **91**, 73 (2001).
4. E. Ostrom, J. Burger, C. B. Field, R. B. Norgaard, D. Policansky, *Science* **284**, 278 (1999).
5. P. Hammerstein, *Genetic and Cultural Evolution of Cooperation* (MIT Press, Cambridge, MA, 2003).
6. T. Yamagishi, *J. Pers. Soc. Psychol.* **51**, 110 (1986).
7. E. Fehr, S. Gächter, *Nature* **415**, 137 (2002).
8. E. Ostrom, J. Walker, R. Gardner, *Am. Polit. Sci. Rev.* **86**, 404 (1992).
9. J. R. Andreoni, W. T. Harbaugh, L. Vesterlund, *Am. Econ. Rev.* **93**, 893 (2003).
10. D. Masclet, C. Noussair, S. Tucker, M.-C. Villeval, *Am. Econ. Rev.* **93**, 366 (2003).
11. M. S. Rege, K. Telle, *J. Public Econ.* **88**, 1625 (2004).
12. E. Fehr, B. Rockenbach, *Nature* **422**, 137 (2003).
13. K. Binmore, *Natural Justice* (Oxford Univ. Press, Oxford, 2005).
14. H. Gintis, *J. Theor. Biol.* **206**, 169 (2000).
15. E. Fehr, U. Fischbacher, S. Gächter, *Hum. Nat.* **13**, 1 (2002).
16. H. Gintis, S. Bowles, R. Boyd, E. Fehr, *Evol. Hum. Behav.* **24**, 153 (2003).
17. R. Boyd, H. Gintis, S. Bowles, P. J. Richerson, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3531 (2003).
18. S. Bowles, H. Gintis, *Theor. Popul. Biol.* **65**, 17 (2004).
19. N. Q. Mahdi, *Ethol. Sociobiol.* **7**, 295 (1986).
20. A. W. Johnson, T. Earle, *The Evolution of Human Societies: From Foraging Group to Agrarian State* (Stanford Univ. Press, Stanford, CA, 1987).
21. P. Wiessner, *Hum. Nat.* **16**, 115 (2005).
22. D. J.-F. DeQuervain *et al.*, *Science* **305**, 1254 (2004).
23. R. Boyd, P. J. Richerson, *Ethol. Sociobiol.* **13**, 171 (1992).
24. J. Henrich, R. Boyd, *J. Theor. Biol.* **208**, 79 (2001).
25. R. Boyd, P. J. Richerson, *J. Theor. Biol.* **215**, 287 (2002).

26. J. Henrich, *J. Econ. Behav. Org.* **53**, 3 (2004).
27. Materials and methods are available as supporting material on *Science* Online.
28. U. Fischbacher, S. Gächter, E. Fehr, *Econ. Lett.* **71**, 397 (2001).
29. R. Kurzban, D. Houser, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1803 (2005).
30. Figure S1 displays the exact flow in both directions between institutions from one period to the next.
31. H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton Univ. Press, Princeton, NJ, 1998).
32. A logistic regression shows that the stay duration in SI in terms of the number of periods has a significantly negative influence on the likelihood of punishing others (table S1). Note, however, that individually exerted punishment may be lowered over time to effectively punish a free-rider because the number of potential punishers becomes larger. Indeed, average payoffs of free-riders decrease over periods, as can be seen from fig. S2.
33. In the last 10 periods, subjects who contribute highly and punish reach on average 98.7% of the payoff of subjects who contribute highly but do not punish.
34. C. Camerer, E. Fehr, *Science* **311**, 47 (2006).

35. We thank S. Bowles, E. Fehr, U. Fischbacher, S. Gächter, H. Gintis, G. Harrison, J. Henrich, M. Peacock, and R. Selten for helpful comments.

Supporting Online Material

www.sciencemag.org/cgi/content/full/312/5770/108/DC1
Materials and Methods
Figs. S1 and S2
Table S1
References

9 December 2005; accepted 14 February 2006
10.1126/science.1123633

Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins

Daniel M. Weinreich,* Nigel F. Delaney,† Mark A. DePristo, Daniel L. Hartl

Five point mutations in a particular β-lactamase allele jointly increase bacterial resistance to a clinically important antibiotic by a factor of ~100,000. In principle, evolution to this high-resistance β-lactamase might follow any of the 120 mutational trajectories linking these alleles. However, we demonstrate that 102 trajectories are inaccessible to Darwinian selection and that many of the remaining trajectories have negligible probabilities of realization, because four of these five mutations fail to increase drug resistance in some combinations. Pervasive biophysical pleiotropy within the β-lactamase seems to be responsible, and because such pleiotropy appears to be a general property of missense mutations, we conclude that much protein evolution will be similarly constrained. This implies that the protein tape of life may be largely reproducible and even predictable.

Resistance to β-lactam antibiotics (e.g., penicillin) is commonly mediated by a bacterial β-lactamase, which hydrolytically inactivates these drugs (*1*). Bacterial resistance to novel β-lactams first arises by point mutations in the β-lactamase gene (*2*, *3*). Five point mutations in an allele of this gene that we designate TEM^{wt} (the reference sequence of the TEM family of β-lactamases) (*4*, *5*) jointly increase resistance by a factor of ~100,000 against cefotaxime (*6*, *7*), a third-generation cephalosporin β-lactam. These consist of four missense mutations [A42G, E104K, M182T, and G238S; numbering as in (*8*)] at clinically important residues (*3*, *9*) and one 5′ noncoding mutation [g4205a; numbering as in (*4*)], and we denote this high-resistance quintuple mutant *TEM**. Thus, five mutations must occur for *TEM** to evolve from TEM^{wt}, and because these can in principle occur in any order, there are 5! = 120 mutational trajectories linking these alleles. However, natural selection for heightened cefotaxime resistance may not regard all trajectories equivalently (*10*). Here, we determine the prevalence with which these mutations only conditionally increase drug resistance, a form of interaction previously designated sign epistasis (*10*). Sign epistasis is both necessary and sufficient for one or more trajectories to *TEM** to be selectively inaccessible (*10*).

To characterize the effect on drug resistance of each mutation on all allelic backgrounds, we first constructed the 32 combinations of these five mutations (*11*, *12*). We next determined their resistance to cefotaxime (*12*) in *Escherichia coli* strain DH5α (Table 1); because the sign of the mutational effect on drug resistance determines the selective accessibility of each trajectory (*10*), we also report the rank order of drug resistance values exhibited by all alleles. *TEM** exhibits the highest resistance and, because at least one mutation increases resistance in all other alleles, the fitness landscape is single-peaked (*13*). Thus, in the case of cefotaxime resistance evolution, populations cannot become trapped (*13*) at suboptimal alleles between TEM^{wt} and *TEM**, as was recently also shown for isopropylmalate dehydrogenase (IMDH) evolution from a nicotinamide adenine dinucleotide phosphate (NADP)–dependent form to a nicotinamide adenine dinucleotide (NAD)–dependent form (*14*).

To estimate the relative probabilities with which evolution by natural selection for heightened cefotaxime resistance will realize each of the 120 possible mutational trajectories from TEM^{wt} to *TEM**, we assumed that the time to fixation or loss of individual mutations is far less than the time between mutations [the "strong selection/weak mutation" model of (*15*)]. Thus, the relative probability of realizing any particular mutational trajectory is the product of the relative probabilities of its constituent mutations, because under our assumption the choice of each subsequent fixation is statistically independent of all previous fixations (*12*). Next, for each allele we partitioned all possible mutations into those that are beneficial, deleterious, or neutral with respect to cefotaxime resistance. The probability of

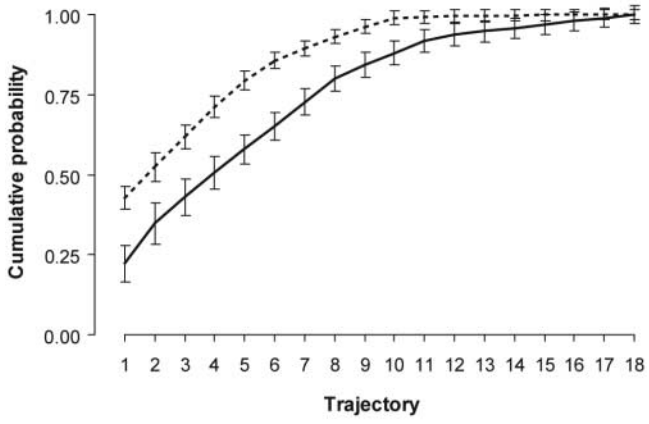

Fig. 1. Estimated cumulative probabilities for all 18 selectively accessible mutational trajectories from TEM^{wt} to *TEM**, under the correlated (broken line) and equal fixation probability (solid line) models, ± SEM. Trajectories are ordered in decreasing probability of realization.

Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.

*To whom correspondence should be addressed. E-mail: dmw@post.harvard.edu

†Present address: Integrative Oceanography Division, Scripps Institute of Oceanography, 9500 Gilman Drive, La Jolla, CA 92037, USA.

SOCIAL SCIENCE

Cooperation, Punishment, and the Evolution of Human Institutions

Given the choice, people prefer institutional arrangements in which those who overconsume common-property resources are punished compared to those in which they go free.

Joseph Henrich

Explaining the scale, diversity, and historical dynamics of human cooperation is increasingly bringing together diverse empirical and theoretical approaches. For decades, this challenge has energized evolutionary and economic researchers to ask: Under what conditions will decision-makers sacrifice their own narrow self-interest to help others? Although classic evolutionary models based on relatedness and reciprocity have explained substantial swaths of the cooperation observed in many species, including our own, theoretical work in the 1980s demonstrated that the puzzle of cooperation in large groups, or in situations without much repeated interaction, remained unsolved and would likely require alterative theoretical formulations (*1, 2*).

Enhanced online at www.sciencemag.org/cgi/content/full/312/5770/60

Such cooperative dilemmas, or "public goods" problems, involve situations in which individuals incur a cost to create a benefit for the group. In our society, think of recycling, buying a hybrid car, valor in combat, voting, and donating blood. The dilemma arises from free-riders who enjoy the group benefits created by the contributions of others without paying the costs. Even if nearly everyone is initially cooperative and contributes, free-riders can profit and proliferate, leading to the eventual collapse of cooperation. So, understanding how public goods problems can be solved has provoked great interest, both because human societies have somehow managed to solve many such problems to varying degrees, and because some of the world's most pressing issues, such as global climate change, are essentially public goods dilemmas. On page 108 of this issue, Gürerk *et al*. (*3*) take an important step in understanding how self-sustaining cooperative institutions may have emerged over the course of human history.

Recent models have demonstrated how evolutionary processes (genetic or cultural) can maintain cooperation in large groups or without repeated interaction. Costly signaling models have shown how cooperation by "high-quality individuals" (those who are potentially desirable as allies or mates) can be sustained if such individuals can accurately signal their quality by making substantial cooperative contributions to public goods (*4*). For example, great hunters might supply all the meat for a public feast, or millionaires might donate a recreational center to their community. Similarly, reputation-based models have shown how cooperation can be sustained if individuals' reputations for not contributing to public goods reduce their payoffs (or fitness) by altering how others treat them in certain dyadic social interactions (*5*). Finally, models that allow individuals to both contribute to the public good and to sanction noncontributors have revealed stable cooperative solutions, especially when the strategies for cooperation and punishment are influenced by social learning (*6*). Thus, a number of possible stable solutions to the puzzle of cooperation in large groups, or cooperation without repeated interaction, have now emerged.

Free-riders not wanted. Those who do not contribute but benefit from the efforts of others can cause the collapse of cooperation. Groups that sanction such free-riders stabilize cooperative behavior and outcompete groups that do not.

It turns out, however, that finding a stable solution is only the first step in confronting the dilemma of cooperation. Each of the above approaches can actually stabilize any behavior or practice, independent of whether it delivers any benefit to anyone. This includes behaviors that reduce the payoff or fitness of the group. For example, instead of public goods contributions, costly signaling could maintain behaviors involving dangerous physical feats (like scaling icy mountain peaks), aggressive displays (like beating up your neighbor), or extravagantly wasteful feasts. Similarly, the same reputational and sanctioning mechanisms that can stabilize cooperation can also sustain maladaptive practices such as consuming the brains of dead relatives, flattening the foreheads of infants, or binding the feet of young girls. Thus, there are actually a multitude of stable equilibria, only some of which are cooperative. What determines which equilibria emerge and/or spread?

Three broad theoretical approaches confront the problem of equilibrium selection. The first, and perhaps the most intuitive, is that rational, forward-looking individuals recognize the long-term payoffs available at stable cooperative equilibria, assume others are similarly sensible, and choose the cooperative state (*7*). The second approach is based on the stochasticity inherent in any interaction. Different stable equilibria are more or less susceptible to this stochasticity, meaning that in the long-run, some equilibria will be substantially more common than others (*8*). The third mechanism, cultural group selection, gives priority to the competition among social groups who have arrived at different culturally evolved equilibria. This intergroup competition favors the spread of individuals and practices from groups stabilized at more cooperative equilibria. In humans, competition between groups can take the form of warfare, demographic production (some social groups reproduce faster than others), or more subtle forms in which individuals learn decisions and strategies by

The author is in the Department of Anthropology, Emory University, 1557 Dickey Drive, Atlanta, GA 30322, USA. E-mail: jhenric@emory.edu

CREDIT: J. SUTLIFF

Published by AAAS

preferentially observing more successful individuals, many of whom are more successful because they live in groups at stable cooperative equilibria (*9*). This can lead to a flow of decisions, strategies, and even preferences from more cooperative groups to less cooperative ones (*6*), or to a migration of individuals among groups (*10*) that favors the spread of the more cooperative equilibria.

Gürerk *et al*. address the issue of equilibrium selection with an elegant addition to the existing experimental work on public goods. In their experiment, individuals (the "players") choose between two different "institutions." In one institution, players can contribute money to a group project. The sum of all contributions to the project is augmented by a fixed percentage and then is divided equally among all players, regardless of their contributions. Previous experiments established that when this interaction is repeated, mean contributions to the public good drop to near zero (a noncooperative equilibrium). The other "sanctioning" institution is very similar, except that after players have contributed, they can pay to punish (reduce the payoff of) other players. When this interaction is played repeatedly (*11*) a substantial fraction of players punish low contributors, causing mean contributions to rise and stabilize near full cooperation (a cooperative equilibrium). Both institutions were run concurrently for 30 interactions and players could, initially and after each subsequent interaction (after seeing others' payoffs), choose their institution for the next interaction.

The principal findings of Gürerk *et al*. can be summarized simply. Initially, most players picked the institution without sanctioning possibilities. But, as usual, free-riders in the nonsanctioning institution started driving mean contributions downward, so cooperators, who hate being exploited by free-riders, started reducing their contributions. Meanwhile, in the sanctioning institution, punishers started driving contributions up by inflicting costs on noncontributors, despite the personal cost of punishing. After a few interactions, players from the nonsanctioning institution—presumably seeing the higher payoffs of those choosing the sanctioning institution—increasingly switched institutions. Notably, despite the incoming flow of migrants from the nonsanctioning institution, the mean contributions in the sanctioning institution consistently increased or held stable near full cooperation. In fact, most incoming migrants, consistent with local norms in their new setting, increased their contributions during their first interaction in the sanctioning institution, and a majority administered some punishment.

What does this tell us about equilibrium selection? First, the players' degree of rationality did not permit them to foresee the final outcome and select the higher payoff institution on the first interaction. Second, despite the stochasticity of human decisions, neither institution drifted to another equilibrium. What did happen is that once players from the lower payoff institution observed the higher payoffs of the other institution, they wanted to adopt either the practices of the higher payoff institution, or the decisions and strategies of those other players. Consistent with ethnographic and historical case studies (*12*, *13*), the present work provides an important experimental demonstration of cultural group selection in action, as the two alternative equilibria compete for shares of the total population.

The course charted by Gürerk *et al*. should spur more empirical work on how processes of equilibrium selection influence the evolution of institutional forms. Many questions remain to be tackled: for example, what happens if switching institutions is costly, or if information about the payoffs in the other institution is poor? Or, what happens if individuals cannot migrate between institutions, but instead can vote on adopting alternative institutional modifications? Such work can both help us understand how humans became such a cooperative species, and teach us how to build durable cooperative institutions that solve public goods problems and are readily spread.

References

1. R. Boyd, P. J. Richerson, *J. Theor. Biol.* **132**, 337 (1988).
2. N. V. Joshi, *J. Genet.* **66**, 69 (1987).
3. Ö. Gürerk, B. Irlenbusch, B. Rockenbach, *Science* **312**, 108 (2006).
4. H. Gintis, E. A. Smith, S. Bowles, *J. Theor. Biol.* **213**, 103 (2001).
5. K. Panchanathan, R. Boyd, *Nature* **432**, 499 (2004).
6. J. Henrich, R. Boyd, *J. Theor. Biol.* **208**, 79 (2001).
7. J. C. Harsanyi, R. Selton, *A General Theory of Equilibrium Selection in Games* (MIT Press, Cambridge, MA, 1988).
8. H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions* (Princeton Univ. Press, Princeton, NJ, 1998),.
9. R. Boyd, P. Richerson, *J. Theor. Biol.* **215**, 287 (2002).
10. R. Boyd, P. J. Richerson, *J. Theor. Biol.* **145**, 331 (1990).
11. E. Fehr, S. Gachter, *Am. Econ. Rev.* **90**, 980 (2000).
12. S. Bowles, *Microeconomics: Behavior, Institutions, and Evolution* (Princeton Univ. Press, Princeton, NJ, 2004).
13. P. J. Richerson, R. Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution* (Univ. of Chicago Press, Chicago, 2005).

10.1126/science.1126398

EVOLUTION

Reducible Complexity

Christoph Adami

How does biological complexity arise? The molecular evolution of two hormone receptors was traced from a common ancestral receptor. Through a series of mutations, receptors with distinct hormone binding properties evolved, one before the appearance of its cognate ligand.

If an elaborate lock fits an equally elaborate key, we immediately sense the purpose of design: The key was crafted with the idea of the lock in mind. We would not entertain the possibility that the match is accidental. When we come upon such lock-and-key pairs in nature, it is natural to ask how these pairs could have evolved via Darwinian evolution. At first glance, it seems that the key can only evolve to fit the lock if the lock is already present, and the lock cannot evolve except in the presence of the key (because without the key, it does not open). On page 97 of this issue, Bridgham *et al.* (*1*) take a closer look at this puzzle and discover a different answer in the molecular evolution of hormone-receptor interactions.

Charles Darwin was fully aware of the problems that such lock-and-key systems—should they exist in biology—would present to his theory because the theory relies upon step-by-step changes to a trait. Building a lock-and-key system appears to require at least two changes to happen simultaneously. He famously remarked that "if it could be demonstrated that any complex organ existed which could not possibly have been formed by numerous successive slight modifications, my theory would absolutely break down" (*2*). This concern has been seized upon by proponents of an "intelligent design" alternative to Darwinian evolution that proposes that complex systems—like those that display lock-and-key complexity—cannot evolve. The premise for the argument is that systems of a lock-and-key nature cannot evolve and are thus "irreducibly complex" (*3*), implying that only the lock-and-key combination, but not its parts, is complex. The argument continues that because such systems do exist in nature, and cannot have evolved, they must have been "designed."

Darwin already saw how such thorny issues could be resolved. He further explains in *The Origin of Species* that "if we look to an organ common to all the members of a large class…in order to discover the early transi-

The author is at the Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA. E-mail: adami@kgi.edu

Supporting Online Material for

The Competitive Advantage of Sanctioning Institutions

Özgür Gürerk, Bernd Irlenbusch, Bettina Rockenbach*

*To whom correspondence should be addressed.
E-mail: bettina.rockenbach@uni-erfurt.de

Published 7 April 2006, *Science* **312**, 108 (2006)
DOI: 10.1126/science.1123633

This PDF file includes:

The Competitive Advantage of Sanctioning Institutions

Supporting Online Material

Özgür Gürerk*, Bernd Irlenbusch[§] & Bettina Rockenbach*

** University of Erfurt, Nordhäuser Straße 63, 99089 Erfurt, Germany,*

[§] London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Correspondence:

Bettina Rockenbach, Chair in Microeconomics, University of Erfurt, Nordhäuser Straße 63, 99089 Erfurt, Germany. Phone: +49 361 7374521; fax: +49 361 7374529; e-mail: bettina.rockenbach@uni-erfurt.de

SUPPORTING ONLINE MATERIAL

Materials and Methods

Methods

84 undergraduate students from the University of Erfurt voluntarily participated in the experiments. Special care was exerted to recruit students from many different disciplines to increase the likelihood that the subjects had never met before. Each participant was allowed to take part in one session only. In total 7 experimental sessions each involving 12 subjects took place. These sessions constituted the independent observations for the non-parametric statistical analysis. Most of the sessions were run in pairs, i.e. 24 subjects were gathered in the lab.

The game is repeated over 30 periods and participants are not restricted by choices performed in previous periods. Each period consists of three stages: An institution choice stage (S0), a voluntary contribution stage (S1), and a sanctioning stage (S2). In stage S0 the participants simultaneously and independently choose between a sanctioning institution (SI) and a sanction-free institution (SFI) in which neither positive sanctioning (rewards) nor negative sanctioning (punishment) is possible. In stage S1, each participant is informed about the number of participants in each institution and in case the institution is occupied by at least two participants a public goods game is played with all participants who have chosen the same institution in S0. If only one subject joins an institution the subject's total endowment is automatically transferred to her/his private account. The public good's game constitutes a prototypical social dilemma in which each player is endowed with 20 money units (MUs) and may contribute between 0 and 20 MUs to a project which benefits the entire group. Each MU contributed to the public good is deducted from the contributor's private account and creates a benefit of 1.6 MUs for the entire group. This group benefit is equally distributed among the group members, i.e. if a group consists of n members each member profits by $1.6/n$ MUs from

each 1 MU contributed (1.6/n is the marginal per capita return MPCR). If, for example, only one group member contributes the total endowment of 20 and the other n–1 group members contribute nothing, the public good amounts to 20·1.6 and the contributor's profit is 20·1.6/n while each free-riders' profit is 20+20·1.6/n. If, however, all n group members contribute an identical amount of x, with $0 \leq x \leq 20$, the public good is $n \cdot x \cdot 1.6$ and each member achieves a profit of $20 - x + 1.6 \cdot x = 20 + 0.6 \cdot x$. Hence for an identical contribution x of all group members the net benefit of each group member is $0.6 \cdot x$ independent from the group size n. The MUs not contributed to the public good are transferred to the participant's private account. Thus, the provider's return from one additional MU is less than 1 but the group's return exceeds 1. Since the cost of providing is higher than the individual return, it is always in the material self-interest of any subject to free-ride on the contributions of the others and to keep all MUs for the private account. If all participants follow their material self-interest, nobody contributes to the public good and each participant achieves a payoff of 20 MUs. Because the group's return of each MU invested is greater than 1, it is in the collective interest that all group members contribute their entire endowment to the group project. These diametrically opposed individual and collective interests constitute the social dilemma in public good provision. After the players have simultaneously made their contribution decisions, they are informed about the contributions of each member in the own group.

At the beginning of stage S2 each player receives additional 20 tokens independent of the affiliation choice in S0. In SFI these tokens are directly transferred to the player's private account without any decisions required, i.e. sanctioning was not possible. In SI these tokens may be used to positively or negatively sanction other members of SI by assigning between zero and 20 tokens to other members. Each player is free to choose which of the other members of SI she/he wants to positively and/or negatively sanction and to determine the amount of allocated sanctioning tokens to each of those players. She/he is free to allocate different numbers of sanctioning tokens to different individuals with the only restriction that

the sum of allocated tokens is limited to at most 20. Tokens not used for sanctioning are transferred to the player's private account. Each token employed as a negative sanction costs the punished member 3 MUs and the punishing member 1 MU. Each token employed as a positive sanction yields the receiving member 1 MU and costs the employing member 1 MU. The leverage in the negative sanctioning mechanism is motivated by the understanding that punishment is more costly for the punished individual than for the punisher. We assume that the leverage in positive sanctioning is smaller and does not create any efficiency gains. The efficiency loss of negative sanctioning as well as the efficiency neutrality of positive sanctioning excludes efficiency gains solely by applying these instruments.

At the end of the period each participant receives detailed (but anonymous) information about each individual other participant from both institutions: the contribution, the sum of allocated positive sanctioning and negative sanctioning tokens to others, the sum of received positive sanctioning tokens from others, the sum of received negative sanctioning tokens from others, and the period profit. Players are neither informed about the identities of the other players nor are they able to track the identities over periods, because the order in which the players' details are displayed is known to be randomized in each period. In particular players could not identify the other players who allocated sanctioning tokens to them.

At the beginning of the experiment subjects received written instructions (see the section "Experimental instructions" below). At the end of the experiment subjects privately received their experimental earnings in cash. One experimental session typically lasted for 2.5 hours, and on average subjects earned 24 € per session. All experimental decisions were made on a computer screen using the experimental software z-Tree (*S1*). Each of the 24 computers was located in a booth such that subjects could not see or communicate with each other.

The Effect of the Group Size on the Marginal per Capita Return (MPCR)

The marginal per capita return (MPCR) denotes the individual return a recipient obtains from each token contributed. In a public goods game of n players the MPCR is lower than 1 and exceeds $1/n$. The MPCR being lower than 1 implies that it is individually rational to refrain from contributing since the individual return is lower than the investment. The fact that the MPCR exceeds $1/n$ implies that contributing is collectively rational because the groups' return on each token invested is greater than one. The endogenous group choice in each period of our experiment allows varying group sizes in each period. We constructed the MPCR such that it changes with the group size n, i.e. MPCR = $1.6/n$. Hence, smaller groups have a higher MPCR than larger groups. Thus, the more members choose an institution the lower is the individual return on investment for one contributed token. As a consequence, however, the total "productivity" $n \cdot$MPCR from the perspective of the group is constant, i.e. equal to 1.6. This means that all groups consisting of full cooperators achieve the same individual payoffs (i.e. $20 \cdot 1.6$), no matter how large the groups are. Hence, in the Nash-equilibrium of complete free-riding as well as under full cooperation the individual payoffs do not depend on the group size. From what is known on the interplay of the group size and the MPCR (*S2*) our design favors cooperation in small groups and disfavors cooperation in large groups.

Experimental Instructions

The next pages show a translation of the written experimental instructions. They describe the sequence of events during the experiment and the payoff rules. Subjects received the instructions in German.

Instructions to the Experiment

General Information: At the beginning of the experiment you will be randomly assigned to one of **2 subpopulations each consisting of 12 participants**. During the whole experiment you will interact only with the members of your subpopulation. At the beginning of the experiment, **1,000 experimental tokens** will be assigned to the experimental account of each participant.

Course of Action: The experiment consists of **30 rounds**. Each round consists of 2 stages. In Stage 1, the group choice and the decision regarding the contribution to the project take place. In Stage 2, participants may influence the earnings of the other group members.

Stage 1

(i) The Group Choice: In Stage 1, each participant decides which group she wants to join. There are two different groups that can be joined:

		Influence on the earnings of other group members
Group	A:	No
	B:	Yes, by assigning positive and negative tokens

(ii) Contribution to the Project: In stage 1 of each round, each group member is endowed with 20 tokens. You have to decide how many of the 20 tokens you are going to contribute to the project. The remaining tokens will be kept in your private account.

Calculation of your payoff in stage 1: Your payoff in stage 1 consists of two components:

- **tokens you have kept** = endowment – your contribution to the project
- **earnings from the project** = 1.6 x sum of the contributions of all group members / number of group members

Thus, **your payoff in Stage 1** amounts to:

20 – your contribution to the project

+ 1.6 x sum of the contributions of all group members / number of group members

The earnings from the project are calculated according to this formula for each group member. **Please note:** Each group member receives the same earnings from the project, i.e. each group member benefits from **all** contributions to the project.

Stage 2

Assignment of Tokens: In stage 2 it will be displayed how many tokens each group member contributed to the project. (**Please note: In each round the order of displaying the members' actions will randomly be determined.** Thus, it is not possible to identify any group member by her position on the displayed list throughout different rounds.) By the assignment of tokens in stage 2 you can increase or reduce the payoff of a group member or keep it unchanged.

In each round each participant receives additional 20 tokens in stage 2. You have to decide how many of these 20 tokens you are going to assign to other group members. The remaining tokens are kept in your private account. You can check the costs of your token assignment by pressing the button "Calculation of Tokens".

- Each **positive token** that you assign to a group member **increases her payoff by 1 token**.
- Each **negative token** that you assign to a group member **reduces her payoff by 3 tokens**.
- If you assign **0 tokens** to a group member her **payoff won't change**.

Calculation of your payoff in stage 2: Your payoff in stage 2 consists of three components:

- **tokens you kept in your private account** = 20 – sum of the tokens that you have assigned to the other group members
- **increased by the number of positive tokens** you have received from other group members
- **diminished by the threefold number of negative tokens** you have received from other group members

Thus, **your payoff in Stage 2** amounts to:

20 – sum of the tokens that you assigned to other group members

+ the number of positive tokens you received from other group members

– 3x (the number of negative tokens you received from other group members)

Calculation of your round payoff: Your round payoff is composed of

	Your payoff from Stage 1	20 – your contribution to the project + 1.6 x sum of the contributions of all group members / number of group members
+	Your payoff from Stage 2	20 – sum of the tokens that you have assigned to other group members + number of positive tokens you have received from other members – 3 x (the number of tokens you have received from other members)
=	Your round payoff	

Special case: a single group member: If it happens that you are the only member in your group you receive 20 tokens in Stage 1 and 20 tokens in Stage 2, i.e. your round payoff sums up to 40. You do not have to take any action neither on Stage 1 nor on Stage 2.

Information at the end of the round: At the end of the round you receive a detailed overview of the results obtained in all groups. For every group member you are informed about her: Contribution to the project, payoff from the Stage 1, assigned tokens (if possible), received positive tokens (if possible), received negative tokens (if possible), payoff from Stage 2, round payoff.

History: Starting from the 2nd round, in the beginning of a new round you receive an overview of the average results (as above) of all previous rounds.

Total Payoff: The total payoff from the experiment is composed of the initial endowment of 1,000 tokens plus the sum of round payoffs from all 30 rounds. At the end of the experiment your total payoff will be converted into Euro with an exchange rate of 1 € per 100 tokens.

Please notice: Communication is not allowed during the whole experiment. If you have a question please raise your hand out of the cabin. All decisions are made anonymously, i.e. no other participant is informed about the identity of someone who made a certain decision. The payment is anonymous too, i.e. no participant learns what the payoff of another participant is.

We wish you success!

Supporting Figures

Figure S1

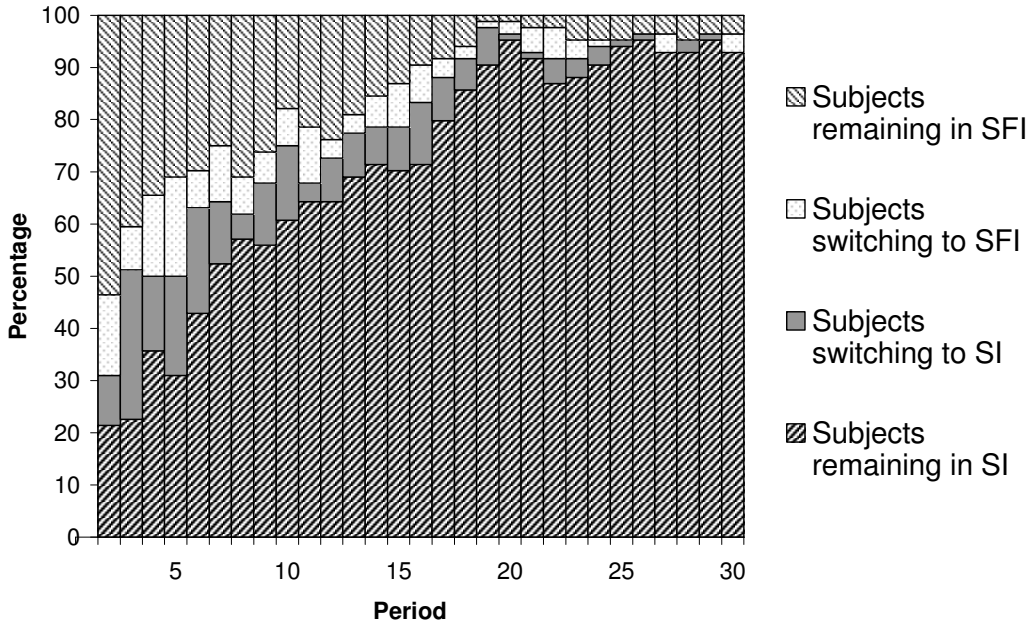

Fig. S1. Subjects' choices of institutions and their switching behavior in both directions.

Figure S1 displays the percentage of subjects who remain in the institutions or switch between the institutions.

Figure S2

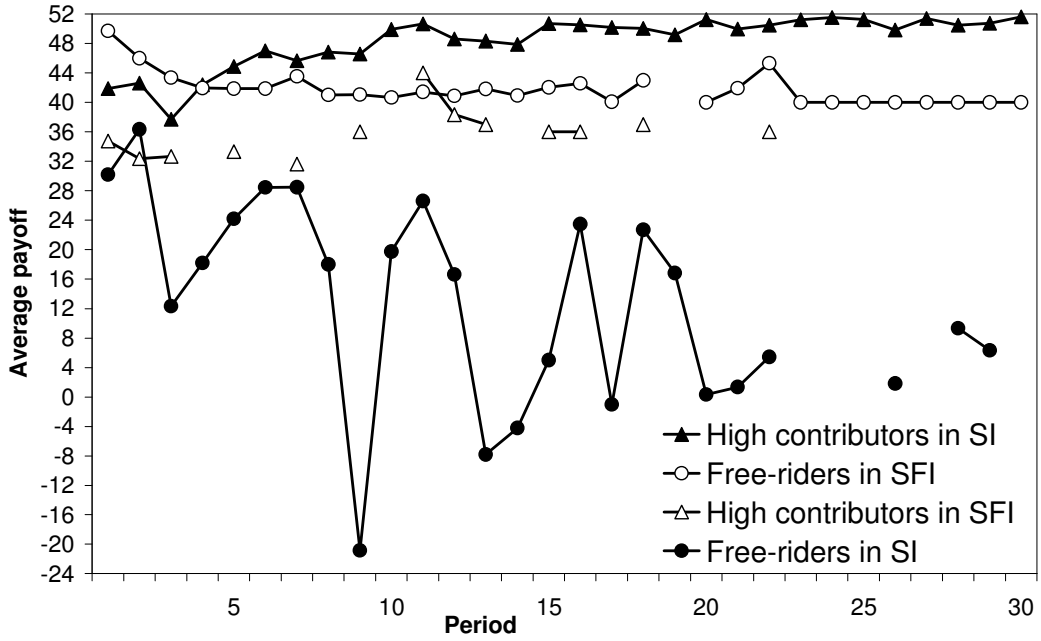

Fig. S2. Subjects' average payoffs dependent on their contribution behavior.

Supporting Tables

Table S1. Logit analysis of the punishment probability dependent on the experience in SI (using individual dummies for subjects, 778 observations in total)

	Coefficient		*Z* value
Number of periods in SI when punishing	-.074	(.026)	- 2.82***
Constant	.798	(.278)	2.87***

*** denotes significance at 1%. The values in parentheses denote the standard errors.

Supporting References and Notes

S1. U. Fischbacher, Working Paper No. 21, *Institute for Empirical Research in Economics*, University of Zurich (1999).

S2. R. M. Isaac, J. Walker, S. Thomas, *Public Choice* 43, 113 (1984).