doi:10.1111/j.1558-5646.2010.01194.x

POWER AND CORRUPTION

Francisco Úbeda[1,2,3] and Edgar A. Duéñez-Guzmán[4]

[1]*Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville Tennessee 37996*

[2]*Madrid Institute of Advanced Studies: Social Sciences (IMDEA: Ciencias Sociales) , Madrid, Spain*

[3]*E-mail: fubeda@utk.edu*

[4]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge Massachusetts 02138*

Received July 13, 2010
Accepted October 25, 2010

Cooperation is ubiquitous in the natural world. What seems nonsensical is why natural selection favors a behavior whereby individuals would lose out by benefiting their competitor. This conundrum, for almost half a century, has puzzled scientists and remains a fundamental problem in biology, psychology, and economics. In recent years, the explanation that punishment can maintain cooperation has received much attention. Individuals who punish noncooperators thrive when punishment does not entail a cost to the punisher. However when punishment is costly, cooperation cannot be preserved. Most literature on punishment fails to consider that punishers may act corruptly by not cooperating when punishing noncooperators. No research has considered that there might be power asymmetries between punishers and nonpunishers that turn one of these type of individuals more or less susceptible to experiencing punishment. Here, we formulate a general game allowing corruption and power asymmetries between punishers and nonpunishers. We show that cooperation can persist if punishers possess power and use it to act corruptly. This result provides a new interpretation of recent data on corrupt policing in social insects and the psychology of power and hypocrisy in humans. These results suggest that corruption may play an important role in maintaining cooperation in insects and human societies. In contrast with previous research, we contend that costly punishment can be beneficial for social groups. This work allows us to identify ways in which corruption can be used to the advantage of a society.

KEY WORDS: Altruism, cooperation, corruption, economic policy, evolutionary games, game theory, political philosophy, Prisoner's dilemma, spite.

Making sense of cooperation remains one of the fundamental problems for evolutionary biologists (Hamilton 1963; Hamilton 1970; Axelrod and Hamilton 1981; West et al. 2002; Doebeli and Hauert 2005; Jansen and van Baalen 2006; West et al. 2007b; Leimar and Hammerstein 2010; Brosnan and Bshary 2010). Natural selection favors those individuals who achieve greater than average fitness. By definition, cooperation improves the fitness of an individual other than the actor, that is the recipient (Hamilton 1963; Hamilton 1970). Hence, all else being equal, cooperation reduces the relative fitness of cooperators and is disfavored by natural selection. Cooperation is justified when it occurs between related individuals (kin selection) (Hamilton 1963) or when mechanisms exist to enforce this behavior (Lehmann and Keller 2006; West et al. 2007b). One mechanism that can sustain cooperation among unrelated individuals is "punishment" (inflicting a cost on defectors) (Axelrod 1986) specifically when it does not entail a cost to the punisher. Yet, when punishment imposes a cost (costly punishment), cooperation cannot be preserved because punishing cooperators are outcompeted by nonpunishing cooperators (Axelrod 1986) who gain the advantage of cooperating without incurring the cost of punishing. Costly punishment, as a solution, has attracted much attention in recent years both from an experimental (Fehr and Gachter 2002; Wenseleers and Ratnieks 2006; Wu et al. 2009; Janssen et al. 2010) and theoretical perspectives (Frank 2003; Gardner and West 2004; Lehmann et al. 2007; Dreber et al. 2008).

Most literature on costly punishment (except Sigmund et al. 2001; Gardner and West 2004; Nakamaru and Iwasa 2006; Eldakar et al. 2007; Lehmann et al. 2007; Eldakar and Wilson 2008) assumes that punishers act "honestly" by cooperating when

punishing defectors. Here, we explore the possibility that punishers act "corruptly" by defecting when punishing defectors. We, moreover, consider that there might be "power" asymmetries when the cost of punishing a nonpunisher is not the same as that of punishing a punisher; and/or the punishment inflicted on a nonpunisher (henceforth civilian) is not the same as that rendered to a punisher (henceforth policer). Corruption and power asymmetries are well documented in social insects (Wenseleers et al. 2005; Stroeymeyt et al. 2007) and pervasive in human societies (Shleifer and Vishny 1983; Lammers et al. 2010).

Corrupt policers are not uncommon in social wasps and ants (Ratnieks et al. 2006). Within communities of the wasp *Dolichovespula sylvestris,* the queen lays eggs but workers normally do not, although they can (Ratnieks et al. 2006). To ensure reproduction is exclusive to the queen, workers police other workers, punishing those who lay eggs either by attacking the workers or removing their eggs (Wenseleers et al. 2005). Policing workers, however, are known to act corruptly by laying eggs (Wenseleers et al. 2005). Workers who assume the policing role often possess greater power (Monnin and Ratnieks 2001; Saigo and Tsuchida 2004; Stroeymeyt et al. 2007). The sources of power vary between communities, ranging from policers with better fighting skills to policers better able to disguise their defection (Monnin and Ratnieks 2001; Saigo and Tsuchida 2004; Stroeymeyt et al. 2007).

Punishment and corruption are pervasive in human societies. Humans show a tendency to punish individuals who do not cooperate; this is true even when social partners interact only once and no room for reciprocity exists (Heinrich et al. 2006; Sigmund 2007). Increments in punishment increase cooperative behaviors across human societies (Heinrich et al. 2006). Cases of corruption in humans fill the newspapers columns everyday. In these times of economic crisis, politicians have been preaching social sacrifices while their own expense accounts have fattened up. This anecdotal evidence has been established scientifically by recent empirical work showing that the greater the power, the greater the tendency to condemn the transgression of others more than one's own (Lammers et al. 2010).

Methods

In the Prisoner's Dilemma, after two individuals choose between cooperating or defecting, they receive a payoff based on their combined action (Axelrod 1984). When both individuals cooperate, they secure a higher payoff (r) than when both defect (0). Although both players are better off cooperating ($r > 0$), defecting against a cooperator yields the highest payoff (t) while cooperating against a defector yields the lowest payoff ($-s$). Individuals are selected to defect against cooperators ($t > r$) and defection takes over the population (Fig. 1) (Axelrod 1984). One of the earliest extensions of the Prisoner's Dilemma was allowing cooperators to punish defectors (Punishment Game).

Corruption Game

		C	D	H	K
Cooperating Civilians	C	r	$-s$	r	$-s$
Defecting Civilians	D	t	0	$t-p$	$-p$
Honest Policers	H	r	$-s-c$	r	$-s-d$
Corrupt Policers	K	t	$-c$	$t-q$	$-q-d$

Figure 1. Pay-off matrix of corruption game. Values r, s, and t correspond to the payoffs of the Prisoner's Dilemma. These payoffs satisfy the inequalities $t > r > t - s$ and $t - s > 0$. Values p and c are the punishment endured by defecting civilians and the cost of punishing civilians, with $p, c > 0$. We color in gray the payoff matrix of the Prisoner's Dilemma and in yellow the payoff matrix of the Punishment Game. Values q and d are the punishment endured by defecting policers and the cost of punishing policers, with $q, d > 0$.

Here, we extend the Punishment Game to allow punishers to either cooperate or defect. We refer to this as the "Corruption Game." We contemplate four types of individuals to allow for the evolution of corrupt policing: cooperating civilians (C), defecting civilians (D), honest policers (H), and corrupt policers (K). To allow for power asymmetries, we consider that the punishment inflicted on a defecting civilian (p) may differ from the punishment inflicted on a defecting policer (q); and the cost of punishing a defecting civilian (c) may differ from the cost of punishing a defecting policer (d) (Fig. 1). Because punishment involves costs both to the actor (the cost of punishing) and the recipient (the cost of being punished), it is a spiteful behavior (Hamilton 1970; West et al. 2007a). We can aggregate these costs and define the "social cost of punishing a civilian" ($p + c$), and the "social cost of punishing a policer" ($q + d$).

Defecting punishers have been considered in the context of public good games (Sigmund et al. 2001; Eldakar and Wilson 2008) and spatial games (Nakamaru and Iwasa 2006), but these punishers have been ignored in the Prisoner's Dilemma, a game that captures the essence of the cooperation problem. Power asymmetries are considered for the first time in models of cooperation and punishment.

We analyze our Corruption Game within the context of continuous time replicator dynamics, which is customary in evolutionary game theory (Hofbauer and Sigmund 1998). We derive analytically the equilibria of the Corruption Game and provide their stability conditions (see Appendix for details).

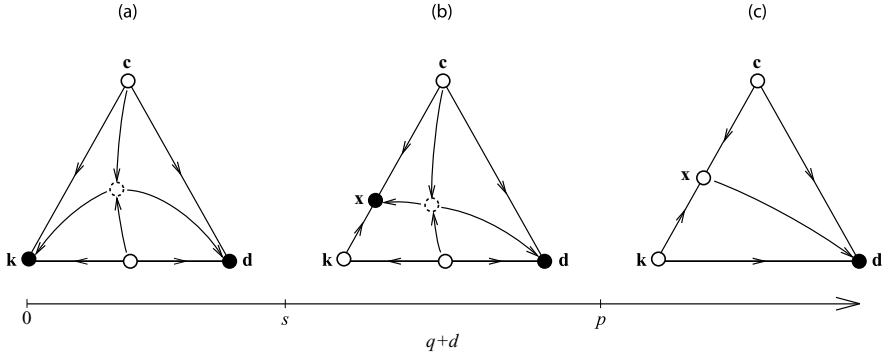

Figure 2. **Dynamics of the corruption game. Each triangle corresponds to the face of the simplex formed by equilibria {c, d, k}. For simplicity only one face is represented but the analysis corresponds to the complete case in which all four strategies are present. Full circles correspond to stable equilibria and empty circles correspond to unstable equilibria. Dashed circles indicate equilibria that may or may not be present depending on the parameter values considered (see Appendix for details). Arrows point in the direction of the dynamics. There are three possible cases depending on the value taken by the social cost of punishing a policer: (a) When $0 < q + d < s$ equilibrium k is stable and equilibrium x does not exist; (b) When $s < q + d < p$ equilibrium k is unstable and equilibrium x exists and is stable; (c) When $p < q + d$ equilibrium k is unstable and equilibrium x exists but is unstable.**

Results

A population of defecting civilians (**d**) is always stable whereas a population of cooperating civilians (**c**) is always unstable; this result corresponds to the Prisoner's Dilemma (Axelrod 1984). There is a neutrally stable set of equilibria involving honest policers and cooperating civilians. The Corruption Game allows two additional stable equilibria: a population of corrupt policers (**k**), and a mixed population of cooperating civilians and corrupt policers (**x**) (see Appendix).

Equilibrium **k** exists and is stable when $q + d$ is in the range $(0, s)$. More interestingly, equilibrium **x** exists and is stable when $q + d$ is in the range (s, p). When $q + d$ is greater than p, neither equilibrium **k** nor equilibrium **x** are stable (Fig. 2) (see Appendix). The existence and stability of equilibrium **x** requires that the punishment of policers is lower than the punishment of civilians ($q < p$). Equal punishment of policers and civilians ($q = p$) precludes their co-existence. It is important to notice that at equilibria **k** and **x** honest policers cannot prosper.

Thus a necessary condition for the evolution of a mixed population of corrupt policers and cooperating civilians is that there are power asymmetries, particularly policers have greater power than civilians.

The frequency of cooperating civilians x_C and corrupt policers x_K at equilibrium $\mathbf{x} = (x_C, x_K)$, is determined by the value of $q + d$ relative to s. The greater $q + d$ is relative to s, the greater x_C (Fig. 3) (see Appendix). The greater x_C is, the greater the population mean payoff $\bar{W}_{\mathbf{x}}$. The value of $\bar{W}_{\mathbf{x}}$ grows rapidly with $q + d$. Values of $q + d$ slightly greater than s bring $\bar{W}_{\mathbf{x}}$ above the population mean payoff for a population of defecting civilians (0). As the value of $q + d$ grows, the value of $\bar{W}_{\mathbf{x}}$ comes closer to the population mean payoff for a population of cooperating civilians (r) (Fig. 4).

When $q + d$ is in the range (s, p) equilibria **x** and **d** are present in the population simultaneously. The likelihood of attaining equilibrium **x** (as opposed to **d**) is given by the size of its basin of attraction $B_{\mathbf{x}}$: the greater $B_{\mathbf{x}}$ is, the greater the likelihood that equilibrium **x** evolves. The size of $B_{\mathbf{x}}$ is determined by the value of $q + d$ relative to p, and by the value of c. The smaller $q + d$ (or the greater p) is, the greater $B_{\mathbf{x}}$. Also the smaller c is, the greater $B_{\mathbf{x}}$ (Fig. 3) (see Appendix). Analytical results concerning $B_{\mathbf{x}}$ are only an approximation. We test the validity of these results through a simulation that introduces random mutants in a population of cooperating civilians and evaluate which equilibrium evolves. Analytic and simulation results are qualitatively equivalent although simulations indicate that the probability that equilibrium **x** evolves is slightly greater than indicated by analytic results (Fig. 3) (see Appendix).

Therefore the social cost of punishing a policer has opposite effects on two desirable properties of a mixed population of corrupt policers and cooperating civilians. The greater the social cost of punishing a policer: (1) the greater the population mean payoff (because it is necessary for less-corrupt policers to sustain a larger population of cooperating civilians), (2) the smaller the likelihood that cooperating civilians and corrupt policers co-exist.

Discussion

We show that power corrupts. In our game, a necessary condition for the evolution of corrupt policing is that there are power asymmetries, particularly policers receive less punishment than civilians. Small power asymmetries (in particular $s < q + d < p$) promote limited corruption, leading to a population where cooperating civilians and corrupt policers co-exist. Large power asymmetries (in particular $0 < q + d < s$) promote absolute

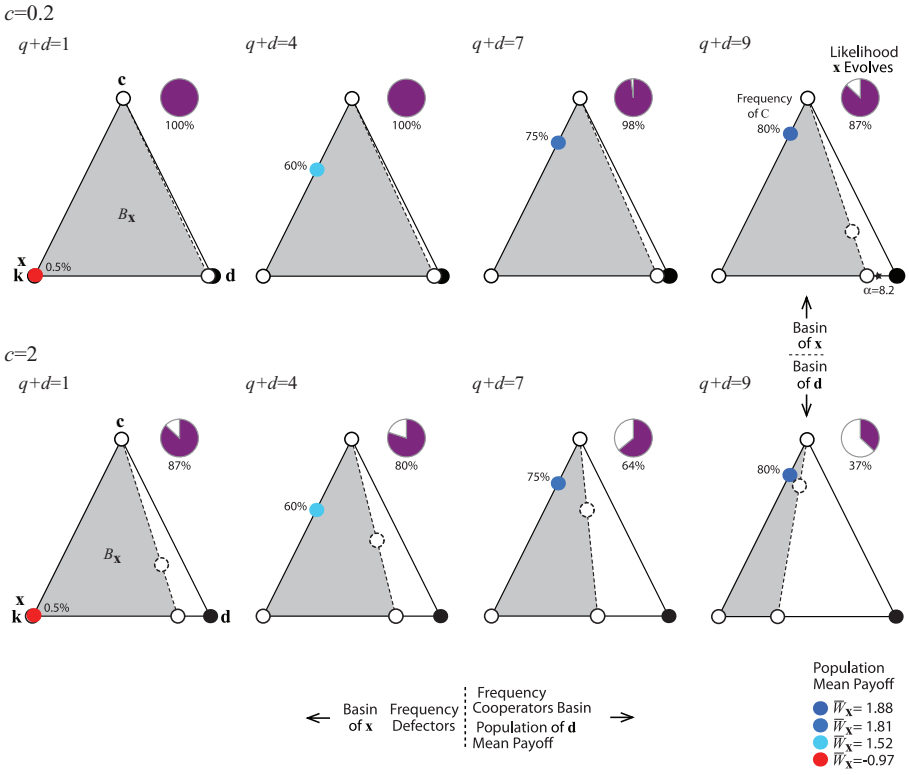

Figure 3. **Frequency of cooperators, population mean payoff and basin of attraction. This figure presents the actual value taken by equilibrium x for the set of payoffs $r = 2$, $s = 1$, and $t = 4$ (corresponding to the a standard parametrization of the Prisoner's Dilemma), the value $p = 10$, and different values of c and $q + d$. In particular, the top row corresponds to the value $c = 0.2$ and the bottom row to the value $c = 2$. Each of the four columns correspond to the $q + d$ values 1, 4, 7, and 9, respectively. We only consider the case in which $s < q + d < p$ that allows the existence of x. We use a color code (see legend) to indicate the value of the population mean payoff corresponding to equilibrium x. We show in gray the approximate size of the basin of attraction of equilibrium x. The likelihood that equilibrium x evolves when the starting point is a population of cooperating civilians only (c) is given by a pie chart to the right of each triangle. We use purple to indicate the likelihood that equilibrium x evolves and white for the likelihood that equilibrium d evolves. Arrows indicate the direction in which frequency of cooperators, population mean fitness, and basin of attraction increase as we move along the grid $c \times (q + d)$.**

corruption, leading to a population of corrupt policers only. These results are best summarized by the words of 19th century historian Lord Acton who claimed "Power corrupts, absolute power corrupts absolutely."

We show that corruption can maintain cooperation. When power asymmetries are small, cooperation can persist. Intuitively, corrupt policers have no incentive to stop punishing because the punishment of defecting civilians outweighs the punishment of defecting policers coupled with the cost of punishing other policers. Cooperating civilians have no incentive to defect because the punishment inflicted by policers outweighs the gain derived from defecting. Notice that our model solves the problem of incentives linked to models of costly punishment (Axelrod 1986) without requiring multiple interactions between social partners, unlike most models of costly punishment (Hagen and Hammerstein 2006; Sigmund 2007).

These results provide insights into recent findings on the evolution of cooperation in social insects and the link between power and hypocrisy in humans. Entomologists found evidence indicating that mechanisms of enforcement might be as important as relatedness when it comes to maintaining cooperation in some social insects (Wenseleers and Ratnieks 2006). Our game provides a framework in which cooperation can evolve and fits well with the evidence available, namely corrupt policing and power asymmetries are well documented in social insects (Monnin and Ratnieks 2001; Saigo and Tsuchida 2004; Wenseleers et al. 2005; Stroeymeyt et al. 2007). Notice, however, that punishment in social insects does not need to be costly.

Psychologists have proved that power promotes the morally corrupt tendency of condemning the defection of other people more than condemning the same behavior in oneself (Lammers et al. 2010). Similarly in our game, power promotes corruption among policers who punish defectors even though they also defect. Therefore, corruption may be an important force to sustain cooperation in human societies. This could be further explored by extending our model of corruption to consider repeated

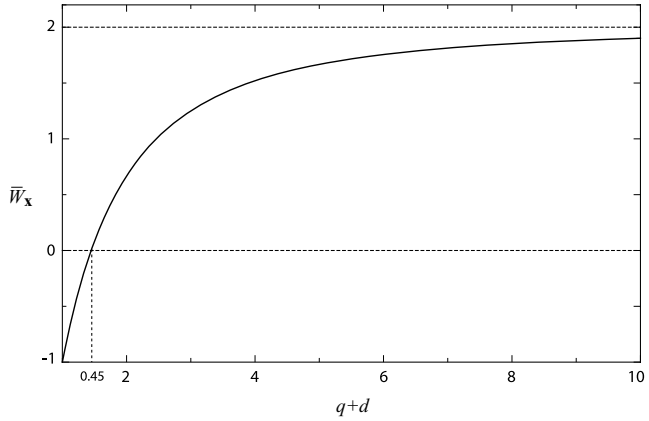

Figure 4. Population mean payoff. This figure presents the population mean payoff $\bar{W}$ corresponding to equilibrium x as a function of $q + d$; $\bar{W}_x = rx_c^2 + 2(t - s)x_c x_k - (q + d)x_k^2$. Its value ranges between the $\bar{W}$ corresponding to equilibrium k, that is $-(q + d) = -1$, and the $\bar{W}$ corresponding to equilibrium c, that is $r = 2$. The $\bar{W}$ corresponding to equilibrium x is greater than the population mean payoff $\bar{W}$ corresponding to equilibrium d, that is 0, when $q + d > 0.45$.

interactions between individuals, as opposed to single interactions that we consider in this work.

We show that corruption often increases the population mean payoff. Power and corruption make policing a viable strategy and policing, by favoring cooperation, is good for the population: even when policers are corrupt! Recently, it has been argued that costly punishment is inefficient because even when it increases the amount of cooperation in a social group, it does not increase the mean payoff of its members (Dreber et al. 2008; Ohtsuki et al. 2009). In contrast, here, costly punishment can increase the amount of cooperation while increasing the population mean payoff. Thus our work demonstrates that costly punishment can benefit all members of a social group and improve the efficiency of the outcome.

Our model has implications for economic policy. It provides directions in which a central planner could modify some parameter values to use corruption to the advantage of a society. From our model, we learn that increasing the punishment inflicted on defectors (civilians and policers) is beneficial by increasing the population mean payoff at equilibrium. Surprisingly, we also learn that the punishment inflicted on policers should always be lower than the punishment inflicted on civilians, or that increasing the cost of punishing a policer is also beneficial. In our model, notice that corruption does not bring about social inequality or exploitation; cooperating civilians and corrupt policers receive the same payoff at equilibrium.

Our results contribute to philosophical debate. Philosophical Anarchism promotes egoist individuals who do not believe in any form of authority (Stirner 1845). Individuals bounded by authority cooperate whereas self-interested individuals (who ignore authority) defect and take advantage of the individuals bounded by authority. Such system would be unstable because eventually all individuals will end up disregarding authority (Miller 1984). Our model indicates that the society championed by Stirner (1845) would persist if the individuals who disregard all authority encourage fellow citizens to respect authority (keeping their arguments to themselves). This is in line with the arguments of Miller 1984).

Our results further sharpens the expectations from a theory of cooperation through enforcement, which has the potential to be applied on many biological levels. Our model predicts behavioral polymorphism with some individuals willing to engage in costly punishment (but others not) and with some individuals willing to cooperate (but others not). Individuals not willing to cooperate, however, are expected to be in positions of power, advocate cooperation, and even punish noncooperators.

ACKNOWLEDGMENTS

E. Akcay, A. Gardner, M. M. Patten, R. Serrano, D. N. Tran, S. Sadedin, and K. Voigt provided comments on an early draft the manuscript.

LITERATURE CITED

Axelrod, R. 1984. The evolution of cooperation. Basic Books, New York.

———. 1986. An evolutionary approach to norms. Am. Polit. Sci. Rev. 80:1095–1111.

Axelrod, R., and W. D. Hamilton. 1981. The evolution of cooperation. Science 211:1390–1396.

Brosnan, S. F., and R. Bshary. 2010. Cooperation and deception: from evolution to mechanisms. Philos. Trans. R. Soc. Lond. B 365:2593–2598.

Doebeli, M., and C. Hauert. 2005. Models of cooperation based on the prisoner's dilemma and the snowdrift game. Ecol. Lett. 8:748–766.

Dreber, A., D. G. Rand, D. Fudenberg, and M. A. Nowak. 2008. Winners don't punish. Nature 452:348–351.

Eldakar, O. T., D. L. Farrell, and D. S. Wilson. 2007. Selfish punishment: altruism can be maintained by competition among cheaters. J. Theor. Biol. 249:198–205. URL < Go to ISI >://000251520500003.

Eldakar, O. T., and D. S. Wilson. 2008. Selfishness as second-order altruism. Proc. Natl. Acad. Sci. USA 105:6982–6986.

Fehr, E., and S. Gachter. 2002. Altruistic punishment in humans. Nature 415:137–140.

Frank, S. A. 2003. Repression of competition and the evolution of cooperation. Evolution 57:693–705.

Gardner, A., and S. West. 2004. Cooperation and punishment, especially in humans. Am. Nat. 164:753–764.

Hagen, E., and P. Hammerstein. 2006. Game theory and human evolution: a critique of some recent interpretation of experimental games. Theor. Popul. Biol. 69:339–348.

Hamilton, W. D. 1963. The evolution of altruistic behavior. Am. Nat. 97:354–356.

———. 1970. Selfish and spiteful behaviour in an evolutionary model. Nature 228:1218.

Hardin, G. 1960. The competitive exclusion principle. Science 131:1292–1297.

Heinrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Heinrich, et al. 2006. Costly punishment across human societies. Science 312:1767–1770.

Hofbauer, J., and K. Sigmund. 1998. Evolutionary games and population dynamics. Cambridge Univ. Press, Cambridge.
Jansen, V. A. A., and M. van Baalen. 2006. Altruism through beard chromodynamics. Nature 440:663–666.
Janssen, M. A., R. Holahan, A. Lee, and E. Ostrom. 2010. Lab experiments for the study of social-ecological systems. Science 328:613–617.
Lammers, J., D. A. Stapel, and A. D. Galinsky. 2010. Power increases hypocrisy. moralizing in reasoning, inmorality in behavior. Psychol. Sci. 21:737–744.
Lehmann, L., and L. Keller. 2006. The evolution of cooperation and altruism—a general framework and a classification of models. J. Evol. Biol. 19:1365–1376.
Lehmann, L., F. Rousset, D. Roze, and L. Keller. 2007. Strong reciprocity or strong ferocity? a population genetic view of the evolution of altruistic punishment. Am. Nat. 170:21–36.
Leimar, O., and P. Hammerstein. 2010. Cooperation for direct fitness benefits. Philos. Trans. R. Soc. Lond. B 365:2619–2626.
Miller, D. 1984. Anarchism. J. M. Dent & Sons Ltd.
Monnin, T., and F. L. W. Ratnieks. 2001. Policing in queenless ponerine ants. Behav. Ecol. Sociobiol. 50:97–108.
Nakamaru, M., and Y. Iwasa. 2006. The coevolution of altruism and punishment: role of the selfish punisher. J. Theor. Biol. 240:475–488.
Ohtsuki, H., Y. Iwasa, and M. A. Nowak. 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457:79–82.
Ratnieks, F. L. W., K. R. Foster, and T. Wenseleers. 2006. Conflict resolution in insect societies. Annu. Rev. Entomol. 51:581–608.
Saigo, T., and K. Tsuchida. 2004. Queen and worker policing in monogynous and monandrous colonies of a primitively eusocial wasp. Proc. R. Soc. Lond. B 271(Suppl. 6):S509–S512.
Shleifer, A., and R. Vishny. 1983. Corruption. Q. J. Econ. 108:599–617.
Sigmund, K. 2007. Punish or perish? retaliation and collaboration among humans. Trends Ecol. Evol. 22:593–600.
Sigmund, K., C. Hauert, and M. A. Nowak. 2001. Reward and punishment. Proc. Natl. Acad. Sci. USA 98:10757–10762.
Stirner, M. 1845. The ego and its own. Cambridge Univ. Press, Cambridge.
Stroeymeyt, N., E. Brunner, and J. Heinze. 2007. "selfish worker policing" controls reproduction in a temnothorax ant. Behav. Ecol. Sociobiol. 61:1449–1457.
Wenseleers, T., and F. L. W. Ratnieks. 2006. Enforced altruism in insect societies. Nature 444:50.
Wenseleers, T., A. Tofilski, and F. L. W. Ratnieks. 2005. Queen and worker policing in the tree wasp *dolichovespula sylvestris*. Behav. Ecol. Sociobiol. 58:80–86.
West, S., A. Griffin, and A. Gardner. 2007a. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. 20:415–432.
West, S. A., A. S. Griffin, and A. Gardner. 2007. Evolutionary explanations for cooperation. Curr. Biol. 17:R661–R672.
West, S. A., I. Pen, and A. S. Griffin. 2002. Conflict and cooperation: cooperation and competition between relatives. Science 296:72–75.
Wu, J., B. Zhang, Z. Zhou, Q. He, X. Zhang, R. Cressman, and Y. Tao. 2009. Costly punishment does not always increase cooperation. Proc. Natl. Acad. Sci. USA 106:17448–17451.

Associate Editor: S. West

Appendix

CORRUPTION GAME

Let

$$\mathbf{A} = \begin{pmatrix} r & -s & r & -s \\ t & 0 & t-p & -p \\ r & -s-c & r & -s-d \\ t & -c & t-q & -q-d \end{pmatrix} \tag{A1}$$

be the payoff matrix of our Corruption Game where each row corresponds to one of the four strategies available, namely: cooperating civilian (C), defecting civilian (D), honest policer (H), and corrupt policer (K). Bold letters represent nonscalar variables with capital and lower case letters corresponding to matrices and vectors, respectively.

Parameters r, s, $t > 0$ correspond to the payoffs of the Prisoner's Dilemma where $t > r$. For simplicity, we assume that $t - r - s > 0$. Parameters p, $q > 0$ correspond to the cost experienced by a defecting civilian (p) and a corrupt policer (q) when punished. Parameters c, $d > 0$ correspond to the cost experienced by a policer when punishing a defecting civilian (c) and a corrupt policer (d).

Following Hofbauer and Sigmund (1998), we define a system of differential equations that describe the continuous time *replicator dynamics* of our Corruption Game:

$$\dot{x_i} = x_i((\mathbf{Ax})_i - \mathbf{x}^T\mathbf{Ax}) \tag{A2}$$

where subscript i corresponds to each of the four strategies available $\{C, D, H, K\}$, a dot corresponds to the time derivative, a T superscript corresponds to the transpose of a vector or matrix. We proceed to characterize the equilibria of this system and to determine the stability of these equilibria.

EQUILIBRIA AND STABILITY

The equilibria may rest in the interior, corners, edges, or faces of the simplex formed by the frequency of each strategy x_i where $\sum_i x_i = 1$.

Interior

Interior equilbria must satisfy the system of equations

$$(\mathbf{Ax})_C = (\mathbf{Ax})_D = (\mathbf{Ax})_H = (\mathbf{Ax})_K. \tag{A3}$$

Therefore, these equilibria are normalized solutions to the system $\mathbf{Ax} = \mathbf{1}_4$, where $\mathbf{1}_n$ is the unit vector of size n.

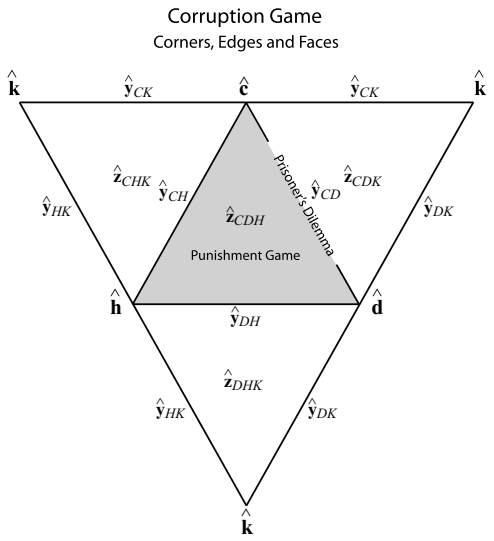

Figure A1. Summary of notation referring to points in the corner, edge, and faces of the simplex formed by equilibria {ĉ, d̂, ĥ, k̂}.

Solving the linear system, we find that its equilibria take the form

$$x = \frac{1}{(t-r-s)(d-c)} \begin{pmatrix} c(r+s-t)-ds \\ (t-r)d \\ (t-r)c \\ -(t-r)c \end{pmatrix} \quad \text{(A4)}$$

which cannot lie in the interior of the simplex. Therefore, all equilibria must lie in the corners, edges, or faces of the simplex, and, by the exclusion principle (Hardin 1960), all interior trajectories converge to the simplex boundary.

Corners

The corners of the simplex are always equilibria (see Equation A2): $\hat{\mathbf{c}} = (1,0,0,0)^T$, $\hat{\mathbf{d}} = (0,1,0,0)^T$, $\hat{\mathbf{h}} = (0,0,1,0)^T$, $\hat{\mathbf{k}} = (0,0,0,1)^T$ (see Fig. A1). Hats denote equilibria. Notice that, for simplicity, we dropped the hats in the main text.

We analyze the local stability of each corner making use of the Jacobian matrices of the system evaluated in each corner

$$\mathbf{J}(\hat{\mathbf{c}}) = \begin{pmatrix} -r & -t & -r & -t \\ 0 & t-r & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t-r \end{pmatrix},$$

$$\mathbf{J}(\hat{\mathbf{d}}) = \begin{pmatrix} -s & 0 & 0 & 0 \\ s & 0 & s+c & c \\ 0 & 0 & -s-c & 0 \\ 0 & 0 & 0 & -c \end{pmatrix},$$

$$\mathbf{J}(\hat{\mathbf{h}}) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & t-r-p & 0 & 0 \\ -r & -t-p & -r & q-t \\ 0 & 0 & 0 & t-r-q \end{pmatrix},$$

$$\mathbf{J}(\hat{\mathbf{k}}) = \begin{pmatrix} q+d-s & 0 & 0 & 0 \\ 0 & q+d-p & 0 & 0 \\ 0 & 0 & q-s & 0 \\ s & p & s+d & q+d \end{pmatrix}. \quad \text{(A5)}$$

Notice that $\mathbf{J}(\hat{\mathbf{c}})$ and $\mathbf{J}(\hat{\mathbf{k}})$ are triangular while $\mathbf{J}(\hat{\mathbf{d}})$ and $\mathbf{J}(\hat{\mathbf{h}})$ are not, but have a structure simple enough that permits their analysis.

The eigenvalues associated with eigenvectors lying on the simplex are

$$\{t-r, 0, t-r\},$$
$$\{-s, -s-c, -c\},$$
$$\{0, t-r-p, t-r-q\},$$
$$\{q+d-s, q+d-p, q-s\},$$

for the corner equilibria $\hat{\mathbf{c}}$, $\hat{\mathbf{d}}$, $\hat{\mathbf{h}}$, and $\hat{\mathbf{k}}$, respectively. Notice that to establish whether an equilibrium is locally stable or not, we can ignore the eigenvalue associated with an eigenvector that does not lie on the simplex (this is the eigenvector $\mathbf{e}$ such that $\mathbf{e}^T \cdot \mathbf{1} \neq 0$).

It is now possible to analyze the local stability of these equilibria. Only $\hat{\mathbf{d}}$ is always stable, and $\hat{\mathbf{c}}$ is always unstable (keep in mind that $t > r$). $\hat{\mathbf{h}}$ can be stable in the directions pointing toward $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$ if $p > t - r$ and $q > t - r$, respectively. However, $\hat{\mathbf{h}}$ is always neutrally stable in the direction pointing toward $\hat{\mathbf{c}}$. Finally, if $q + d < s$, $q + d < p$, and $q < s$, then $\hat{\mathbf{k}}$ is stable in the directions of $\hat{\mathbf{c}}$, $\hat{\mathbf{d}}$ and $\hat{\mathbf{h}}$, respectively (these results are summarized in Fig. A2).

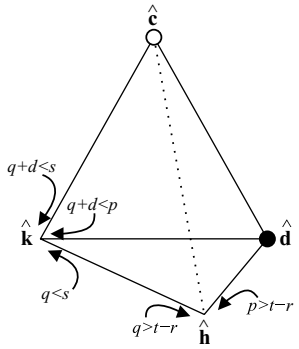

Figure A2. Conditions for stability of the four corners of the simplex. If the condition is satisfied, then the direction pointed by the arrow behaves as a local attractor. The dashed line shows the neutral stability between ĉ and ĥ. d̂ is always stable, denoted by the filled circle, whereas ĉ is always unstable, denoted by the open circle. Filled and open circles denote stable and unstable equilibria in the full game (game that includes all strategies).

Edges

For convenience, we use $\hat{\mathbf{y}}$ to represent an equilibrium in an edge of the simplex, and subscripts *ij* to represent which set of strategies is present at equilibria, for example $\hat{\mathbf{y}}_{DH} = (\hat{y}_D, \hat{y}_H)$ which corresponds to equilibrium $\hat{\mathbf{x}}_{DH} = (0, \hat{x}_D, \hat{x}_H, 0)$ in the complete game (see Fig. A1).

Game between C and D. This is the case corresponding to the well-known Prisoner's Dilemma.

$$\mathbf{A}_{CD} = \begin{pmatrix} r & -s \\ t & 0 \end{pmatrix} \tag{A6}$$

Game between C and H. This game is characterized by the payoff matrix

$$\mathbf{A}_{CH} = \begin{pmatrix} r & r \\ r & r \end{pmatrix} \tag{A7}$$

which corresponds to a degenerate game.

We can characterize the behavior of the neutral line joining $\hat{\mathbf{c}}$ and $\hat{\mathbf{h}}$. We parameterize this line as $(\lambda, 0, 1-\lambda, 0)$ where $0 \leqslant \lambda \leqslant 1$. The eigenvalues of the Jacobian matrix evaluated in this line are $\{t-r-(1-\lambda)q, t-r-(1-\lambda)p, 0, -r\}$ where the eigenvector associated with eigenvalue $-r$ lies outside of the simplex and can be ignored.

Notice that $t-r-(1-\lambda)\gamma < 0$ if and only if $\frac{\gamma-(t-r)}{\gamma} > \lambda$ (where $\gamma = \{p, q\}$). Given that λ belongs to [0, 1], the latter condition requires that $\frac{\gamma-(t-r)}{\gamma} < 1$, which is true in all cases, and that $\frac{\gamma-(t-r)}{\gamma} > 0$, which is true only when $\gamma > t+r$. Therefore, there is a sufficiently large λ such that a segment of the neutral line is unstable. This segment is stable when both p and q are greater than $t-r$. Therefore, there is always a small enough λ such that $t-r-(1-\lambda)p < 0$ and $t-r-(1-\lambda)q < 0$ (i.e., there is a segment that is stable). Notice that $q > t-r$ and $p > t-r$ are the conditions for the stability of $\hat{\mathbf{h}}$ in the reduced games between H and K, and H and D, respectively (see Fig. A2).

In summary, when: (1) $p, q < t-r$ the neutral line joining $\hat{\mathbf{c}}$ and $\hat{\mathbf{h}}$ is unstable; (2) $p < t-r < q$ or $q < t-r < p$ the segment of the neutral line starting at $\hat{\mathbf{h}}$ and ending in $(\frac{\gamma_{mx}-(t+r)}{\gamma_{mx}}, 0, \frac{(\gamma_{mx}+r)}{\gamma_{mx}}, 0)$ (where $\gamma_{mx} = max\{p, q\}$) is a saddle. The dynamics of the system near this segment is pushed toward $\hat{\mathbf{d}}$ or $\hat{\mathbf{k}}$ depending on whether $p < q$ or $q < p$, respectively. The rest of the line is unstable ending in $\hat{\mathbf{c}}$; (3) $t-r < p, q$ the segment of the neutral line starting in $\hat{\mathbf{h}}$ and ending in $(\frac{\gamma_{mn}-(t+r)}{\gamma_{mn}}, 0, \frac{(\gamma_{mn}+r)}{\gamma_{mn}}, 0)$ (where $\gamma_{mn} = min\{p, q\}$) is stable. The rest of the line is first a saddle (which can be degenerate of length zero if $p = q$) and later unstable as in the previous case. These results are summarized in Figure A3.

Game between H and K. This game is characterized by the payoff matrix

$$\mathbf{A}_{HK} = \begin{pmatrix} r & -s-d \\ t-q & -q-d \end{pmatrix} \tag{A8}$$

and has an interior equilibrium (derived from solving $\mathbf{A}_{HK}\hat{\mathbf{y}}_{HK} = \mathbf{1}_2$) of the form

$$\hat{\mathbf{y}}_{HK} = \frac{1}{t-r-s}\begin{pmatrix} q-s \\ t-r-q \end{pmatrix} \tag{A9}$$

if and only if $t-r > q > s$.

These conditions are equal to the instability conditions for $\hat{\mathbf{h}}$ and $\hat{\mathbf{k}}$ in the edge joining them (see Fig. A2). Therefore, we expect equilibrium $\hat{\mathbf{y}}_{HK}$ to be stable in the game restricted to edge $\hat{\mathbf{h}}$ and $\hat{\mathbf{k}}$. The eigenvalues of $\mathbf{J}(\hat{\mathbf{x}}_{HK})$ are $\{(t-r-q)d, (t-r-q)d-(r+s)q+ts, d(t+q)+(s-t)(q-p)-(p-(q+d))r, (s-q)(t-r-q)\}$ up to the positive factor $(t-r-s)^{-1}$. The eigenvector associated with the second eigenvalue, $(0, 0, \frac{q-s}{t-r-q}, 1)^T$, lies outside the simplex, thus this eigenvalue can be ignored. Eigenvalue $(t-r-q)d$ is always positive and then equilibrium $\hat{\mathbf{y}}_{HK}$ is unstable in the full game.

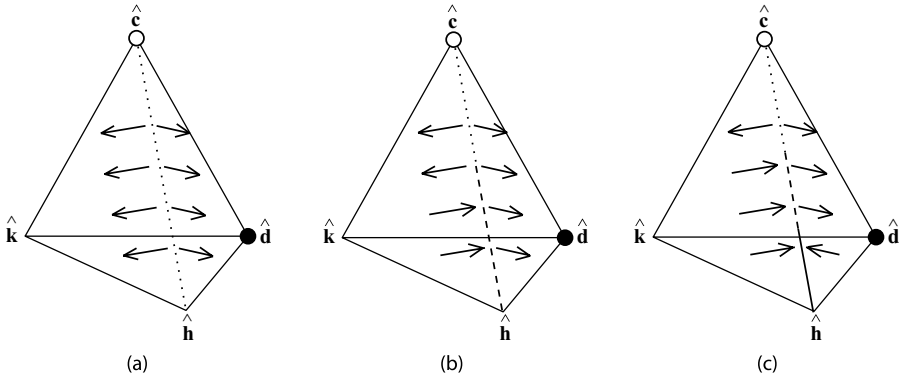

Figure A3. Dynamics on the neutral line between $\hat{\mathbf{c}}$ and $\hat{\mathbf{h}}$. Fine dash represents unstable points, coarse dashed represents saddle and solid represents stable. The cases are in the same order as in the text. For illustration purposes, we assume $p < q$. There are three possible cases: (a) $p, q < t-r$, (b) $p < t-r < q$, (c) $t-r < p, q$.

Game between D and K. This game is characterized by the payoff matrix

$$\mathbf{A}_{DK} = \begin{pmatrix} 0 & -p \\ -c & -q-d \end{pmatrix} \tag{A10}$$

and has an interior equilibrium of the form

$$\hat{\mathbf{y}}_{DK} = \frac{1}{p+c-(q+d)} \begin{pmatrix} p-q-d \\ c \end{pmatrix} \tag{A11}$$

if and only if $p > q + d$.

This condition is equal to the stability condition for $\hat{\mathbf{k}}$ in the edge between $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$ (see Fig. A2). The eigenvalues of $J(\hat{\mathbf{x}}_{DK})$ are $\{c(p-s) - s(p-q-d), cp, c(p-q-d), c(q-s) - s(p-q-d)\}$ up to the positive factor $(p+c-q-d)^{-1}$. Eigenvalue $c(p-q-d)$ (with eigenvector $(0, -1, 0, 1)^T$ in the simplex) is always positive and equilibrium $\hat{\mathbf{y}}_{DK}$ is unstable in the full game.

Game between D and H. This game is characterized by the payoff matrix

$$\mathbf{A}_{DH} = \begin{pmatrix} 0 & t-p \\ -s-c & r \end{pmatrix} \tag{A12}$$

and has an interior equilibrium of the form

$$\hat{\mathbf{y}}_{DH} = \frac{1}{p+c-(t-r-s)} \begin{pmatrix} p+r-t \\ s+c \end{pmatrix} \tag{A13}$$

if and only if $p > t - r$.

This condition is equal to the stability condition for $\hat{\mathbf{h}}$ in the edge between $\hat{\mathbf{d}}$ and $\hat{\mathbf{h}}$ (see Fig. A2). Therefore, we expect equilibrium $\hat{\mathbf{y}}_{DH}$ to be unstable in the game restricted to edge $\hat{\mathbf{d}}$ and $\hat{\mathbf{h}}$. The eigenvalues of $\mathbf{J}(\hat{\mathbf{x}}_{DH})$ are $\{(p+r-t)(s+c), (s+c)(p-t), (r+q-t)c + (q-p)s, c(p+r-t)\}$ up to the positive factor $(p+c+r+s-t)^{-1}$. Eigenvalues $(p+r-t)(s+c)$ and $c(p+r-t)$ are positive and equilibrium $\hat{\mathbf{y}}_{DH}$ is unstable in the full game.

Game between C and K. This game is characterized by the payoff matrix

$$\mathbf{A}_{CK} = \begin{pmatrix} r & -s \\ t & -q-d \end{pmatrix} \tag{A14}$$

and has an interior equilibrium of the form

$$\hat{\mathbf{y}}_{CK} = \frac{1}{t-r-s+q+d} \begin{pmatrix} q+d-s \\ t-r \end{pmatrix} \tag{A15}$$

if and only if $q + d > s$.

This condition corresponds to the instability condition for $\hat{\mathbf{k}}$ in the edge between $\hat{\mathbf{c}}$ and $\hat{\mathbf{k}}$ (see Fig. A2). The eigenvalues of $\mathbf{J}(\hat{\mathbf{x}}_{CK})$ are $\{d(r-t), (r-t)(q+d-s), ts - r(q+d), (r-t)(p-q-d)\}$ up to the positive factor $(q+d+t-s-r)^{-1}$. The eigenvector associated with the third eigenvalue, $(\frac{q+d-s}{t-r}, 0, 0, 1)^T$, lies outside the simplex thus this eigenvalue can be ignored. The first and second eigenvalues are always negative. From the fourth eigenvalue, we derive the condition for the stability of $\hat{\mathbf{y}}_{CK}$ in the full game that is $p > q + d$. This condition is equal to the stability condition of $\hat{\mathbf{k}}$ in the edge between $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$.

Faces

For convenience, we use $\hat{\mathbf{z}}$ to represent an equilibrium in a face of the simplex, and subscripts *ijk* to represent which set of strategies are present at the equilibrium, for example $\hat{\mathbf{z}}_{DHK} = (\hat{z}_D, \hat{z}_H, \hat{z}_K)$ which corresponds to $\hat{\mathbf{x}}_{DHK} = (0, \hat{x}_D, \hat{x}_H, \hat{x}_K)$ in the complete game (see Fig. A1).

Game without C. This game is characterized by the payoff matrix

$$\mathbf{A}_{DHK} = \begin{pmatrix} 0 & t-p & -p \\ -s-c & r & -s-d \\ -c & t-q & -q-d \end{pmatrix}. \tag{A16}$$

Solutions to the system $\mathbf{A}_{DHK}\hat{\mathbf{z}}_{DHK} = \mathbf{1}_3$ take the form

$$\hat{\mathbf{z}}_{DHK} = \alpha_{DHK} \begin{pmatrix} (t-r-s)(p-q-d) + (q-s)d \\ c(q-s) - s(p-q-d) \\ c(t-r-q) + s(p-q) \end{pmatrix} \tag{A17}$$

where $\alpha_{DHK} = (qd + (c+p-q-d)(t-r-s))^{-1}$. If $p > q + d$, then $qd + (c+p-q-d)(t-r-s) > 0$, and equilibrium $\hat{\mathbf{z}}_{DHK}$ is in the interior of the simplex if and only if $q > s$. The former condition, $p > q + d$, implies that there is an unstable equilibria on the edge between $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$. The latter condition, $q > s$, implies that $\hat{\mathbf{k}}$ is unstable in the direction of $\hat{\mathbf{h}}$. Internal equilibrium $\hat{\mathbf{z}}_{DHK}$ must be unstable in the direction of $\hat{\mathbf{d}}$. There might be a stable equilibrium in the edge between $\hat{\mathbf{h}}$ and $\hat{\mathbf{k}}$, but its stability is not preserved in the presence of cooperating civilians (shown in the analysis of the game between *H* and *K*). The actual dynamics (shown in Fig. A4) depends on the values of p and q with respect to $t - r$. If $p < q + d$ there are two possible dynamics depicted in Figure A4.

Game without D. This game is characterized by the payoff matrix

$$\mathbf{A}_{CHK} = \begin{pmatrix} r & r & -s \\ r & r & -s-d \\ t & t-q & -q-d \end{pmatrix}. \tag{A18}$$

Solutions to the system $\mathbf{A}_{CHK}\hat{\mathbf{z}}_{CHK} = \mathbf{1}_3$ take the form

$$\hat{\mathbf{z}}_{CHK} = \frac{1}{q} \begin{pmatrix} t-r-q \\ r-t \\ 0 \end{pmatrix} \tag{A19}$$

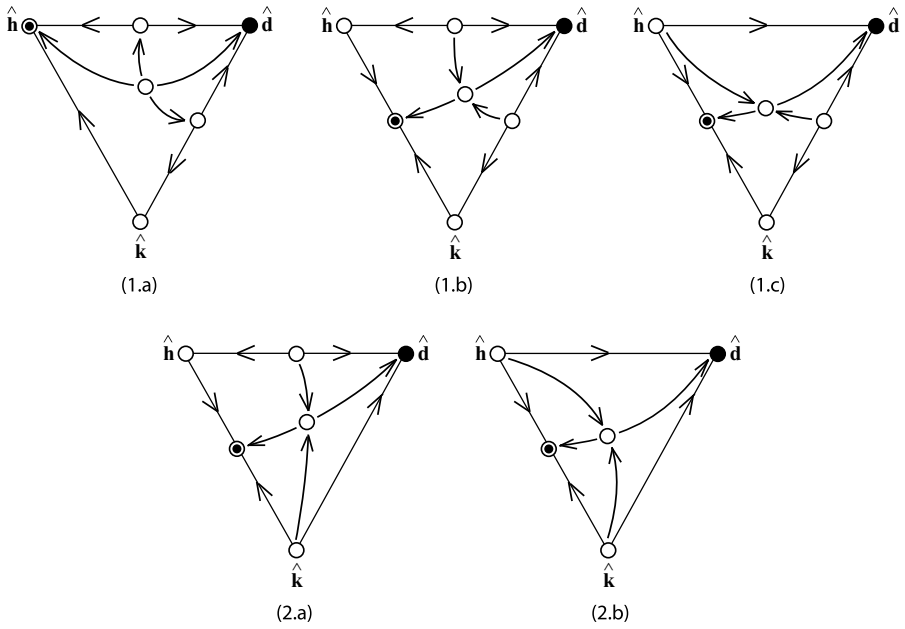

Figure A4. Dynamics of the reduced game without *C*. For simplicity, we illustrate the cases when there is an internal equilibrium. (1) Condition $p > q + d$ is satisfied. (1.a) $p > t - r > s > q$, (1.b) $p > t - r > q > s$, (1.c) $q > t - r > p$. (2) Condition $p < q + d$ is satisfied. (2.a) $p > t - r > q > s$ and, (2.b) $q > t - r > p$. When considering reduced games we continue using filled and open circles to denote stable and unstable equilibria in the full game. There is an additional possibility corresponding to the case of an equilibrium showing stability in the reduced game but instability in the full game that we denote by a filled circle contained in an open circle.

which cannot be an interior equilibrium. The dynamics of this reduced game (shown in Fig. A5) depend on values of p with respect to $q + d$ as well as of q with respect to $t - r$ and s. For simplicity, Figure A5 shows only the case when $p > q + d$.

Game without H. This game is characterized by the payoff matrix

$$\mathbf{A}_{CDK} = \begin{pmatrix} r & -s & -s \\ t & 0 & -p \\ t & -c & -q-d \end{pmatrix}. \tag{A20}$$

Solutions to the system $\mathbf{A}_{CDK}\hat{\mathbf{z}}_{CDK} = \mathbf{1}_3$ take the form

$$\hat{\mathbf{z}}_{CDK} = \alpha_{CDK} \begin{pmatrix} c(p-s) - s(p-q-d) \\ (t-r)(p-q-d) \\ (t-r)c \end{pmatrix} \tag{A21}$$

where $\alpha_{CDK} = (cp + (t - r - s)(c + p - q - d))^{-1}$. Because $(t - r)c > 0$, equilibrium $\hat{\mathbf{z}}_{CDK}$ is in the interior of the simplex when $cp + (t - r - s)(c + p - q - d) > 0$. This implies that both $c(p - s) - s(p - (q + d))$ and $(t - r)(p - q - d)$ must be positive, which requires that $q + d < p$. This condition is equal to the instability condition of $\hat{\mathbf{k}}$ in the edge between $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$.

The eigenvalues of $\mathbf{J}(\hat{\mathbf{x}}_{CDK})$ are $\{c(t-r)(q-p), c(t-r)(p-q-d), (t-r)(s(p-q-d) - c(p-s)), ts(p-q-d) - c(rp-ts)\}$ up to the positive factor α_{CDK}. Eigenvalue $c(t - r)(p - q - d)$ is always positive and has its corresponding eigenvector in the simplex indicating that equilibrium $\hat{\mathbf{z}}_{CDK}$ is unstable. In particular, this internal equilibrium is a saddle point attracting from $\hat{\mathbf{c}}$ and $\hat{\mathbf{y}}_{DK}$ and repelling toward $\hat{\mathbf{d}}$.

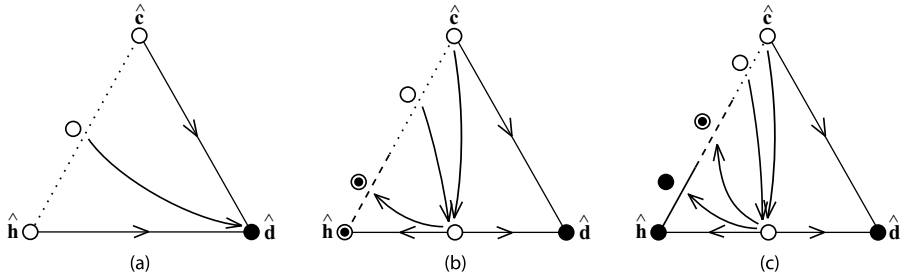

Figure A5. Dynamics of the reduced game without *D*. For simplicity, we illustrate the cases when $p > q + d$. (a) $t - r > s > q$, (b) $t - r > q > s$, (c) $q > t - r > s$.

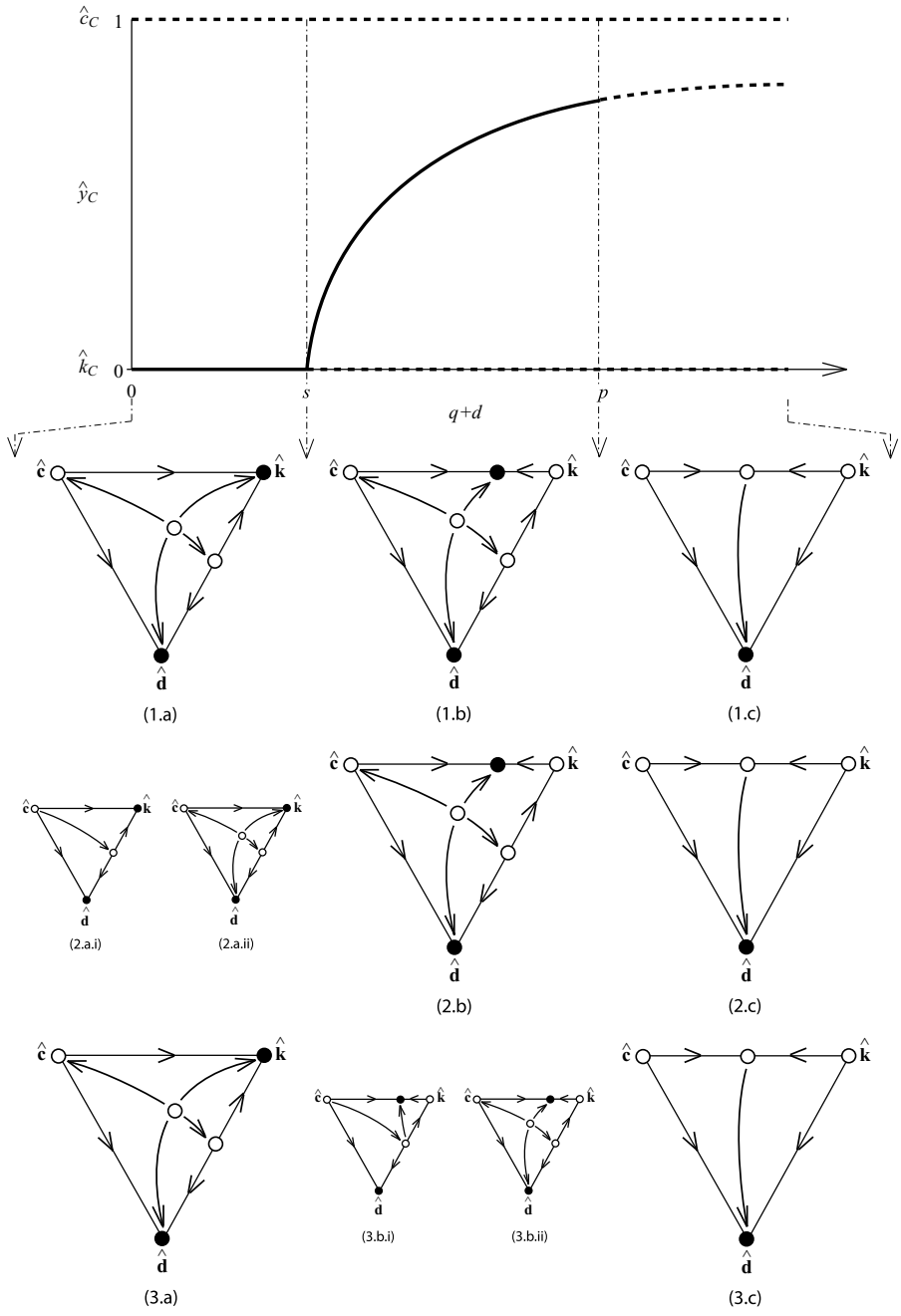

Figure A6. **Dynamics of the reduced game without *H* and bifurcation diagram of the equilibria on the edge between $\hat{\mathbf{c}}$ and $\hat{\mathbf{k}}$. Bifurcation diagram of the equilibria on the edge between $\hat{\mathbf{c}}$ and $\hat{\mathbf{k}}$. The horizontal axis corresponds to the value taken by $q + d$. The vertical axis corresponds to the frequency of *C* in the population at equilibrium. When $0 < q + d < s$ the frequency of *C* is 0. When $s < q + d < p$ the frequency of *C* is greater than 0 and its value is indicated by a solid line. When $p < q + d$ the frequency of *C* is greater than 0 and its value is indicated by a dashed line. Solid and dashed lines in this figure represent stable and unstable equilibria. Dynamics of the reduced game without *H*. (1) Condition $R < 0$ is satisfied. (2) Condition $0 < R < s$ is satisfied. (3) Condition $s < R < p$ is satisfied. Within each of these conditions $q + d$ may be comprised between (a) $0 < q + d < s$, (b) $s < q + d < p$, and (c) $p < q + d$.**

The dynamics of the game without honest policers can be studied by analyzing the stability of the equilibria on the edges of the reduced simplex. There are two possible scenarios. In both cases the interior equilibrium is a saddle point and $\hat{\mathbf{d}}$ is a stable equilibrium. However, if $q + d < s$ then equilibrium $\hat{\mathbf{k}}$ is stable and $\hat{\mathbf{y}}_{CK}$ does not exist, but if $q + d > s$ then equilibrium $\hat{\mathbf{k}}$ is a saddle point and $\hat{\mathbf{y}}_{CK}$ is stable. Therefore, either stable equilibria $\hat{\mathbf{d}}$ and $\hat{\mathbf{k}}$ co-exist and the dynamics of the system converge to one of them, or stable equilibria $\hat{\mathbf{d}}$ and $\hat{\mathbf{y}}_{CK}$ co-exist and the dynamics of the system converge to one of them. The likelihood that the

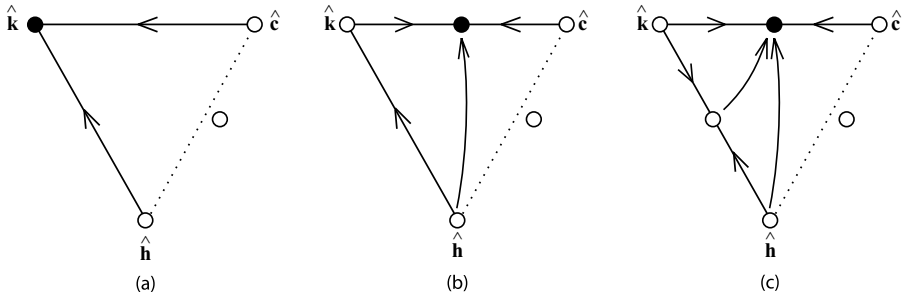

Figure A7. Dynamics of the reduced game without K. (a) $p < t - r$, (b) $p > t - r$ and $q > p > t - r$, (c) $p > q > t - r$.

dynamics of the system converge to one or the other equilibrium is determined by the size of the basins of attraction $B_{\hat{\mathbf{x}}}$ of each co-existing stable equilibria.

In the game without H, the size of the basins of attraction of equilibria $\hat{\mathbf{k}}$ and $\hat{\mathbf{y}}_{CK}$ are determined by the position of equilibria $\hat{\mathbf{y}}_{DK}$ and $\hat{\mathbf{z}}_{CDK}$ (this is an approximation as strategy H is missing. See section Simulations for an analysis of the basin of attraction in the complete game). The bigger $q + d$ respect to p the greater component $\hat{z}_C$ of $\hat{\mathbf{z}}_{CDK}$ but the lesser component $\hat{z}_D$ rendering equilibrium $\hat{\mathbf{z}}_{CDK}$ closer to corner $\hat{\mathbf{d}}$. Equilibrium $\hat{\mathbf{z}}_{CDK}$ only reaches corner $\hat{\mathbf{d}}$ in the limit when p tends to infinity. The bigger $p - q - d$, the closer equilibria $\hat{\mathbf{y}}_{DK}$ and $\hat{\mathbf{z}}_{CDK}$ to corner $\hat{\mathbf{d}}$ and the greater the basins of attraction $B_{\hat{\mathbf{k}}}$ or $B_{\hat{\mathbf{y}}_{CK}}$ versus their alternative basin of attraction $B_{\hat{\mathbf{d}}}$. In contrast, the bigger $q + d$ with respect to s, the closer equilibrium $\hat{\mathbf{y}}_{CK}$ to corner $\hat{\mathbf{c}}$.

Let $R = p - \frac{c}{s}(p - s)$. If $q + d = R$ component $\hat{z}_C$ of $\hat{\mathbf{z}}_{CDK}$ takes the value 0. As $q + d$ crosses threshold R equilibrium $\hat{\mathbf{z}}_{CDK}$ splits from the unstable equilibrium $\hat{\mathbf{y}}_{DK}$. Figure A6 shows the bifurcation diagram of equilibria on the between $\hat{\mathbf{c}}$ and $\hat{\mathbf{k}}$ edge, with respect to $q + d$ in terms of component z_C of $\hat{\mathbf{z}}_{CDK}$. The dynamics of the game without honest policers can be classified using the values taken by $q + d$ and R: (1) When $R < 0$ the dynamics of the game are given by (1.a) when $0 < q + d < s$, (1.b) when $s < q + d < p$, (1.c) when $p < q + d$ in Figure A6; (2) When $0 < R < s$ the dynamics are (2.a.i) when $0 < q + d < R$, (2.a.ii) when $R < q + d < s$, (2.b) when $s < q + d < p$, (2.c) when $p < q + d$ in Figure A6; (3) When $s < R < p$ the dynamics are (3.a) when $0 < q + d < s$, (3.b.i) when $s < q + d < R$, (3.b.ii) when $R < q + d < p$, (3.c) when $p < q + d$ in Figure A6

Game without K. This game corresponds to the Punishment Game and is characterized by the payoff matrix

$$\mathbf{A}_{CDH} = \begin{pmatrix} r & -s & r \\ t & 0 & t-p \\ r & -s-c & r \end{pmatrix}. \tag{A22}$$

Solutions to the system $\mathbf{A}_{CDH}\hat{\mathbf{z}}_{CDH} = \mathbf{1}_3$ take the form

$$\hat{\mathbf{z}}_{CDH} = \frac{1}{p}\begin{pmatrix} p+r-t \\ 0 \\ t-r \end{pmatrix} \tag{A23}$$

which cannot be an interior equilibrium. The dynamics of this reduced game (shown in Fig. A7) depend on values of p and q with respect to $t - r$.

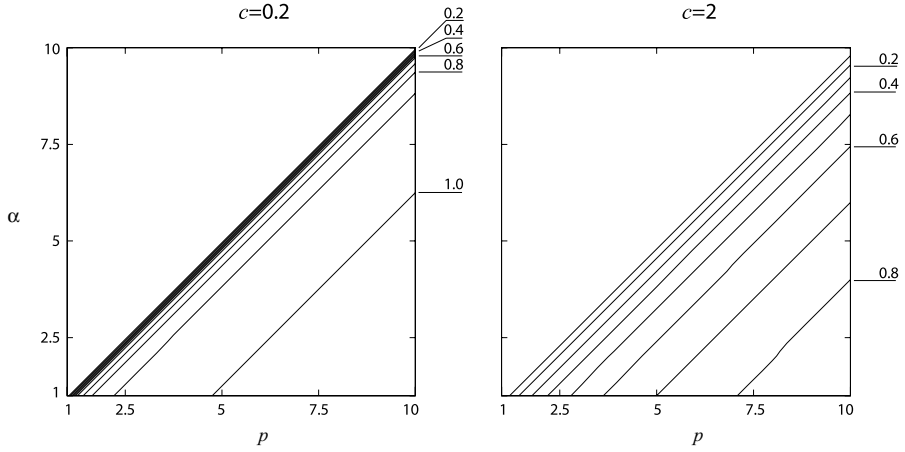

Figure A8. Dynamics of the system in the vicinity of $\hat{\mathbf{c}}$. We present two different cases: low c (in particular $c = 0.2$) and large c (in particular $c = 2$). The horizontal axis corresponds to the value of p. The vertical axis corresponds to the value of α. Isoclines represent the proportion of the population converging to $\hat{\mathbf{y}}_{CK}$ (values indicated to the right of each line).

BASIN OF ATTRACTION: SIMULATIONS

To estimate the basin of attraction of each of the equilibria, we simulated the dynamics corresponding to Equation (A2) numerically. All runs were performed for the set of parameter values $t = 4$, $r = 2$, and $s = 1$.

We conduct the analysis for p and $q + d$ with values between 1 and 10 in increments of 0.25, and c between 0.2 and 2 with increments of 0.2. Given a value of the parameters p, $q + d$ and c, we analyze the dynamics close to the simplex's corners $\hat{\mathbf{c}}$, $\hat{\mathbf{h}}$ and $\hat{\mathbf{k}}$. The corner $\hat{\mathbf{d}}$ is not analyzed for it is always stable, and thus, all small perturbations return to the equilibrium. For each of these three cases, we have a uniform set of initial populations $\mathbf{x} = \mathbf{i} + \mathbf{e}$, where $\mathbf{i} \in \{\hat{\mathbf{c}}, \hat{\mathbf{h}}, \hat{\mathbf{k}}\}$ and $\mathbf{e}$ is a small perturbation so that $\sum_{j=1}^{4} \mathbf{e}_j = \epsilon$. In our case, $\epsilon = 0.01$. Finally, we simulate the dynamical system given by Equation (A2) until the population is close enough to one of the three observed equilibria: $\hat{\mathbf{y}}_{CK}$, $\hat{\mathbf{y}}_{CH}$ or $\hat{\mathbf{d}}$.

We are interested in what can evolve from a population of civilian cooperators. Therefore for corner $\hat{\mathbf{c}}$, we report the proportion of populations that end in each of the three possible equilibria (see Fig. A8). These proportions are a numerical approximation of the equilibrium's basin of attraction.

Note that, as expected, whenever $q + d > p$, the proportion of runs converging to $\hat{\mathbf{y}}_{CK}$ is zero. In general, as long as $q + d > s$, increasing power asymmetries (by increasing p or decreasing $q + d$) increases the basin of attraction of $\hat{\mathbf{y}}_{CK}$. This is seen more clearly in Figure A8, where the basin $B_{\hat{\mathbf{y}}_{CK}}$ increases smoothly from zero when $p = q + d$ to close to 1 when $p >> q + d$.