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.

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.

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.

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

Published by AAAS