



Indirect reciprocity with private, noisy, and incomplete information

Christian Hilbe^{a,1}, Laura Schmid^a, Josef Tkadlec^a, Krishnendu Chatterjee^a, and Martin A. Nowak^{b,c,d}

^aInstitute of Science and Technology Austria, 3400 Klosterneuburg, Austria; ^bProgram for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138; ^cDepartment of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138; and ^dDepartment of Mathematics, Harvard University, Cambridge, MA 02138

Edited by Brian Skyrms, University of California, Irvine, CA, and approved October 16, 2018 (received for review June 19, 2018)

Indirect reciprocity is a mechanism for cooperation based on shared moral systems and individual reputations. It assumes that members of a community routinely observe and assess each other and that they use this information to decide who is good or bad, and who deserves cooperation. When information is transmitted publicly, such that all community members agree on each other's reputation, previous research has highlighted eight crucial moral systems. These "leading-eight" strategies can maintain cooperation and resist invasion by defectors. However, in real populations individuals often hold their own private views of others. Once two individuals disagree about their opinion of some third party, they may also see its subsequent actions in a different light. Their opinions may further diverge over time. Herein, we explore indirect reciprocity when information transmission is private and noisy. We find that in the presence of perception errors, most leading-eight strategies cease to be stable. Even if a leading-eight strategy evolves, cooperation rates may drop considerably when errors are common. Our research highlights the role of reliable information and synchronized reputations to maintain stable moral systems.

cooperation | indirect reciprocity | social norms | evolutionary game theory

Humans treat their reputations as a form of social capital (1–3). They strategically invest into their good reputation when their benevolent actions are widely observed (4–6), which in turn makes them more likely to receive benefits in subsequent interactions (7–12). Reputations undergo constant changes in time. They are affected by rumors and gossip (13), which themselves can spread in a population and develop a life of their own. Evolutionary game theory explores how good reputations are acquired and how they affect subsequent behaviors, using the framework of indirect reciprocity (14–17). This framework assumes that members of a population routinely observe and assess each other's social interactions. Whether a given action is perceived as good depends on the action itself, the context, and the social norm used by the population. Behaviors that yield a good reputation in one society may be condemned in others. A crucial question thus becomes: Which social norms are most conducive to maintain cooperation in a population?

Different social norms can be ordered according to their complexity (18) and according to the information that is required to assess a given action (19, 20). According to "first-order norms," the interpretation of an action depends only on the action itself. When a donor interacts with a recipient in a social dilemma, the donor's reputation improves if she cooperates, whereas her reputation drops if she defects (21–26). According to "second-order norms," the interpretation of an action additionally depends on the reputation of the recipient. The recipient's reputation provides the context of the interaction. It allows observers to distinguish between justified and unjustified defections (27–29). For example, the standing strategy considers it wrongful only to defect against well-reputed recipients; donors who defect against bad recipients do not suffer from

an impaired reputation (30). According to "third-order norms," observers need to additionally take the donor's reputation into account. In this way, assessment rules of higher order are increasingly able to give a more nuanced interpretation of a donor's action, but they also require observers to store and process more information.

When subjects are restricted to binary norms, such that reputations are either "good" or "bad," an exhaustive search shows there are eight third-order norms that maintain cooperation (20, 31). These "leading-eight strategies" are summarized in Table 1, and we refer to them as L1–L8. None of them is exclusively based on first-order information, whereas two of them (called "simple standing" and "stern judging," refs. 32 and 33) require second-order information only. There are several universal characteristics that all leading-eight strategies share. For example, against a recipient with a good reputation, a donor who cooperates should always obtain a good reputation, whereas a donor who defects should gain a bad reputation. The norms differ, however, in how they assess actions toward bad recipients. Whereas some norms allow good donors to preserve their good standing when they cooperate with a bad recipient, other norms disincentivize such behaviors.

Ohtsuki and Iwasa (20, 31) have shown that if all members of a population adopt a leading-eight strategy, stable cooperation can emerge. Their model, however, assumes that the players' images are synchronized; two population members would always agree on the current reputation of some third population member. The assumption of publicly available and synchronized information

Significance

Indirect reciprocity explores how humans act when their reputation is at stake, and which social norms they use to assess the actions of others. A crucial question in indirect reciprocity is which social norms can maintain stable cooperation in a society. Past research has highlighted eight such norms, called "leading-eight" strategies. This past research, however, is based on the assumption that all relevant information about other population members is publicly available and that everyone agrees on who is good or bad. Instead, here we explore the reputation dynamics when information is private and noisy. We show that under these conditions, most leading-eight strategies fail to evolve. Those leading-eight strategies that do evolve are unable to sustain full cooperation.

Author contributions: C.H., L.S., J.T., K.C., and M.A.N. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

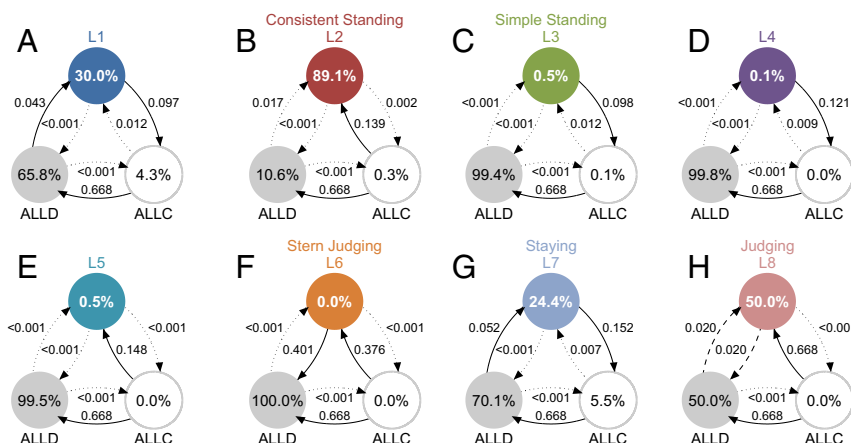
This article is a PNAS Direct Submission.

Published under the PNAS license.

¹To whom correspondence should be addressed. Email: christian.hilbe@ist.ac.at.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810565115/-DCSupplemental.

Published online November 14, 2018.



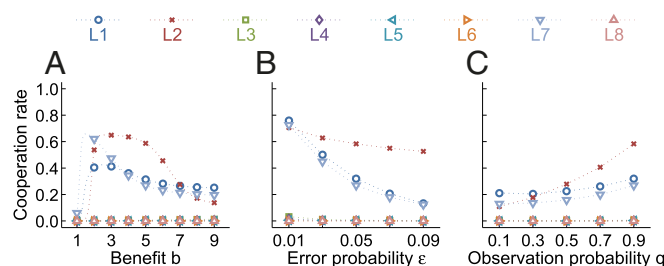
There are only three scenarios in Fig. 3 that allow for positive cooperation rates. The corresponding leading-eight strategies are L1, L2 (“consistent standing”), and L7 (“staying,” ref. 45). For L1 and L7, the evolutionary dynamics take the form of a rock–scissors–paper cycle (46–50). The leading-eight strategy can be invaded by ALLC, which gives rise to ALLD, which in turn leads back to the leading-eight strategy. Because ALLD is most robust in this cycle, the leading-eight strategies are played in less than one-third of the time (Fig. 3 *A* and *G*).

Only consistent standing, L2, is able to compete with ALLC and ALLD in a direct comparison (Fig. 3B). Under consistent standing, there is a unique action in each possible situation that allows a donor to obtain a good standing. For example, when a good donor meets a bad recipient, the donor keeps her good standing by defecting, but loses it by cooperating. Compared with stern judging, which has a similar property (18), consistent standing incentivizes cooperation more strongly. When two bad players interact, the correct decision according to consistent standing is to cooperate, whereas a stern player would defect (Table 1).

Nevertheless, we find that even when consistent standing is common, the average cooperation rate in the population rarely exceeds 65%. To show this, we repeated the previous evolutionary simulations for the eight scenarios while varying the benefit-to-cost ratio, the error rate, and the observation probability (Fig. 4). These simulations confirm that five of the leading-eight strategies cannot maintain any cooperation when competing with ALLC and ALLD. Only for L1, L2, and L7 are average cooperation rates positive, reaching a maximum for intermediate benefit-to-cost ratios (Fig. 4A). If the benefit-to-cost ratio is too low, we find that each of these leading-eight strategies can be invaded by ALLD, whereas if the ratio is too high, ALLC can invade (SI Appendix, Fig. S5). In between, consistent standing may outperform ALLC and ALLD, but in the presence of noise it does not yield high cooperation rates against itself. Even if all interactions are observed ($q = 1$), cooperation rates in a homogeneous L2 population drop below 70% once the error rate exceeds 5% (SI Appendix, Fig. S4). Our analytical results in SI Appendix suggest that while L2 populations always recover from single disagreements, it may take them a substantial time to do so, during

which further errors may accumulate. As a result, whereas L2 seems most robust when coevolving with ALLC and ALLD, it is unable to maintain full cooperation. Furthermore, additional simulation results suggest that even if L2 is able to resist invasion by ALLC and ALLD, it may be invaded by mutant strategies that differ in only one bit from L2 (*SI Appendix, Fig. S6*).

So far, we have assumed that mutations are rare, such that populations are typically homogeneous. Experimental evidence, however, suggests that there is considerable variation in the social norms used by subjects (4, 7–11). While some subjects are best classified as unconditional defectors, others act as unconditional cooperators or use more sophisticated higher-order strategies (11). In agreement with these experimental studies, there is theoretical evidence that some leading-eight strategies like L7 may form stable coexistences with ALLC (36). In *SI Appendix, Figs. S7–S9*, we present further evolutionary results for higher mutation rates, in which such coexistences are possible.



6. Ohtsuki H, Iwasa Y, Nowak MA (2015) Reputation effects in public and private interactions. *PLoS Comput Biol* 11:e1004527.
7. Wedekind C, Milinski M (2000) Cooperation through image scoring in humans. *Science* 288:850–852.
8. Seinen I, Schram A (2006) Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur Econ Rev* 50:581–602.
9. Bolton G, Katok E, Ockenfels A (2005) Cooperation among strangers with limited information about reputation. *J Public Econ* 89:1457–1468.
10. van Apeldoorn J, Schram A (2016) Indirect reciprocity; a field experiment. *PLoS One* 11:e0152076.
11. Swakman V, Molleman L, Ule A, Egas M (2016) Reputation-based cooperation: Empirical evidence for behavioral strategies. *Evol Hum Behav* 37:230–235.
12. Capraro V, Giardini F, Vilone D, Paolucci M (2016) Partner selection supported by opaque reputation promotes cooperative behavior. *Judgment Decis Making* 11:589–600.
13. Sommerfeld RD, Krambeck HJ, Semmann D, Milinski M (2007) Gossip as an alternative for direct observation in games of indirect reciprocity. *Proc Natl Acad Sci USA* 104:17435–17440.
14. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437:1291–1298.
15. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314:1560–1563.
16. Sigmund K (2010) *The Calculus of Selfishness* (Princeton Univ Press, Princeton).
17. Sigmund K (2012) Moral assessment in indirect reciprocity. *J Theor Biol* 299:25–30.
18. Santos FP, Santos FC, Pacheco JM (2018) Social norm complexity and past reputations in the evolution of cooperation. *Nature* 555:242–245.
19. Brandt H, Sigmund K (2004) The logic of reprobation: Assessment and action rules for indirect reciprocation. *J Theor Biol* 231:475–486.
20. Ohtsuki H, Iwasa Y (2004) How should we define goodness?—Reputation dynamics in indirect reciprocity. *J Theor Biol* 231:107–120.
21. Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393:573–577.
22. Nowak MA, Sigmund K (1998) The dynamics of indirect reciprocity. *J Theor Biol* 194:561–574.
23. Ohtsuki H (2004) Reactive strategies in indirect reciprocity. *J Theor Biol* 227:299–314.
24. Brandt H, Sigmund K (2005) Indirect reciprocity, image scoring, and moral hazard. *Proc Natl Acad Sci USA* 102:2666–2670.
25. Berger U (2011) Learning to cooperate via indirect reciprocity. *Games Econ Behav* 72:30–37.
26. Berger U, Grüne A (2016) On the stability of cooperation under indirect reciprocity with first-order information. *Games Econ Behav* 98:19–33.
27. Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. *Proc R Soc B* 268:745–753.
28. Panchanathan K, Boyd R (2003) A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *J Theor Biol* 224:115–126.
29. Suzuki S, Akiyama E (2007) Evolution of indirect reciprocity in groups of various sizes and comparison with direct reciprocity. *J Theor Biol* 245:539–552.
30. Sugden R (1986) *The Economics of Rights, Co-Operation and Welfare* (Blackwell, Oxford).
31. Ohtsuki H, Iwasa Y (2006) The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol* 239:435–444.
32. Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. *J Theor Biol* 241:233–240.
33. Santos FP, Santos FC, Pacheco JM (2016) Social norms of cooperation in small-scale societies. *PLoS Comput Biol* 12:e1004709.
34. Uchida S (2010) Effect of private information on indirect reciprocity. *Phys Rev E* 82:036111.
35. Uchida S, Sasaki T (2013) Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos Solitons Fractals* 56:175–180.
36. Okada I, Sasaki T, Nakai Y (2017) Tolerant indirect reciprocity can boost social welfare through solidarity with unconditional cooperators in private monitoring. *Sci Rep* 7:9737.
37. Okada I, Sasaki T, Nakai Y (2018) A solution for private assessment in indirect reciprocity using solitary observation. *J Theor Biol* 455:7–15.
38. Martínez-Vaquero LA, Cuesta JA (2013) Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation. *Phys Rev E* 87:052810.
39. Traulsen A, Nowak MA, Pacheco JM (2006) Stochastic dynamics of invasion and fixation. *Phys Rev E* 74:011909.
40. Stewart AJ, Plotkin JB (2013) From extortion to generosity, evolution in the iterated prisoner's dilemma. *Proc Natl Acad Sci USA* 110:15348–15353.
41. Reiter JG, Hilbe C, Rand DG, Chatterjee K, Nowak MA (2018) Crosstalk in concurrent repeated games impedes direct reciprocity and requires stronger levels of forgiveness. *Nat Commun* 9:555.
42. Hilbe C, Šimsa Š, Chatterjee K, Nowak MA (2018) Evolution of cooperation in stochastic games. *Nature* 559:246–249.
43. Fudenberg D, Imhof LA (2006) Imitation processes with small mutations. *J Econ Theor* 131:251–262.
44. Wu B, Gokhale CS, Wang L, Traulsen A (2012) How small are small mutation rates? *J Math Biol* 64:803–827.
45. Sasaki T, Okada I, Nakai Y (2017) The evolution of conditional moral assessment in indirect reciprocity. *Sci Rep* 7:41870.
46. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, UK).
47. Szolnoki A, Szabó G (2004) Phase transitions for rock-scissors-paper game on different networks. *Phys Rev E* 70:037102.
48. Claussen JC, Traulsen A (2008) Cyclic dominance and biodiversity in well-mixed populations. *Phys Rev Lett* 100:058104.
49. Szolnoki A, et al. (2014) Cyclic dominance in evolutionary games: A review. *J R Soc Interface* 11:20140735.
50. Stewart AJ, Parsons TL, Plotkin JB (2016) Evolutionary consequences of behavioral diversity. *Proc Natl Acad Sci USA* 113:E7003–E7009.
51. Resnick P, Zeckhauser R, Swanson J, Lockwood K (2006) The value of reputation on eBay: A controlled experiment. *Exp Econ* 9:79–101.
52. Restivo M, van de Rijt A (2012) Experimental study of informal rewards in peer production. *PLoS One* 7:e34358.
53. Lieberman E, Hauert C, Nowak MA (2005) Evolutionary dynamics on graphs. *Nature* 433:312–316.
54. Szabó G, Fáth G (2007) Evolutionary games on graphs. *Phys Rep* 446:97–216.
55. Szabó G, Tóke C (1998) Evolutionary prisoner's dilemma game on a square lattice. *Phys Rev E* 58:69–73.