Public Choice (2008) 135: 399–413
DOI 10.1007/s11127-007-9269-3

Functional unpleasantness: the evolutionary logic of righteous resentment

William B. Heller · K.K. Sieberg

Received: 7 June 2007 / Accepted: 4 December 2007 / Published online: 8 January 2008

Abstract Economics experiments and everyday experience cast doubt on the assumption that people are self-interested. In divide-the-dollar ultimatum games, participants turn down offers that would make them objectively better off. Similarly, drivers stuck in a traffic jam fume at cars cruising by on the shoulder. Many stuck drivers would punish the moving ones if they could, even at some cost to themselves. Such strategies appear irrational because they make the punisher worse off than accepting the situation or offer. We examine explanations for costly punishment and relax the presumption that punishers themselves prefer cooperation, using evolutionary game theory to show how uncooperative punishers can support cooperation.

Keywords Cooperation · Punishment · Evolutionary games · Altruism

Humans were social animals before they were political animals. People have been cooperating in large groups, without the benefit of small-group dynamics (Olson 1965; Henrich and Boyd 2001), since well before the invention of government (Fehr and Rockenbach 2003; Rubin 2002). Given resource scarcity, conflicts of interests, and such collective dilemmas as the free-rider problem, such cooperation is curious. How does cooperation survive in the face of collective-action problems, mutual suspicion, and people's general—and justified (Cook et al. 2005)—distrust for one another?

Typically, the answer hinges on two factors. First, people have to believe that cooperation is better than a situation where everyone cheats their fellows at every opportunity. Second, people have to be willing to punish such cheaters, *even at a personal cost*. Both ring true:

W.B. Heller (✉)
Department of Political Science, Binghamton University, P.O. Box 6000, Vestal Parkway East, Binghamton, NY 13902, USA
e-mail: wbheller@post.harvard.edu

K.K. Sieberg
Erkko Chair of North American Studies, University of Tampere, Tampere, Finland
e-mail: Katri.sieberg@uta.fi

numerous economics experiments show, and a little honest self-reflection probably will attest, that we humans readily sacrifice some personal well-being (at least when the stakes are low; Henrich 2000) to punish others whose conduct we see as unfair or otherwise inconsistent with societal norms (Axelrod 1984; Bowles and Gintis 2002; Boyd and Richerson 1992; Ensminger 2001; Fehr and Gächter 2002; Fehr and Rockenbach 2003; Henrich and Boyd 2001; Henrich et al. 2001). But paying costs to punish is irrational when the benefits are dispersed throughout society. To make punishment coherent, scholars assume that some actors value cooperation enough to punish noncooperators (see, e.g., Boyd et al. 2003; Boyd and Richerson 1992; Fehr and Fischbacher 2003; Fehr et al. 2002; Fehr and Gächter 2000, 2002; Fehr and Henrich 2003; Fehr and Schmidt 1999; Fowler 2005; Henrich 2004; Ostrom et al. 1992). The rationales for such punishment can vary, as we note in the next section.

We relax the assumption that punishers cooperate. Actors who punish need not play well with others when not punishing; rather, we posit a new type—Unpleasant players who cheat all others and yet punish Cheaters. We find that Unpleasant players are functional for society under a fairly inclusive set of conditions: they can help support cooperation, even though they neither desire nor seek it. One might expect players who are nasty to everyone to affect society negatively: ironically, Functional Unpleasantness supports the survival of naïve Fair players who could not compete against Cheaters alone.

The next section briefly reviews the literature and experimental work exploring the foundations of human cooperation. The third section replaces the altruistic punisher usually found in the literature with a new type. We show that the presence of Unpleasant players induces significant levels of cooperation in society. The final section concludes.

1 Literature review

People expect others to cheat and are very good at recognizing cheating behavior (Fehr and Fischbacher 2003). No one really wants to cooperate in the absence of some means of ensuring that cheating does not pay (Sieberg 2005), precisely because absent some kind of punishment cheating *does* pay. Nonetheless, "large groups of unrelated individuals" cooperate to build societies without the benefit of government (Wedekind 1998; see also, e.g., Bowles and Gintis 2002; Fehr and Fischbacher 2003; Fehr and Gächter 2002; Henrich and Boyd 2001; Rubin 2002). The key conditions for such cooperation are two. First, social norms define what people consider "fair" behavior (Fehr and Rockenbach 2003), which can vary across societies (see Vogel 2004). And second, a willingness on the part of individuals or groups to punish violations of those social norms. As long as the threat to punish those who cheat on their societal obligations is credible (Levi 1989; and cf. Flack et al. 2005a, 2005b, 2006), the expected benefits of cooperation increase. That the latter condition holds (which in turn implies the former) is clear from experimental results of ultimatum games, in which two players have to agree on a division of a sum of money, with one player proposing a division and the other accepting or rejecting. If the proposal is accepted, the money is divided according to the proposal; if the proposal is rejected, neither player gets anything. Researchers around the world have found that "about half" of all players in myriad plays of the game reject offers that yield them less than 20% of the initial sum (Nowak et al. 2000; and cf. Fehr and Gächter 2002; Güth and Tietz 1990; Henrich and Boyd 2001). Skyrms (2003) shows, using evolutionary game theory, that players in societal ultimatum games generally will settle with splits (the specifics of his model yield a 50–50 split).

When action is collective, each person's behavior can affect everybody else. Even when actions are only minimally collective, as in the case of drivers on a freeway or sunbathers on a beach—for whom others are basically irrelevant unless they get in the way—each individual's enjoyment depends on the comportment of others. Most individuals, moreover, have incentives to cheat: the driver who weaves in and out of lanes on the freeway gets where she is going more quickly, but forces other drivers to take defensive actions and slow down; the beachgoer who takes the easiest path to the water kicks sand on sunbathers, causing discomfort.

It is to be expected that those who suffer the effects of such behavior might want to retaliate. Retaliation can be costly, however—a nasty glare could provoke a violent reaction—and often those who suffer are not in a position to retaliate: the driver who has to brake or swerve to avoid a car cutting in front of her can do little to affect the offender, who is already long gone. That said, people do retaliate, even at a cost to themselves, to behavior they perceive to be unfair (Nowak et al. 2000; Herbert Gintis, quoted in Vogel 2004). Moreover, experiments by Fehr and Fischbacher (2004) show that people who observe but are unaffected by cheating behavior are willing to punish it, even at a cost to themselves. To reiterate, the general conclusion that researchers draw from results such as these is that people value cooperation and are willing to expend resources to support it. The motives underpinning this willingness to punish can vary.

A frequent explanation is culture. Cultural familiarity can breed trust, creating the repeated play conditions that support opportunities for cooperation. Alesina and La Ferrara (2000, 2002) find that group heterogeneity increases the transaction costs of social experiences with members of other groups. Barr (1999) and Coleman (1990) argue that "familiarity breeds trust." The cultural effect, however, relies on the assumptions that groups are small and homogeneous enough for players to have sufficient information to sanction one another for cheating (Saari-Sieberg 1998). It also relies on the assumptions that members value cooperation and that they care enough about other members in their group that they will both cooperate with them and will punish cheaters. We agree that these assumptions are possible, but find them unnecessarily strong. The cultural idea finds mixed experimental support. In games that tested willingness to contribute to a group, Brandts, Saijo, and Schram (2004) found little evidence for cross-country differences in behavior in Japan, the Netherlands, and Spain. However, Ockenfels and Weimann (1999), in comparisons of games between East and West Germans, found more selfish behavior in East Germans than in their West German counterparts.

Cooperators need not be selfless altruists who both cooperate and punish noncooperators, even though "cheating would be economically beneficial for them" (Fehr and Rockenbach 2003). Some scholars focus on the conditional nature of cooperation, which is worthwhile only if enough *other* people are cooperating to make cooperation worthwhile (see, e.g., Boyd and Richerson 1992; Camerer and Fehr 2006; Fehr and Fischbacher 2003; Hibbing and Alford 2004; Sigmund and Nowak 2000). Others take more nuanced views, arguing in the spirit of the chainstore paradox (Selten 1978) that punishment is in essence a strategic investment in personal reputation that reduces the chance of being cheated in the future (see, e.g., Fehr and Fischbacher 2003; Fehr and Henrich 2003). Alternatively, if one response to cheating is for non-cheaters to stop participating in production of a public good, punishers keep cheating at a low enough level that the increase in the public good makes up for the cost of punishing (Fowler 2005; and cf. Sethi and Somanathan 1996). Underpinning this latter approach is the notion that there are two levels of selection and payoffs: on one hand, competition among individuals means that individuals who punish free riders suffer reduced fitness; on the other hand, competition among

groups favors those that include such punishers and thus are better equipped to engage in collective action to produce, employ, and consume collective goods (Henrich 2004; Sethi and Somanathan 1996; and cf. Rubin 2002).

We seek to extend understanding of cooperation by relaxing the assumption that actors who punish noncooperators would themselves prefer to cooperate. The key is the willingness to punish, not the desire to cooperate. The ideal situation for a free rider, after all, is to be the only cheater in the group. The possible origins of the drive to punish are many—it could be ingrained and emotional (Bowles and Gintis 2002; Sanfey et al. 2003; and see Bewley 2003), strategic (Fehr and Fischbacher 2003; Fehr and Henrich 2003), or based on reciprocity built on an altruistic concern for fairness (Dawes et al. 2007; Fehr and Schmidt 1999; Pulkkinen 2007). This unpleasantness factor, which appears to be a simple character flaw, can function to allow cooperation to persist. Our analysis focuses only on within-group selection, but clearly establishes that cooperation can survive in the face of widespread, unapologetic cheating. In the next section we use evolutionary game theory to explore the logic of such unintended social harmony.

2 Modeling cooperation

We begin with the simplest case, a society composed of individuals who are either pure Cheaters (C) or pure Fair players (F) (Table 1). Consider, for example, drivers on a crowded, but not jammed, freeway. As long as each driver moves with the flow of traffic and does not interrupt anyone else's progress, everyone arrives at their destination with all deliberate speed. The payoff to F players in such a situation is $\alpha \in [0, 1)$. This is the baseline payoff. When most drivers are playing F, a cheating driver—driving faster than others, passing on both left and right, cutting close enough to other cars that their drivers have to slow or swerve, or both—gets to his destination at least as quickly as he would have had he played F, so we normalize the payoff to a Cheater interacting with a Fair player to 1. If β is the cost of being cheated—e.g., the time lost from having to slow down, or even being involved in an accident occasioned by, but not involving, a C player—then the payoff to an F player interacting with a Cheater is $\alpha - \beta$ ($\alpha > \beta$, by assumption).[1] When two Cheaters interact with each other, one receives a payoff of 1 and the other receives a payoff of $1 - \beta$. This is the situation when one speeding driver is overtaken and cut off by another, faster driver: a player can either cheat or be cheated, but not both. If everyone plays C, then everyone's net payoff is $\frac{1+1-\beta}{2}$.

When the only possible strategies are pure C and pure F, the result is straightforward: C is the unique evolutionarily stable strategy (ESS) if $2(1 - \alpha) > -\beta$, or $\beta > 2(\alpha - 1)$. Since $\alpha < 1$ by definition, the right-hand side of this equation is negative. Thus, as long as $\beta \geq 0$,

Table 1 Pure-strategy payoffs

When meeting	Expected payoff to	
	F	C
F	α	$\alpha - \beta$
C	1	$\frac{2-\beta}{2}$

[1] Playing F when most everyone else plays C is a naïve and unprofitable strategy. The payoff to C is the highest possible payoff, normalized to 1.

C will be the ESS. A few C players inserted into a population of F players would thrive, as most of their interactions would be with F players and only rarely would they meet with C players and receive a payoff of less than 1. Even when C players mostly interact with other C players they still do better than F players, who also mostly interact with F players. If we take payoffs as a measure of survival fitness, C players have higher fitness than F players and, over time, would entirely displace them.

Note that behavior that hurts the interests of others need not be purposeful. It is, rather, essentially selfish: most people probably do not bother going out of their way to upset others but they might be willing to upset others in pursuit of their own interests—e.g., getting to the head of a queue, bypassing slower drivers, or taking the quickest route to the water over hot sand. Whether such behavior upsets others depends on expectations. If everyone tries to cut to the head of the line, for example, there will be no line. If there is no line and everyone is pushing to get to the front of a chaotic crowd, then cheating is expected, hence accepted; if there is a line, someone who cuts to the front is likely to receive nasty glares and harsh comments, and, possibly, more tangible punishments.

Given a minimal, but plausible set of assumptions about payoffs, a world where the only possible behaviors are cheating and playing fair is a world without cooperation. Given the apparent tendency of at least some people to cheat when they can, cooperation requires something more than an appreciation of its benefits to survive. Adding Fair players who also punish Cheaters to the mix can do the trick, as numerous experiments have shown (for a general model and theoretical exploration, see Boyd and Richerson 1992). As long as punishments impose costs on Cheaters that are greater than the benefits they gain from cheating, a population with players that punish will not evolve to pure C.

The mechanics of punishment by players who prefer to play fair have been explored extensively elsewhere, as documented above, so we do not expound in detail on them here. To reiterate, the addition of players who play fair but punish cheating makes cheating less profitable. Cheaters will survive as long as punishments are not too onerous; so also will both Fair and Altruistic players, provided that the cost of imposing punishment is not too great. The ESS will be a mix of player types, with proportions depending on the values attached to punishing, being punished, and being cheated.

2.1 Equilibria in populations with three player types

Fair players and Cheaters can coexist when Altruistic players are part of the societal mix. Underpinning this conclusion, however, is the assumption that individuals can sublimate their personal interests to those of society. We think this assumption is too strong: taken to extremes, it is tantamount to claiming that the Prisoner's Dilemma and the Tragedy of the Commons are illusions. We relax this assumption by positing a new type of Unpleasant punishing player (U). Unlike the Altruistic punisher, this player makes no sacrifices for society: s/he cheats when interacting with other players (thereby earning the highest possible payoff for those interactions) and punishes when interacting with (identified) Cheaters.

Table 2 provides a summary of our basic notation. If the costs of punishing and being punished are ϕ and γ, respectively, then Table 3 shows the payoff matrix for a society comprising Fair players, Cheaters, and Unpleasant players. Unpleasant can be an ESS if

$$\frac{2-\phi-\gamma}{2} > 1-\gamma \quad \text{and} \quad \frac{2-\phi-\gamma}{2} > \alpha-\beta.$$

The first condition holds if $\beta > \phi$; the second holds if $2-\phi-\gamma > 2(\alpha-\beta)$, which simplifies to $2(1-\alpha)+\beta-\phi > \gamma-\beta$. In other words, U is an ESS only if the cost of being cheated

Table 2 Basic notation and parameter limits

$\alpha \in [0, 1)$	Payoff to fair player meeting fair player (baseline)
$\beta, \beta < \alpha$	Cost of being cheated
$\phi, \phi < \gamma$	Cost of punishing
$\gamma, \gamma > \beta, 1 - \gamma < \alpha$	Cost of being punished

Table 3 Pure-strategy payoffs with punishment

When meeting	Expected payoff to		
	F	C	U
F	α	$\alpha - \beta$	$\alpha - \beta$
C	1	$\frac{2-\beta}{2}$	$\frac{2-\gamma-\beta}{2}$
U	1	$\frac{2-\phi}{2}$	$\frac{2-\phi-\gamma}{2}$

is greater than the cost of punishing and the difference between the cost of being punished and the cost of being cheated is less than twice the difference between (a) the payoff to C and the payoff to F plus (b) the difference between the cost of being cheated and the cost of punishment. The result is intuitive. If being cheated is costly and a Cheater does better than a Fair player, then an Unpleasant player who both cheats and punishes can do better among its own type than would either a pure-Fair or pure-Cheating player.

Unpleasantness is not the sole ESS. Cheating also can be an evolutionarily stable strategy if $\frac{2-\beta}{2} > \frac{2-\phi}{2}$ (that is, if $\phi > \beta$), because punishment is so costly that Unpleasant types do not thrive. Cheaters cannot coexist with Unpleasant players if U is an ESS because of the opposing constraints regarding the cost of punishment and the cost of being cheated ($\beta > \phi$ or $\phi > \beta$).[2] The cost of punishment is lower than the cost of being punished, for example, in the case of an Unpleasant speeding driver who swerves in and out of traffic, cutting off others. If that driver sees another Cheater coming from behind, the cost to her of slowing down in front of the second driver, blocking him and forcing him to slow as well, can be lower than the cost of being blocked. The blocked-in driver has to slow more than the driver doing the blocking. The situation would be very different, however, if the blocked driver might succumb to road rage and respond with lethal force—thus inflicting an even higher cost on the punisher.[3] Alternatively, the costs of punishing could be higher than the costs of being cheated in face-to-face confrontations, where for instance a person might want to punish a larger person for cutting in line. Accordingly, for the same reason, there cannot be a mixed, polymorphic ESS in a population with p_1 F players, p_2 C players, and $(1 - p_1 - p_2)$ U players.

2.2 Global equilibria in populations with three player types

Thus far, we have identified conditions for local ESS with various population-proportion mixes of the three types of players. Using the qualitative approach toward evolutionary game theory developed by Saari (2002), we also can employ these equations to extract global

[2] Of course, punishing when $\phi > \beta$ seems insane. Here, however (consistent with Sanfey et al. 2003), punishing Cheaters is an imperative, not a choice, for U types.

[3] We are exploring the implications of allowing mixed strategies in another paper.

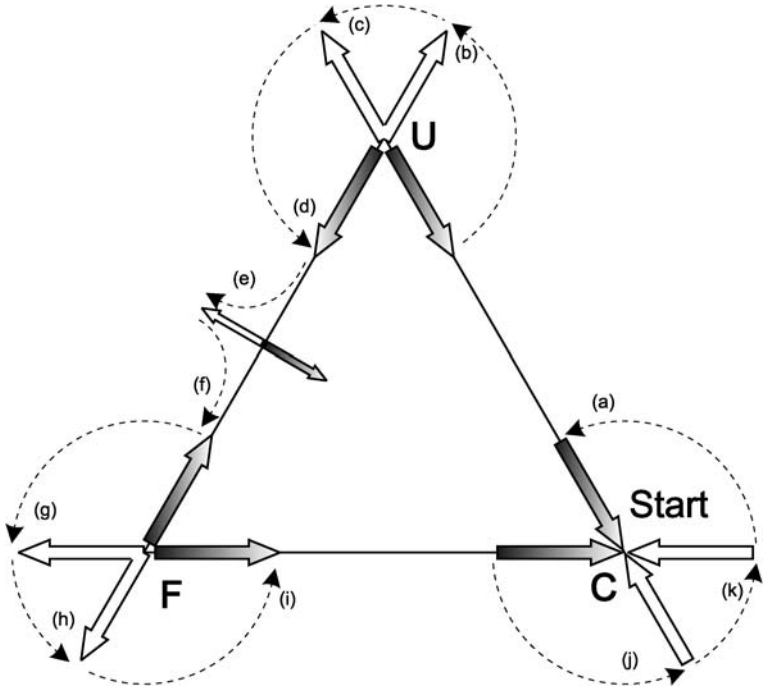

Fig. 1 Population simplex for fair, cheating, and unpleasant types ($\beta < \phi$)

information about the entire population and its dynamics. The equilateral triangle in Fig. 1 represents the simplex $\{(x, y, z) \mid x + y + z = 1; x, y, z \geq 0\}$ (Saari 2002), which we can use to depict the population proportions of F, C, and U types.

In the discussion below, we examine only the case where Cheaters are readily identifiable, and punishing precludes being cheated. In light of humans' ability to recognize Cheaters (Fehr and Fischbacher 2003) and our traffic analogy, we think this case worthy of detailed discussion. We analyze the more general case, where U players punish all Cheaters, but sometimes are cheated and sometimes are not, in Appendix B.[4] For discussion of the case where Cheaters can be identified, the simplex in Fig. 1 depicts the circumstance where the cost of punishing is higher than the cost of cheating.

2.3 When punishing is more costly than being cheated

In Fig. 1, the vertices of the triangle depict situations where the population consists of only one type of players. The lower left vertex, labeled F, is all Fair players; the lower right vertex is all Cheaters; and the upper vertex is all U types. Any point along one of the sides of the triangle (not at a vertex) represents a mix between the two population types at the angles defining the side, and any point *inside* the triangle represents a mix of all three types. The arrows indicate which type fares best overall in interacting with the other types, in the proportions defined by the point in question. Thus, the bottom side of the triangle represents a population composed of F and C players, and C players do better than F players in any

[4] Appendices are available at http://bingweb.binghamton.edu/~wheller/quorum/FU1.

population that comprises any combination of only those two types (see Table 3 and the accompanying discussion).

The C–U side is similar, as U players cannot coexist with Cheaters if being cheated is less costly than punishing ($\beta < \phi$). U players would exhaust all their resources punishing Cheaters, but because they both cheat and punish they also inflict costs on each other and realize no net gain from punishing Cheaters. Cheaters thus dominate in any population of only C and U players; a pure-U population also is vulnerable to F (as well as C) players, as indicated by the arrows pointing away from the U vertex on both sides.

The dynamics along the U–F side, representing population proportions of Fair and Unpleasant players, are more complex. Unpleasant players thrive in a population dominated by F players because they benefit from cheating the Fair players and are punished only if they meet each other, which would happen rarely. Fair players do well in a population of mostly U players, however, because even though they are cheated they neither incur costs of punishing others nor pay penalties for cheating.[5] There is thus an equilibrium along the U–F side, as indicated by the arrows that point toward each other. This equilibrium is stable, barring invasion from Cheating players, as the dynamics of U–F interactions automatically correct for any disruption that might move the population mix away from it along the side (see Appendix A, available online at the URL in footnote 4, for conditions supporting a U–F equilibrium). In this case, being punished and punishing both are costly. The intuition is straightforward. Unpleasant types will either experience punishment or dole it out whenever they meet others of their own type; they are better off in encounters with Fair players, whom they cheat. Fair players do better in encounters with their own type. The mixed ESS looks stable, because as Unpleasant types increase, they reduce their own fitness with each encounter with one another.

There is more to a simplex than its sides. To fully understand the local equilibria along the legs of the triangle in Fig. 1, we need to consider the consequences if a member of the third group were to enter the population. Consider, for example, what would happen if a C player were to invade the U–F equilibrium described above, where the cost of punishing is higher than the cost of being cheated. Because C players dominate both U and F players, as indicated by the arrows pointing to the C vertex along the bottom and right-hand sides of the simplex, the U–F equilibrium is highly susceptible to invasion by Cheaters, and any such invasion would lead to the ESS at C.

In circumstances like those shown in Fig. 1, where $\beta < \phi$, there can be no internal (polymorphic) equilibrium. More generally, to ascertain whether a global internal equilibrium exists we evaluate local equilibria in combination with winding numbers (Milnor 1997; the discussion below is adapted from Saari 2002). First, the arrows outside the triangle identify all known local equilibria. An equilibrium exists where no two arrows on a single axis point the same direction (cf. Saari 2002). Arrows that point away from an equilibrium indicate that a slight perturbation along its axis would upset the equilibrium; arrows that point toward an equilibrium indicate that equilibrium would be reestablished after such a perturbation. Thus, the U–F equilibrium is stable with respect to perturbations that affect the mix of U and F players, but not with respect to invasions by C players; the C equilibrium resists invasion by either or both U and F players.

[5] There are conditions, described in Appendix A, under which U players will eliminate F players, but the dynamics are virtually identical to the dynamics along the C–F or C–U sides of the simplex depicted in Fig. 1. We focus here on the more analytically and substantively interesting situations, in which Fair and Unpleasant players can coexist.

Winding numbers are a mathematical concept that we can use to leverage information about local equilibria to learn something about global equilibria. Formally, a winding number is a mapping from a circle to a circle. Intuitively, it represents the number of times a path goes around a fixed reference point. In the child's playground game of tetherball, for example, two players try to wind a cord around a post by hitting a ball attached to the cord in opposite directions. The winding number counts the number of times the cord goes around the post at the end of the game, even though during the game it might have wrapped and unwrapped numerous times in both directions.

In the population simplex depicted in the figure, the information is portrayed as a triangle, but it can be deformed into a circle. The arrows identify local (side) equilibria, which, to continue the tetherball metaphor, disrupt the progress of the ball as it winds the cord around the post. The winding number identifies the winner of the game by indicating whether the cord is wrapped around the post, how many times, and in what direction, even if the cord itself is hopelessly tangled. If the winding number does not match the sum of the equilibrium indices for the side equilibria then there must be at least one equilibrium somewhere in the interior. That mathematical result provides purchase for modeling qualitative social phenomena (Saari 2002). We therefore use this mechanism, instead of more complex dynamical systems mechanisms, to give us an idea of what interior equilibria exist and what types they may be.

To compute winding numbers for the population simplex in Fig. 1, begin at the point labeled "start" and count the number of full revolutions indicated by the arrows as you move counterclockwise around the triangle. Each counterclockwise revolution is worth 1, and each clockwise revolution is worth -1. Beginning at "start" in Fig. 1 trace: (a) a 120° counterclockwise rotation; (b) a second, 120° counterclockwise rotation; (c) a 60° counterclockwise rotation; (d) a 120° counterclockwise rotation; (e) a 90° clockwise rotation, followed by (f) a second 90° clockwise rotation; (g) a 120° counterclockwise rotation (erasing the clockwise gain and yielding a cumulative total of one full counterclockwise rotation); (h) a 60° counterclockwise rotation; (i) a 120° counterclockwise rotation; (j) a 120° counterclockwise rotation followed by (k) a 60° counterclockwise rotation, for a second full counterclockwise rotation. The winding number (the number of counterclockwise rotations minus the number of clockwise rotations) for the triangle in Fig. 1 thus is 2.

The winding number must equal the sum of local equilibria indices, where an equilibrium index is the product of the signs of the two pairs of arrows at each equilibrium. Arrow pairs pointing toward the equilibrium take a negative sign, and arrow pairs pointing away from the equilibrium take a positive sign. The indices for the equilibria in Fig. 1 thus are, beginning with the lower-right vertex of the triangle and moving counterclockwise, as follows: $(-1 \times (-1)) = 1$; $(1 \times 1) = 1$; $(1 \times (-1)) = -1$; $(1 \times 1) = 1$. The sum of equilibrium indices for the simplex in Fig. 1 is 2, equal to the figure's winding number. We therefore know that there can be no internal equilibrium in the situation depicted, where the cost of punishing is greater than the cost of being cheated.[6]

Slight alterations in parameters could occasion significant changes in system dynamics. If the punishment to eliminate a small invasion of Cheaters in a primarily U–F population were costly enough, for instance, then the arrows on the U–F side would point inward. The winding number then would be 3, and the sum of equilibrium indices 4, implying an internal equilibrium of -1. This equilibrium has two arrows pointing out and two pointing in (corresponding to the eigenvectors). This new equilibrium divides the simplex into two

[6]This result is for the simplest model allowing these boundary conditions. Adding an equilibrium of degree 1 and another of degree -1 would yield an internal equilibrium.

regions. To the left of lines drawn from the U and F vertices to an internal equilibrium, any mix of the three population types would evolve to the (stable) equilibrium on the U–F side; on the other side, any mix of the three types would collapse to the pure C vertex. In this case, then, if C players can be eliminated through punishment when they are few (i.e., when the problem is emerging, but small), the reward is a nicer society. If punishment is insufficient to achieve this, or if the number of C invaders is large, then as Fig. 1 shows, society devolves to a state of (Cheating) nature.

It makes sense that Unpleasant types cannot survive when the cost of punishing is high relative to the cost of being cheated. If the cost of punishing is lower than the cost of being cheated ($\beta > \phi$), by contrast, Unpleasant players can and do survive, and a population mix of all three player types is possible. We begin, as we did in our discussion of the opposite case above, by analyzing the dynamics in populations of two player types.

2.4 When punishing less costly than being cheated

Changing the relationship between the cost of punishing and the cost of cheating is irrelevant if U players are not in the mix. The dynamics along the bottom of the simplex thus are the same as in the case illustrated in Fig. 1: Cheaters eliminate F types. Similarly, a mix between U and F players (barring, as above, invasion by Cheaters) can result in a stable mix of the two types (see Appendix A). In contrast to the case where $\beta < \phi$, however, Cheaters cannot survive in a two-type mix with Unpleasant players.

Figure 2 shows the side (two-player) dynamics when being cheated is worse than punishing. As indicated by the arrow pairs that point away from each other, on no side is there an ESS that is resistant to invasion. A population made up solely of F types, for example, could easily be invaded by either Cheaters or U types, while a population only of U types is vulnerable to invasion by F players, and a pure C population would be eradicated by contact with U players. There is a two-player equilibrium along the U–F side of the simplex, but, as in the previous case, it is vulnerable to invasion by Cheaters.

What are the conditions that define a mix of Fair and Unpleasant players? Obviously, the population proportions of both types must be greater than 0; specifically, if p is the proportion of U players in the population, $p = \frac{\alpha-1}{\beta+(2-\phi-\gamma)/2-1}$. The numerator is negative (by construction), so $p > 0$ and a U–F mix can exist iff $\beta + \frac{2-\phi-\gamma}{2} - 1 < 0$ (that is, $1 - \phi + 1 - \gamma < 1 - \beta + 1 - \beta$). Since $\beta > \phi$ and $\gamma > \beta$, the inequality reduces to $\beta - \phi < \gamma - \beta$. In other words, a U–F mix is possible if the difference between the cost of being cheated and the cost of punishing is smaller than the difference between the cost of being punished and the cost of being cheated—that is, as long as punishing or being punished is relatively costly. Otherwise, Unpleasant types would overrun Fair players.

For the U–F mix to be an ESS, the proportion of F types in the population $(1 - p)$ must be positive, so

$$\frac{2 - 2\alpha + 2\beta - \gamma - \phi}{2\beta - \gamma - \phi} > 0.$$

In line with the assumptions for $p > 0$, the denominator is negative, so the numerator also must be negative. Simplifying the numerator, we see that $1 - p > 0$ as long as $2(1 - \alpha) + \beta - \phi < \gamma - \beta$. An ESS consisting of Fair and Unpleasant types therefore can exist only if the difference between the cost of being punished and the cost of being cheated is greater than the sum of twice the difference between the Unpleasant and Fair player's payoffs and the difference between the cost of being cheated and the cost of punishment. This is less complicated than it seems: Fair types can survive in a world of U types as long

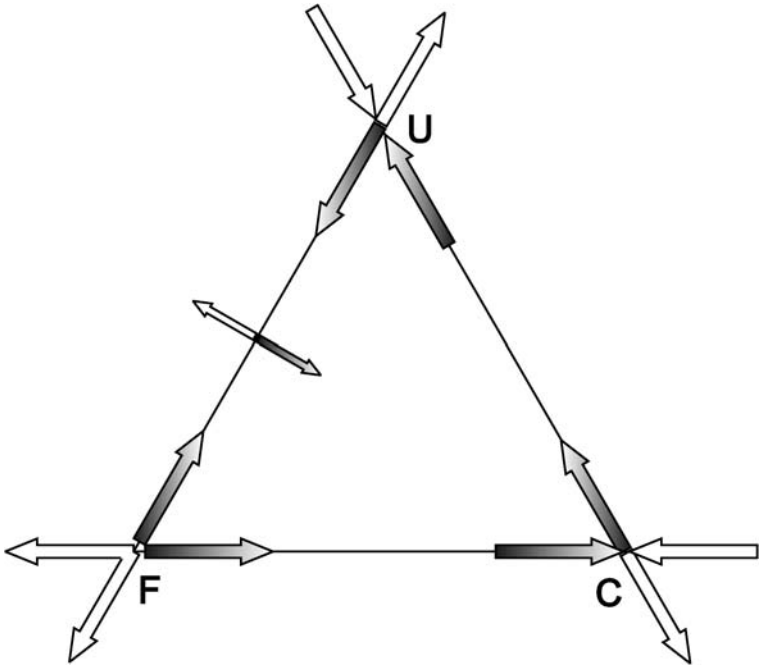

Fig. 2 Population simplex for fair, cheating, and unpleasant types ($\beta > \phi$)

as punishments imposed on cheating U players are much more onerous than the pain they inflict by cheating.

As noted, the U–F mix is stable only as long as there are no Cheaters to invade it. To understand the global dynamics when $\beta > \phi$, we need to know what would happen if a Cheater were to enter the population. Given the conditions for the local ESS, and since a Cheater does better against an F player than against a U player, a sole Cheater would thrive if the proportion of F players were greater than the proportion of U players, but not if $p > 1 - p$, which requires that

$$\frac{\alpha - 1}{\beta + (2 - \phi - \gamma)/2 - 1} > \frac{2 - 2\alpha + 2\beta - \gamma - \phi}{2\beta - \gamma - \phi} > 0.$$

This condition holds only if $(2\beta - \gamma - \phi)(-4 + 4\alpha - 2\beta + \gamma + \phi) > 0$. The second term on the left side of the inequality is negative, however, which implies that there are more Fair than Unpleasant players in a mixed U–F population. The prospects for survival of a lone Cheater therefore are good, as indicated graphically in Fig. 2 by the arrow pointing away from the U–F equilibrium.

Clearly, something interesting could be happening in the interior of the simplex. To understand the interior dynamics, we again need to examine winding numbers and equilibrium indices. The winding number for the case illustrated in Fig. 2 is -1, and the sum of local indices is -2. We thus know that there must be an internal equilibrium with a positive sign—all arrows pointing inward (an attractor), or all arrows pointing outward (a repellor). If the equilibrium is a repellor, as illustrated in Fig. 3, the dynamic of the game moves from the side equilibrium toward the C vertex, from there toward the U vertex, and around again. The population mix will cycle around the repellor point in the top portion of the simplex. If

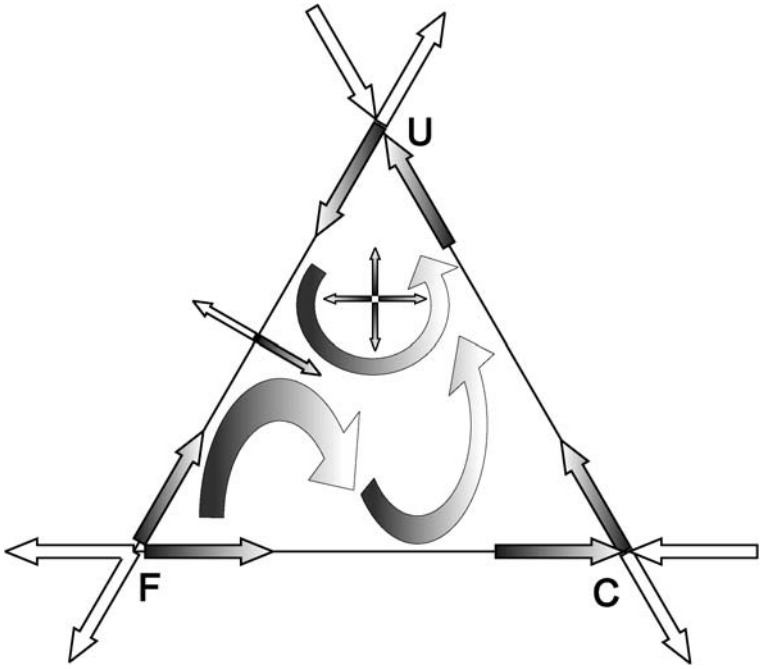

Fig. 3 Internal dynamics when $\beta > \phi$

the internal equilibrium is an attractor, by contrast, any mix of the three types would evolve to the internal equilibrium. In the former case, society would consist of a mix of all three types, but their relative population proportions would fluctuate. In the latter case, population proportions would settle into a stable equilibrium.

The lesson to be drawn from Fig. 3 extends the basic conclusion of most studies of the role of punishment in supporting cooperation. Cooperation can survive in the face of cheating as long as some people are willing to punish cheating. We part company with other scholars, however, in our counterintuitive finding that punishers need not value or desire cooperation *per se*. Indeed, our punishers are truly unpleasant types whose existence nonetheless underpins the survival of cooperation in an imperfect world. In short, the "unpleasantness factor" in human interaction is functional for society.

3 Conclusion

That people do cooperate is not in contention. *Homo sapiens* probably have had to cooperate in groups in order to survive throughout human history. It is not surprising, then, that suggestions that people might not want to cooperate are seen as artificial. The Prisoner's Dilemma is, after all, a contrived situation designed to eliminate cooperation. The evidence from ultimatum- and dictator-game experiments (e.g., Andreoni et al. 2002; Camerer and Fehr 2006; Ensminger 2001; Henrich 2000; Mace 2000; Nowak et al. 2000; Fehr and Fischbacher 2004; for useful reviews, see Fehr and Fischbacher 2003; Güth and Tietz 1990) that people tend to offer even or near-even splits of available money, and that those who are offered something less than what they see as fair are likely to refuse the split

("offers below 20% are almost always rejected", Mace 2000) is unequivocal; the common (albeit often implicit) inference that those who refuse "unfair" offers, even at a personal cost, do so because they value cooperation is rather more dubious.

Scholars generally accept that individuals often face temptations to cheat against each other and against society. Even so, people seem to be upset by the thought that cheating might be irresistible. Anyone who has taught the Prisoner's Dilemma in an introductory Economics or Political Science class knows that many students find it difficult to accept that people who share preferences, interests, and fates would sell each other out. Underlying most treatments of the ultimatum game and other games like it is the presumption that most people *want to cooperate* as long as they can reasonably be confident that others will cooperate as well. They are "wary cooperators" who, if cooperation from others is not forthcoming, "cease cooperating and look for avenues to punish noncooperators even if punishment is personally costly" (Hibbing and Alford 2004; see also, e.g., Boyd and Richerson 1992; Dawes et al. 2007; Fehr et al. 2002).

In our view, to assume that the people doing the punishing want to cooperate is to ignore an important part of the puzzle. Cooperation is a common-pool resource (Hardin 1968 and Ostrom 1990), and people who are willing to punish free riders *and not yield to the temptation to free ride themselves* are prima facie irrational, because they both punish and do not free ride. This seems a weak foundation for theory. We start from the premise that people who punish cheating do so because they are unpleasant, not good. They punish, but they also cheat. They punish not because they value cooperation—although each would benefit most if she were in a position to cheat while everyone else cooperated—but rather because they get annoyed when other people cheat.

Acknowledgements We thank Massimiliano Landi, Michael McDonald, Scott Page, Adriana Stoian, Julie VanDusky, and two anonymous reviewers for helpful comments and suggestions. We accept full responsibility for all errors or omissions, although each of us reserves the right to blame the other whenever possible.

References

Alesina, A., & La Ferrara, E. (2000). Participation in heterogeneous communities. *Quarterly Journal of Economics*, *115*(3), 847–904.

Alesina, A. & La Ferrara, E. (2002). Who trusts others? *Journal of Public Economics*, *85*(2), 207–234.

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2002). The carrot or the stick: rewards, punishments, and cooperation. Working paper. University of Oregon, Economics Department.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Barr, A. (1999) Familiarity and trust: an experimental investigation. Working paper series, Center for the Study of African Economies.

Bewley, T. (2003). Fairs fair. *Nature*, *422*, 125–126.

Bowles, S., & Gintis, H. (2002). Homo reciprocans. *Nature*, *415*, 125–128.

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*(3), 171–195.

Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. (2003). The evolution of altruistic punishment. *Proceedings of the National Academy of Sciences*, *100*(6), 3531–3535.

Brandts, J., Saijo, T., & Schram, A. (2004). How universal is behavior? A four country comparison of spite and cooperation in voluntary contribution mechanisms. *Public Choice*, *119*(3–4), 381–424.

Camerer, C., & Fehr, E. (2006). When does economic man dominate social behavior? *Science*, *311*, 47–52.

Coleman, J. (1990). *Foundations of social theory*. Cambridge: Harvard University Press.

Cook, K. S., Hardin, R., & Levi, M. (2005). *Cooperation without trust?* New York: Russell Sage Foundation.

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796.

Ensminger, J. (2001). Market integration and fairness: evidence from ultimatum, dictator, and public goods experiments in East Africa. California Institute of Technology, Pasadena, CA.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791.
Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.
Fehr, E., & Henrich, J. (2003). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In P. Hammerstein (Ed.), *Genetic and cultural evolution of cooperation* (pp. 55–82). Cambridge: MIT Press.
Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137–140.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868.
Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1–25.
Flack, J. C., Krakauer, D. C., & de Waal, F. B. M. (2005a). Robustness mechanisms in primate societies: a perturbation study. *Proceedings of the Royal Society B*, *272*(1568), 1091–1099.
Flack, J. C., de Waal, F. B. M., & Krakauer, D. C. (2005b). Social structure, robustness, and policing cost in a cognitively sophisticated species. *The American Naturalist*, *165*, E126–E139.
Flack, J. C., Girvan, M., de Waal, F. B. M., & Krakauer, D. C. (2006). Policing stabilizes construction of social niches in primates. *Nature*, *439*, 426–429.
Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, *102*(19), 7047–7049.
Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: a survey and comparison of experimental results. *Journal of Economic Psychology*, *11*, 417–449.
Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243–1248.
Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *The American Economic Review*, *90*(4), 973–979.
Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*(1), 3–35.
Henrich, J., & Boyd, R. (2001). Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*(1), 79–89.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, *91*(2), 73–78.
Hibbing, J. R., & Alford, J. R. (2004). Accepting authoritative decisions: humans as wary cooperators. *American Journal of Political Science*, *48*(1), 62–76.
Levi, M. (1989). *Of rule and revenue*. Berkeley: University of California Press.
Mace, R. (2000). Fair game. *Nature*, *406*, 248–249.
Milnor, J. W. (1997). *Topology from a differentiable viewpoint*. Princeton: Princeton University Press.
Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*, 1773–1775.
Ockenfels, A., & Weimann, J. (1999). Types and patterns: an experimental East–West–German comparison of cooperation and solidarity. *Journal of Public Economics*, *71*(2), 275–287.
Olson, M. (1965). *The logic of collective action: public goods and the theory of groups*. Cambridge: Harvard University Press.
Ostrom, E. (1990). *Governing the commons: the evolution of institutions for collective action*. Cambridge: Cambridge University Press.
Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *The American Political Science Review*, *86*(2), 404–417.
Pulkkinen, O. (2007). Emergence of cooperation and systems intelligence. In R. P. Hämäläinen & E. Saarinen (Eds.), *Systems intelligence in leadership and everyday life* (pp. 251–266). Espoo: Systems Analysis Laboratory, Helsinki University of Technology.
Rubin, P. H. (2002). *Darwinian politics: the evolutionary origin of freedom*. New Brunswick/London: Rutgers University Press.
Saari, D. G. (2002). Mathematical social sciences; An Oxymoron? *PIMS distinguished chair lecture* (Vol. 2006). Pacific Institute for Mathematical Sciences.
Saari-Sieberg, K. (1998). *Rational violence: an analysis of corruption*. New York: Department of Political Science, New York University.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755–1758.

Selten, R. (1978). The chain store paradox. *Theory and Decision*, *9*(2), 127–159.
Sethi, R., & Somanathan, E. (1996). The evolution of social norms in common property resource use. *The American Economic Review*, *86*(4), 766–788.
Sieberg, K. K. (2005). *Criminal dilemmas: understanding and preventing crime*. Berlin/Heidelberg: Springer.
Sigmund, K., & Nowak, M. A. (2000). Evolution and social science: enhanced: a tale of two selves. *Science*, *290*(5493), 949–950.
Skyrms, B. (2003). *The stag hunt and the evolution of social structure*. Cambridge: Cambridge University Press.
Vogel, G. (2004). The evolution of the golden rule. *Science*, *303*, 1128–1131.
Wedekind, C. (1998). Game theory: give and ye shall be recognized. *Science*, *280*(5372), 2070–2071.