CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Research Papers in Economics

Vengefulness Evolves in Small Groups

Daniel Friedman and Nirvikar Singh

Department of Economics

University of California, Santa Cruz

September 2003

Abstract

We argue that evolution will tend to erode vengefulness as a preference trait, due to the costs associated with its exercise. We identify two problems: threshold and mimicry. The first asks how vengeance can evolve from low values where it has a negative fitness gradient. The second asks why cheap imitators do not evolve who look like highly vengeful types but do not bear the costs of actually wreaking vengeance. We discuss how these problems may be overcome in small group interactions where encounters with outsiders are also important. We analyze the role of within-group social norms in overcoming these evolutionary problems.

Acknowledgements

We have benefited greatly from the comments, on our related work, of Ted Bergstrom, Robert Boyd, Bryan Ellickson, Jack Hirshleifer, Steffen Huck, Peter Richerson, and Donald Wittman. This paper draws on much of our previous work on this topic.

1. The Evolutionary Puzzle of Vengefulness

Vengefulness is a powerful human motive: when some culprit harms you or your loved ones, you may choose to incur a substantial personal cost to harm him in return. There can be major economic and social consequences, positive and negative. A taste for vengeance, the desire to ”get even,” is so much a part of daily life (and the evening news) that it is easy to miss the evolutionary puzzle. We shall argue that indulging your taste for vengeance in general reduces your material payoff or fitness. Absent countervailing forces, the meek (less vengeful people) should have inherited the earth long ago, because they had higher fitness. Why then does vengeance persist?

Figure 1 elucidates the fundamental social dilemma in terms of net material benefit ($x > 0$) or cost ($x < 0$) to “Self” and benefit or cost ($y > 0$ or < 0) to counterparties, denoted “Other”. Economists think most often about the mutual gains quadrant I, where actions simultaneously benefit Self and Other. Such symbiotic actions increase social efficiency.[1]

Social dilemmas arise from the fact that evolution directly supports behavior that benefits Self, i.e., outcomes $x > 0$ in quadrants IV (or I) but not $x < 0$ in II (or III), while in contrast, efficiency requires outcomes above the diagonal [$x + y = 0$]. Social creatures (such as humans) thrive on cooperation, by which we mean devices that support outcomes in II+ and discourage outcomes in IV-. Such devices somehow internalize Other’s costs and benefits.

Quadrant IV is the well-studied opportunistic region, where Self benefits at Other’s expense; the biological terms are parasitism and predation. The flip side is the altruism quadrant II, where Self bears a personal cost in order to benefit Other. Quadrant III is anomalous. In fact, Cipolla (1976) refers to behavior producing outcomes in Quadrant III as “stupidity.” Behavior producing such outcomes harms both Self and Other, contrary to efficiency as well as self-interest. How can it persist? We shall argue that the threat of visits to quadrant III (wreaking vengeance) helps discipline opportunistic behavior and helps encourage efficient altruism. But first we should point out that there are several devices.

[1] For simplicity we neglect here possible effects on third parties such as customers of a cartel. Extensions of the present diagram could replace “other” by “average of everyone else affected” or could look explicitly at all affected types.

2. Genetics or Repeated Interactions as Explanations

Biologists emphasize the device of genetic relatedness. If Other is related to Self to degree $r>0$, then a positive fraction other's payoffs are internalized via "inclusive fitness" (Hamilton, 1964) and evolution favors outcomes above the line $[x + \ ry = 0]$. For example, the unusual genetics of insect order *hymenoptera* lead to $r=3/4$ between sisters, so it is no surprise that most social insects (including ants and bees) belong to this order and that the workers are sisters. For humans and most other species, r is only ½ for full siblings and for parent and child, is 1/8 for first cousins, and goes to zero exponentially for more distant relations. On average r is rather small in human interactions, as in the steep dashed line in Figure 1, since we typically have only a few children but work and live in groups with dozens of individuals. Clearly non-genetic devices are needed to support human social behavior.

Economists emphasize devices based on repeated interaction, as in the "folk theorem" (Fudenberg and Maskin, 1986; Sethi and Somanathan, 2003). Suppose that Other returns the benefit ("positive reciprocity") with probability and delay summarized in discount factor $\delta \in [0, 1)$. Then that fraction of other's payoffs are internalized (Trivers, 1971) and evolution favors behavior producing outcomes above the line $[x + \ \delta y = 0]$.[2] This device can support a large portion of socially efficient behavior when δ is close to 1, i.e., when interactions between two individuals are symmetric, predictable and frequent. But humans specialize in exploiting once-off opportunities with a variety of different partners, and here δ is small, as in the same steep dashed line. Other devices are needed to explain such behavior.

A variation on the repeated interaction scenario is one where cooperative acts are credibly communicated to others, who are then more likely to be cooperative in interactions with the first individual. This version is referred to as "indirect reciprocity" (e.g., Fehr and Henrich, 2003), and has been discussed or modeled by Alexander (1987) and Nowak and Sigmund (1997), for example.

[2] Another way to think about it is that with positive reciprocity (or genetic relatedness) one takes a weighted average of the first outcome (in II+ or IV+) and the reciprocal outcome (reflected through the 45 degree line, as self and other are interchanged, so now in IV+ or II+). This gives an outcome in the mutual gains quadrant I if the weight δ (or r) on the reciprocal outcome is sufficiently large.

3. Other Regarding Preferences and Indirect Evolution

Here we will emphasize devices based on other-regarding preferences. For example, suppose Self gets a utility increment of ry. Hence Self partially internalizes the material externality, and undertakes behavior that is above the line [$x + ry = 0$]. Friendly preferences, $r \in [0, 1]$, thus can explain the same range of behavior as genetic relatedness and repeated interaction.[3] However, by itself the friendly preference device is evolutionarily unstable: those with lower positive r will tend to make more personally advantageous choices, gain higher material payoff (or fitness), and displace the more friendly types. Friendly preferences therefore require the support of other devices.

Vengeful preferences rescue friendly preferences.[4] Self's material incentive to reduce r disappears when others base their values of r on Self's previous behavior and employ $r < 0$ if Self is insufficiently friendly. Such visits to quadrant III will reduce the fitness of less friendly behavior and thus boost friendly behavior. But visits to quadrant III are also costly to the avenger, so less vengeful preferences seem fitter. What then supports vengeful preferences: who guards the guardians? This is the central question in the present paper.

In answering this question, our analysis must pass the following theoretical test: people with the hypothesized preferences receive at least as much material payoff (or evolutionary fitness) as people with alternative preferences. Otherwise, the hypothesized preferences would disappear over time, or would never appear in the first place. In a seminal piece, G¨uth and Yaari, (1992) described this test as indirect evolution, because evolution operates on preference parameters that determine behavior rather than operating directly on behavior. Precursors of this idea include Becker (1976) and Rubin and Paul (1979), but it is subsequently to Guth and Yaari's work that the literature has exploded, including papers such as Huck and Oechssler (1999), Dekel, Ely and Yilankaya (1998), Ely and Yilankaya (2001), Kockesen, Ok and Sethi

[3] Indeed, in principle we could have $r>1$ and explain inefficient altruistic behavior. The golden rule ("love thy neighbor as thyself") value $r=1$ seems to be a practical upper bound, however, since no evolutionary devices that we know of will tend to push it higher.

[4] Vengeful and friendly preferences are both examples of what is often termed "strong reciprocity". Other authors have tackled the issue of their evolutionary viability, and provided various answers. Henrich and Boyd (2001) use the assumption of small amounts of conformist transmission. Gintis (2000) focuses on group extinction threats. In his model, strong reciprocity is favored in between-group selection, since it increases group survival chances. Still other approaches are possible, e.g., Bowles and Gintis (2001) and Sethi and Somanathan (2001).

(2000), Possajennikov (2002a, 2002b), and Samuelson and Swinkels (2001). Many of these papers focus on positive reciprocity rather than negative reciprocity, or vengeance.

4. Modeling Other Regarding Preferences

Two main approaches can be distinguished in the recent literature. The distributional preferences approach is exemplified in the Fehr and Schmidt (1999) inequality aversion model, the Bolton and Ockenfels (1999) mean preferring model, and the Charness and Rabin (1999) social maximin model. These models begin with a standard selfish utility function and add additional terms capturing self's response to how own payoff compares to other's payoffs. In Fehr-Schmidt, for example, my utility decreases (increases) linearly in your payoff when it is above (below) my own payoff.

The other main approach is to model reciprocal preferences directly. Building on the Geanakoplos, Pearce and Stacchetti (1989) model of psychological games, Rabin (1993) constructs a model of reciprocation for two player normal form games, extended by Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (1998) to somewhat more general settings. The basic idea is that my preferences regarding your payoff depends on my beliefs about your intentions, e.g., if I believe you tried to increase my payoff then I want to increase yours. Such models are intractable except in the simplest settings. Levine (1998) improves tractability by replacing beliefs about others' intentions by estimates of others' type.

We favor a further simplification. Model reciprocal preferences as state dependent: my attitude towards your payoffs depends on my state of mind, e.g., friendly or vengeful, and your behavior systematically alters my state of mind. This state-dependent other-regarding approach is consistent with Sobel (2000) and is hinted at in some other papers including Charness and Rabin. The approach is quite flexible and tractable, but in general requires a psychological theory of how states of mind change. Fortunately a very simple rule will suffice for present purposes: you become vengeful towards those who betray your trust, and otherwise have standard selfish preferences.[5]

Empirical evidence is now accumulating that compares the various approaches. Cox and Friedman (2002), for example, review about two dozen recent papers. Some authors of the distributional models find evidence favoring their models, but all other authors find evidence mainly favoring state-dependent or reciprocal models. Our own reading of the evidence convinces us to focus on state dependent preferences, while noting that distributional preferences and state dependent preferences can coexist (see footnote 5).

To formalize these ideas, the first step is to model the underlying social dilemma explicitly. In the next section, we provide a particularly stark example, based on a betrayal of trust. Specific numbers are used for payoffs, but this can be generalized straightforwardly.

5. Modeling the Underlying Social Dilemma

Many variants of prisoner's dilemma and public goods games are reasonable choices. For expository purposes we prefer a simple extensive form version of the prisoner's dilemma. It is a convenient vehicle to demonstrate that given a vengeance motive, cooperative behavior is no longer dominated and can become part of a Nash equilibrium even when there is no repeat interaction.

Panel A of Figure 2 presents the underlying game, with payoffs graphed in Figure 1. Player 1 (Self) can opt out (N) and ensure zero payoffs to both players. Alternatively Self can trust (T) player 2 (Other) to cooperate (C), giving both a unit payoff and a social gain of 2. However, Other's payoff is maximized by defecting (D), increasing his payoff to 2 but reducing Self's payoff to –1 and the social gain to 1. The basic game has a unique Nash equilibrium found by backward induction (or iterated dominance): Self chooses N because Other would choose D if given the opportunity, and social gains are zero.

To this underlying game we add a punishment technology and a punishment motive as shown in Panel B. Self now has the last move and can inflict harm (payoff loss) *h* on Other at

[5] It is also possible to incorporate distributional concerns into this state-dependent preferences approach. Thus, my vengeful state may be triggered by being treated unfairly, where unfairness is based on some expectations or norms

personal cost *ch*. The marginal cost parameter *c* captures the technological opportunities for punishing others.

Self's punishment motive is given by state-dependent preferences.[6] If Other chooses D then Self receives a utility bonus of $v \ln h$ (but no fitness bonus) from Other's harm *h*. In other states utility is equal to own payoff. The motivational parameter *v* is subject to evolutionary forces and is intended to capture an individual's temperament, e.g., his susceptibility to anger. See R. Frank (1988) for an extended discussion of such traits. The functional forms for punishment technology and motivation are convenient (we will see shortly that *v* parameterizes the incurred cost) but are not necessary for the main results. The results require only that the chosen harm and incurred cost are increasing in *v* and have adequate range.

Using the notation I_D to indicate the event "Other chooses D," we write Self's utility function in terms of own payoff *x* and the reduction *h* in other's payoff as $U = x + v\, I_D \ln h$. When facing a "culprit" ($I_D = 1$), Self chooses *h* to maximize $U = -1 - ch + v \ln h$. The unique solution of the first order condition is $h^* = v/c$ and the incurred cost is indeed $ch^* = v$. For the moment assume that Other correctly anticipates this choice. Then we obtain the reduced game in Panel C. For selfish preferences ($v = 0$) it coincides with the original version in Panel A with unique Nash equilibrium (N, D) yielding the inefficient outcome (0, 0). For $v > c$, however, the transformed game has a unique Nash equilibrium (T, C) yielding the efficient outcome (1, 1).[7] The threat of vengeance rationalizes Other's cooperation and Self's trust.

that I possess. In our formulation of the social dilemma, developed in the next section, the betrayal of trust is modeled more starkly than "just" unfairness.

[6] Other's utility function here is simply own payoff. If we were focusing on friendliness instead of vengeance, we might write Other's utility function with a positive component for Self's payoff when Self chooses T. This would also lead to an efficient Nash equilibrium if the relevant coefficient *r* exceeds 0.5.

[7] One can also model the situation corresponding to a simultaneous move game. In that case, the game is symmetric, and D is a dominant strategy for each player in the absence of vengeance. With vengefulness, however, for $v > c$, the transformed game no longer has D as a dominant strategy. When population fraction *s* plays C, the expected fitness of C is $W(C) = 1s - (1+v)(1-s)$ and the expected fitness of D is $W(D) = (2 - v/c)s$. The two expressions are equal at $s^* = (1+1/v)/(1+1/c)$. For $s < s^*$ the expected utility is higher for D and we can expect cooperation to disappear as play converges to the inefficient (fitness 0) all-D equilibrium, as in the basic game. But for $s > s^*$ the expected utility is higher for C and we can expect vengeance to drive out defection, resulting in the Pareto efficient all-C equilibrium. Thus for $v > c$ we have a coordination game with two locally stable pure Nash equilibria and an unstable mixed Nash equilibrium at $s^* < 1$. The analysis of the evolution of vengeance is very similar for the simultaneous game model and the extensive form version described in the body of this paper.

6. Evolution of Vengeful Preferences: Threshold and Mimicry Problems

Consider evolution of the vengeance parameter v in an unstructured population. Assume for simplicity that the marginal punishment cost c is constant. Again for simplicity (and perhaps realism) assume that, given the current distribution of v within the population, behavior adjusts rapidly towards Nash equilibrium but that there is at least a little bit of behavioral noise. Noise is present because equilibrium is not quite reached or just because the world is uncertain. For example, Self may intend to choose N but may twist and ankle and find himself depending on Other's cooperative behavior, and Other may intend to choose C but oversleeps or gets tied up in traffic. Such considerations can be summarized in a behavioral noise amplitude $e \geq 0$. Also, Other may imperfectly observe Self's true vengeance level v; say that the observational noise amplitude is $a \geq 0$.

The task before us is to compute Self's (expected) fitness $W(v; a, e)$ for each value of v at the relevant short run equilibrium given the observational and behavioral noise. The function W defines a fitness landscape in which evolution pushes the evolving trait v uphill (Wright, 1949; Eshel, 1983; Kaufman, 1993). We seek long run evolutionary equilibrium where fitness is maximal throughout the population distribution.

First consider the case $a = e = 0$, where v is perfectly observed and behavior is noiseless. Recall from the previous section that in this case the short run equilibrium (N, D) with payoff W=0 prevails for $v<c$, and (T, C) with W=1 prevails for $v<c$. Thus $W(v; 0, 0)$ is the unit step function at $v=c$. We can show (Friedman and Singh, 2003) that with a little behavioral noise (small $e > 0$) the step function slopes down, and with a little observational noise (small $a > 0$) the sharp corners are rounded off, as in Figure 3. As indicated, evolution pushes v downward towards 0 in the subpopulation initially below a level near c-a, and pushes v in the rest of the subpopulation to a level near c+a.

Thus evolution in this case could lead to two types of individuals. One type is just sufficiently vengeful to deter inefficient defection and has fitness $W \approx 1 - 2e$. The other type, recognizably different, is completely unvengeful and therefore unable to support cooperation. It has fitness $W \approx - e$. The unvengeful type can't coexist in evolutionary equilibrium with the more vengeful type because (with small positive e) it has much lower fitness and therefore tends to die out. But there is a serious problem with the more vengeful type: how could it evolve from low values given the negative fitness gradient? We refer to this as the threshold problem, and will outline a solution in the next section.

Putting aside the threshold problem for the moment, the analysis also assumes that Other can observe Self's type quite accurately. But evolutionary forces affect observability, creating what we shall call the Viceroy problem. Butterflies and insect-eating birds play an instructive variant on the vengeance game. A butterfly can hide from birds (analogous to N) or fly about freely (T), and the bird can prey on it (D) or let it alone (C). Monarch butterflies (*Danaus plexippus*) feed on toxic milkweed and so are very unpalatable ($v > c$). Their striking Halloween markings make them easy for birds to avoid as in the efficient deterrence equilibrium (T, C). However, in Santa Cruz and many other areas where Monarchs are common, an unrelated species called the Viceroy (*Limenitis archippus*) has evolved markings that are (from the bird's viewpoint) virtually indistinguishable from the Monarch, a situation that biologists call Batesian mimicry. The Viceroys free ride on the Monarch's high v reputation and are even fitter because they do not bear the dietary cost.

Note that we have not described evolutionary equilibrium in the butterfly-bird game. Although evolution favors population growth of Viceroys when scarce, it does not favor either species once the Viceroys become common. At that point it is worthwhile for hungry birds to sample the butterflies and spit out the unpalatable. An interior equilibrium with both Viceroys and Monarchs is possible if Monarchs can survive being spit out. If Monarchs can't survive the experience, then two other evolutionary equilibria seem plausible: one where the Monarchs migrate ahead of Viceroys so the latter remains relatively scarce, and a second (called Mullerian mimicry) where Viceroys also evolve unpalatability. The field evidence for all three equilibria seems inconclusive (**cites**).

The Viceroy problem surely arises in the vengeance game. An individual with actual v=0 who could convincingly mimic v>c would gain a fitness increment of approximately (1+v)e over the object of his mimicry, and an increment of approximately 1-e over his candid clone. Such increments are irresistible, evolutionarily speaking, so the assumption of near observability (small a) cannot be maintained in evolutionary equilibrium, unless some mechanism to control this phenomenon is available. We discuss possibilities for overcoming Viceroy and threshold problems in the next section.

7. The Role of Group Interactions

We do not know any way to overcome both the threshold problem and the Viceroy problem within the context of unstructured interactions in a large population. Potentially, group interactions can help. Sober and Wilson (1998) argue that group selection can favor prosocial traits, but Richerson and Boyd (1998) point out that genetic group selection in humans is implausible due to relatively rapid cross-group gene flow rates. Henrich and Boyd (2001) argue that the negative gradient aspect of the threshold problem can be overcome within groups if more popular behavior tends to be imitated.[8] They then argue, as we will below with a different approach, that groups that achieve better internal cooperation will displace other groups. But group interactions provide other modeling opportunities that we will now lay out, for overcoming the threshold problem.

The basic theory of repeated games shows how repeat interaction within a small group improves the adaptive value of sub-threshold *v*. Suppose Other expects that he and Self will switch roles from time to time, and that he can expect Self to reciprocate his current choice (C or D) into the indefinite future. Summarizing in the probability and delay of reciprocation in the discount parameter δ, Other compares an immediate payoff of $2\text{-}v/c$ and 0 continuation value if he chooses D, to immediate payoff 1 and continuation value $\delta + \delta^2 + \delta^3 + \dots = \delta/(1\text{-}\delta)$ if he chooses C. Simple calculations reveal that it is advantageous to choose C if $\delta > ½$ in the v=0 case, and if $\delta > (c\text{-}v)/(2c\text{-}v)$ in case of positive *v*. The last expression decreases towards 0 as *v* increases towards *c*. Thus small increments of $v < c$ increase the range of Others who will find it in their interest to play C. This boosts Self's fitness and (depending on the distribution of δ within the group) can more than offset the increment's fitness cost (of order –*e*, as seen earlier.)

[8] The issue in Henrich and Boyd is the same as here, why people would bear the personal cost to punish defectors. The paper notes the game theory device of higher order punishments, e.g., third order is punishing those who don't punish defectors. The modeling goal is to stabilize punishments at finite order, and the key insight is that under reasonable conditions the need (hence cost) for higher order punishment decreases exponentially as the order increases. If conformist transmission has a positive constant impact, then even if it is rather small it can reverse the negative payoff gradient at some sufficiently high order of punishment, and hence stabilize lower orders of punishment and cooperation. This does seem to be a possible solution, but its appeal to an economist is reduced by two considerations. First, if conformist transmission is modeled explicitly, it might be difficult to make it independent of the order. For example, if third order punishments are relevant, an imitator would only rarely observe the difference between his own third order behavior and that of the majority. The transmission rate parameter alpha thus might also decline exponentially in the punishment order and may never reverse the negative payoff gradient. Second, economists tend to think that actual payoffs trump conformity when they point in opposite directions. (Psychologists and other social scientists are unlikely to share this prejudice.)

Repeat interaction can also reduce the marginal cost of punishing culprits within the group. One does not have to retaliate immediately and directly as assumed in Panels B and C of Figure 2. Instead, one can tell other group members about the culprit, and they can choose other partners for mutually productive activities at little or no cost to anyone except the culprit. If so, the effective value of *c* is quite small within the group. (Later we will describe another group punishment technology with even lower cost.) Thus within the group, the threshold is lower and moderate positive values of *v* have positive incremental fitness, and the threshold problem is solved.

At first it seems that similar considerations also solve the Viceroy problem. Given lots of repeat interaction and communication among group members, and a small amount of behavioral noise, a player's true *v* would soon be revealed to his group. Mimicry is not viable in this setting, but reputations are. Thus there are devices for overcoming first order cooperation and second order enforcement problems within the group.

Hence, the real problems arise from players' interactions outside the group. Assume, as might be reasonable, that a typical individual does not have significant repeat interaction with any particular person outside the group, but the interactions with all people outside the group collectively do have a significant effect on her fitness.[9] Assume also that individuals can fairly reliably assess any individual's group affiliation and know the reputation of the group. Then we have a free rider problem with respect to group reputation. Each individual would benefit from using low *v* in interactions outside the group but the group's reputation and hence its members' fitness would suffer. The group must somehow regulate its members' behavior or things will

[9] Across-group encounters are also frequent, but a given individual will encounter a specific non-group member only very sporadically. An individual in such encounters cannot reliably signal her true *v* because outward signs can be mimicked at low cost, but neither (due to the large numbers of sporadic personal encounters) can she easily establish a reputation for her true *v*. A specific assumption that would capture these considerations is that the perceived vengeance parameter of one's opponent v^e is equal to the true value *v* in encounters within the group, but in encounters outside the group $v^e = \lambda \overline{v} + (1-\lambda)E\overline{v} + \varepsilon$, an idiosyncratic error plus the weighted average of the partner's group average $\overline{v}$ and overall population average $E\overline{v}$, with the weight λ on the group average an increasing function of group size. The idea is that v^e is a Bayesian posterior, with sample information on any individual overwhelming priors for internal matches and sample information on the relevant group being important for external matches. Implicit in this formulation is a theory of group size. Very large groups would violate the assumptions that everyone knows everyone well and monitors the all-C equilibrium, so there are diseconomies of scale. At the margin, these diseconomies should balance the economies arising from the dependence of λ on group size. We shall not attempt to develop such a theory here, but simply will assume the existence of moderate size groups.

unravel. We hypothesize that groups themselves possess traits that evolve to solve such problems.

Note that social groups, unlike butterflies, use conscious mechanisms to control mimicry. Gangs may have secret handshakes and other codes of communication, but these are relevant only for identifying membership of the group. In Indian villages, one aspect of enforcing caste distinctions involves codes of dress and bodily decoration, so that lower castes cannot mimic upper castes, in general interactions, including with third parties. In that case, the higher caste is protecting its group reputation. In large anonymous settings such as towns and cities, these codes are harder, if not impossible, to enforce, and mimicry is more common, with lower castes redefining their identities to be able to claim higher caste status.

In general, we may identify three different possible responses to the Viceroy problem. The first, and the one most familiar to economists, would be the use of costly signaling. In the standard signaling model, one type (say, High) has a lower cost of signaling than another type (say, Low), and in a separating equilibrium, the Low type chooses not to mimic the High type. For example, “toughness” may be signaled by acquiring tattoos, which would be too painful for those who are not “tough”. Depending on the parameters of the situation, however, there may also be pooling equilibria, where the two types cannot be distinguished. In the kinds of situations we are interested in (across-group interactions where group reputations matter), signaling might be enforced by the group, when group benefits to signaling exceed individual benefits. As noted, certain kinds of dress codes and bodily decorations may be enforced within groups.

A second possible response to the Viceroy problem is evasion, so that mimicry is avoided by physical separation. This is plausible in the context of migratory butterflies, but it is not clear how relevant it might be for human groups. One might also conceive of evasion and pursuit taking place in the space of characteristics, with the mimicked species or group evolving new traits as the old ones lose their distinctiveness. This would be akin to a dynamic signaling model, where multiple signals are possible: as the signaling characteristics of the Viceroy evolve toward those of the Monarch, the Monarch may evolve new distinguishing markers. Note once more that in the nonhuman species case, the evolution is necessarily through genetic mutation and selection, whereas in the case of human groups, conscious choices are involved, in choosing signal levels – evolution in this latter case would be cultural, and could be the result of learning.

The third and final possibility in responses to the Viceroy problem is that of group enforcement. Here we mean enforcement across groups, rather than within groups, which we discussed in the context of the signaling model. Thus high-caste groups may be willing to incur costs of punishing low-caste groups that try to mimic them in encounters with third parties. The benefits are protection of reputation, and fitness gains associated with that protection. Note that this enforcement also requires overcoming free-rider problems within the group, but, as we have discussed, within-group interactions that are frequent allow repeated game mechanisms to come into play.

8. Group Traits and Individual Fitness

In earlier writings on the subject (Friedman and Singh, 1999, 2001) we have used the word meme, but to some readers that suggests an individual trait transmitted via imitation. A group trait is a characteristic of the group rather than an individual characteristic. Perhaps the sort of group trait most discussed in recent literature is a convention or norm: a Nash equilibrium of a coordination game in which it is in each member's interest to play a certain way given that the other group members are doing so, e.g., observe Sabbath on Saturday. But this is unnecessarily restrictive. Majority rule and primogeniture (or school mascots such as aggies or banana slugs) are group traits that need not be modeled as Nash equilibria of individual behavior. Likewise for group traits such as use of a particular flag design, or language, or (closer to home) peer review protocols or the use of special jargon. Group traits are often discussed in the context of corporate culture and organizational routines (Nelson and Winter, 1982), and recent experiments by Weber and Camerer (2003) empirically bear out the existence of some facets of organizational culture that are created by organization members, but survive changes in individual membership of organizations.[10]

The relevant group traits for the present discussion are prescriptions on how individuals *should* behave in social dilemmas. Such prescriptions, when widely shared by group members,

[10] In these experiments, the relevant dimension of organizational culture is a specialized homemade language developed by organization members to complete a task efficiently. This kind of group trait is not relevant for encounters with outsiders, but only for within-group interactions. Corporate dress codes would matter for outsiders, but are copied very easily. However, it is easy to think of being "hard-nosed" as a corporate trait that might be valuable in dealings with outsiders, difficult to imitate, and enforced by internal norms of status.

are group traits that are logically distinct from, but that co-evolve with, the individual traits that determine actual behavior. For example, the group trait might be the shared belief that the appropriate level of the vengeance parameter is 3 and (as we will see in Section 10) that group trait might be in evolutionary equilibrium with actual behavior governed by the individual trait with a somewhat lower value, say, $v = 2$.

One can imagine several different mechanisms by which group traits affect the fitness of an individual's traits. Perhaps the mechanism most familiar to game theorists is higher order punishment strategies: deviations of actual behavior from prescriptions are punished, as are failures to punish, failures to punish non-punishers, etc., *ad infinitum*. We prefer to emphasize a different mechanism, mediated by status (e.g., Catanzaro, 1992; Nisbett and Cohen, 1996). The mechanism has two parts: (a) the group's traits and the individual's behavior affect status, and (b) status affects fitness.

To elaborate on (a), we recognize that status may depend on individual traits of all sorts, including age, sex, birth order and parental status. In all societies we know about, it also depends on contribution to local public goods. Local public goods include access to resources such as water supplies, sites for shelter and foraging, and military capabilities. Also included are intangibles such as the group's reputation among other groups, and its internal cohesiveness. (**cites**) Adherence to the group's prescribed level of vengefulness v^n contributes to that group's internal cohesiveness and external reputation. Thus it is reasonable to postulate that, other things equal, an individual will have higher status when his behavior reflects v closer to v^n. Such behavior upholds the group's identity; see Akerlof and Kranton (2000).

Part (b) is straightforward. The group allocates many rival resources; depending on the context, these might include marriage partners, home sites, access to fishing holes and plots of land. Status is a device for selecting among the numerous coordination equilibria: the higher status individuals get the first choice on available home sites etc. (**cites**) The model below uses a single parameter t to combine the sensitivity of fitness to status with the sensitivity of status to behavior.

9. Evolution of Group Traits

Several authors recently have discussed the evolution of individual traits whose fitness depends on their prevalence in the group (e.g., Sober and Wilson, 1998) and other authors have discussed the evolution of conventions (**cites**), but our question is a bit different. Unlike individual traits such as *v*, group traits cannot differ across individuals within a group: everyone knows how he is supposed to behave in that group and knows the likely consequences of a deviation. Individuals of various sorts may enter or leave a group, and the group may grow or shrink, but these changes have no direct impact on group traits. Rather, over time a particular group's trait may drift or occasionally change abruptly as the members' common understanding reacts to experience.

A detailed micro-dynamic evolutionary model of a group trait would have to consider the joint time path of the traits across groups and the group sizes. Such detail seems awkward and unnecessary. We need to know which group traits will displace others, but it does not much matter whether the displacement occurs through changes in group size or the numbers of groups. It seems sufficient to use aggregate dynamics that track the population shares for each group trait.

In specifying even aggregate dynamics one must consider a variety of transmission mechanisms for group traits including imitation, proselytization, migration and conquest, as well as fertility and mortality. It is possible for horizontal transmission to increase the share of a group trait that reduces fitness (e.g., encouraging tobacco consumption), but we do not believe that such considerations play a central role for the group traits of present interest. For simplicity we will just hypothesize that the population shares respond positively to the average fitness of its members relative to the overall population average. In the evolutionary games literature for discrete traits, this is usually referred to as monotone dynamics, of which the leading example is replicator dynamics (Weibull, 1995). In the much smaller literature on continuous traits, the relevant dynamics are gradient (e.g., Friedman and Yellin, 1997): the value of the continuous trait increases where the payoff gradient is positive and decreases where it is negative.

Given the underlying social dilemma considered here, the relevant group traits are prescriptions for responding to culprits and cooperators,[11] and for responding to deviations from the first level prescriptions. Prescriptions for all permutations and combinations could be cumbersome, but are mostly irrelevant for present purposes. Given devices discussed earlier that ensure a high degree of cooperation within the group, the relevant group traits can be summarized in two parameters: the prescribed degree of vengefulness v^n towards culprits outside the group, and the tolerance parameter t for dealing with deviations by group members from v^n.

Recall, from the previous section, that deviations $x = v - v^n$ of actual from prescribed behavior are dealt with by reducing status, which leads to an adverse redistribution of resources and reduced fitness for the deviator. We assume simply that the fitness reduction $\rho(x)$ is smooth and convex (i.e., the incremental fitness reduction increases with the magnitude of the deviation) and is minimized with value 0 at $x = 0$. The second order Taylor expansion approximation therefore can be written $\rho(x; t) = x^2/(2t)$, where deviations are treated less harshly the larger is the tolerance parameter $t>0$.

10. Evolutionary Time Scales and Equilibrium

A few remarks may be in order about fitness, monotone dynamics and time scales. The analysis becomes very simple if there is a hierarchy of time scales so only one sort of trait is evolving significantly in any time scale. One can assume that individual levels of v adjust rapidly within the genetically feasible range $[0, v^{max}]$; the idea is that people learn and accommodate themselves to the group's meme within a short period, say weeks or months. For example, according to stories in the media, children raised in Belfast and Beirut brought to the US have no problem adapting with a few months to the US norm and then adapting back when they return. Group traits also adjust, but in the medium run of years to decades. The capacity for vengeful behavior v^{max} can be thought of as mainly genetic and thus it too can adjust in the long run, over several generations.

The dynamics are trivial in this case because in each time scale only a single scalar variable is adjusting, the fitness functions are single peaked, and the direction of change is immediate from the definition of fitness. First, individual values of v converge to the level that maximizes individual fitness given v^n and t. Then v^n adjusts (for t fixed) to the level that maximizes the group average fitness given the error and noise rates and v^{max}; the individual v's trail along with the adjustments in v^n. (To be a bit more sophisticated, one could let t adjust at the

[11] In the case of the Viceroy problem, we can add prescriptions for dealing with mimicry by other groups as another

same time, or separately, and possibly also allow the error and noise rates to evolve.) Finally, if the values of *v* are constrained by v^{max} then it too evolves, with the other variables moving in its wake.

Of course, time scales actually are not so hierarchical, and there may be nontrivial co-evolution of individual *v* (social regulation of emotions), group traits, and emotional capacity. We conjecture that such co-evolution would not affect the relevant evolutionary equilibria nor alter their stability in the present case, although it certainly can in more general settings.

We will now sketch how equilibrium norms of vengeance might evolve in our setting. We use the basic trust game with observational and behavioral errors, as discussed in Sections 5 and 6 (see Figure 2),[12] and we assume that Self and Other belong to different groups. We will assume a simple, two-point distribution of types for Self: they can either have vengeance parameter 0 or $v > 0$. This assumption was discussed in Section 6 (see Figure 3). We will assume that there is a Perfect Bayesian Equilibrium (PBE) that is separating. As shown in Friedman and Singh (2003b), this requires that the proportion of vengeful types of Self that are encountered by Other is neither too small nor too large. The intuition is that if there are too few vengeful types, then Other has an insufficient incentive to ever cooperate, whereas if there are too many vengeful types, either there is a pooling equilibrium with only trust and cooperation, or Other's cooperation gets Self to always trust, but this would lead Other to defect sometimes, so that there is no (pure-strategy) equilibrium.[13]

Focusing on the separating equilibrium, we can summarize the possible payoffs and probabilities in the following table. Note that the probability α combines two error possibilities: an accurate observation followed by a behavioral error, and an observation error followed by intended behavior. To the fitness payoffs in Table 1, we add the consequences of the social norm discussed in Section 8. In particular, if the individual vengeance parameter, *v*, deviates from the group norm, v^n, then the individual suffers a fitness loss, through loss of status, given by $\rho(x; t) = x^2/(2t)$, where $x = v - v^n$. Incorporating this additional term, then, using the payoffs and probabilities in Table 1, the vengeful Self's expected fitness is given by

component of the relevant group traits.

[12] Thus, our assumptions here are an alternative to those discussed in footnote 9.

[13] Details of this logic, as well as the precise bounds on the proportion of vengeful types in the population, are in Friedman and Singh (2003b).

$$W(v;\, v^{n}) = 0.e + 1.(1-e)(1-\alpha) - (1+v).(1-e)\,\alpha + \rho(v - v^{n}).\,(1-e)\,\alpha$$

Table 1: PBE Probabilities

		Fitness Payoff	**Equilibrium Probability**
	Choice	**Self, Other**	**Strategies: (NT, DC)**
$v > 0$	(N, .)	0, 0	e
	(T, C)	1, 1	$(1-e)(1-\alpha)$
	(T, D)	$-(1+v)$, $2-v/c$	$(1-e)\alpha$
$v = 0$	(N, .)	0, 0	$1 - e$
	(T, C)	1, 1	$e\alpha$
	(T, D)	-1, 2	$e(1-\alpha)$

Notes:

Other observes s = 1 with probability a in (0, ½) when $v = 0$, and observes s = 0 with probability a when $v > 0$. Other chooses his less preferred action with probability $\alpha = a(1-e) + e(1-a) = e + a - 2ae$.

Now dynamics in the current setting, which assumes hierarchical time scales, converge the individual's vengeance parameter toward the value that maximizes individual expected fitness. A simple calculation yields the first order condition $\rho'(v - v^{n}) = 1$. This, in turn, for the quadratic case reduces to the condition $v = v^{n} - t$.[14] We see that, in this equilibrium, groups will enjoin an exaggerated version of the optimal v, but the individually optimal v prevails. That optimum is the same as in Fig 3, since group reputations have only small observational error.

Note the comparative statics: the punishment technology for out-group interactions is the relevant c, and the prevailing v tracks the optimum given that value of c. Thus the model implies that easier detection and punishment of culprits will lower people's taste for the amount of

[14] The identical first-order condition and individual equilibrium are derived in Friedman and Singh (1999, 2001). Our earlier results are derived for the simultaneous-move case that was described in footnote 7. There we also allow for somewhat more complex group interactions and perceptions, as well as including a more detailed discussion of reputation and status issues, and of relevant biological constraints on vengeance. In particular, in the simultaneous move game, there is symmetry, and either player can choose to defect. Nevertheless, the outcome in terms of the individual we focus on is similar, and is based on the marginal logic of trading off the cost of individual retaliation and the impact on status within the group.

punishment in long run equilibrium. Also, higher tolerance t in a group correlates with higher v^n, although there is not really a causal relationship either way.

In the analysis to this point, we can assume that everyone in Self's group is identical, so that $v^n - t$ is also the group average vengeance parameter. Examining the dynamics of this group average proceeds as follows. In the case of group average fitness, status losses represented by the function $\rho(x; t)$ are netted out for the group, so that average fitness is

$$W^g(\bar{v}) = 0.e + 1.(1 - e)(1 - \alpha) - (1 + \bar{v}).(1 - e)\,\alpha\ .$$

In our earlier work, we assumed that the probability of facing someone who would defect was described by an encounter function that was decreasing in $\bar{v}$. In the current formulation, we adopt the argument of Friedman and Singh (2003b), that the probability of an observational error is negatively related to the level of the group average vengeance parameter, so that $a = A(\bar{v})$, with a negative first derivative.[15]

Now it is easy to derive the first-order condition for the group optimum, which is denoted v^o. The condition is $A'(v^o)(2 + v^o) + A(v^o) = 0$. If, for example, $A(v) = 0.5\exp(-v/b)$,[16] then it is easy to see that the first order condition reduces to $v^o = b - 2$.[17] Alternative assumptions on $A(v)$ are also possible, such as a Gaussian rather than exponential form: we used this alternative in Friedman and Singh (2003b) to derive a long run version of PBE (focusing in that case on populations in general, rather than interactions among members of small groups). In any case, in the medium run, quality of the individual and group optima will yield an expression for the vengeance norm for the group. In the case of our specific functions used, we have the result that $v^n = t + b - 2$.[18]

We have, in this section, provided some analytical details of a stylized model of the evolution of vengeful preferences that are supported by within-group social norms, but that operate in encounters with outsiders. To some extent this is a partial equilibrium, since we have

[15] Note that this dependence did not come into play at the individual level, since we assume that observability is driven by group traits. The justification here is that more individuals from vengeful groups are more likely to be correctly perceived.

[16] Note that the factor 0.5 ensures that as vengefulness goes to 0, the observation becomes completely noisy, which is as it should be.

[17] In our earlier work, this expression was derived from an 'encounter function' that assumed that the probability of encounters with defecting players from other groups would be decreasing in the individual's own-group vengeance parameter. We see that our alternative assumption, where the probability of correct observation is affected, has a similar implication, as one might expect.

[18] Note that we assume throughout our discussion that these expressions are valid interior solutions.

not worked out how the entire distribution of vengeance parameters over different groups might evolve. Note that various groups may differ in their environments and the frequency of their interactions with each other. For example, pastoral and agricultural groups may end up with different equilibrium levels of vengeance (Nisbett and Cohen, 1996, and references therein).

11. Discussion

In this paper, we have discussed some of the issues that arise in considering the evolutionary viability of the trait of vengefulness. We have offered an explanation that is based on interactions among small groups. These interactions are not frequent enough to support the use of repeated game or related mechanisms for reciprocity. However, they are important enough in the aggregate to affect fitness. We have argued that small groups can enforce relatively costless norms of vengeance, and shown how individual adherence to these norms, while imperfect, can be strong enough in evolutionary equilibrium to sustain cooperative outcomes in inter-group encounters. We have discussed how the theoretical problem of evolving a small degree of vengeance, insufficient for the above equilibrium, can be overcome through its role in within-group encounters.

Our analysis is not the only approach to the problem of explaining vengefulness. Several authors have encountered the viability problem in one form or another, and have found ways to finesse it. As noted earlier, Henrich and Boyd (2001) argue that the negative gradient aspect of the threshold problem can be overcome within groups if more popular behavior tends to be imitated, even when this conformity effect is very weak. As a result, groups with this property achieve better internal cooperation, and displace other groups.

Rosenthal (1996) considers a limited form of vengeance in which a player can detect culprits and shun them after the first encounter. The payoffs of such players (called TBV for "trust but verify") are all reduced by verification costs. Rosenthal begins with a basic stage game like ours and then modifies it by expressing payoffs as present values of the continuing relationship. The harm a TBV player inflicts on a culprit is the present value of payoffs the culprit foregoes after the initial temptation payoff. The punishment cost is the present value of verification less the present value of the avoided (sucker payoff) loss, which for relevant parameter values is negative. Thus punishment brings a net personal *benefit* and the all-C strategy (corresponding to our *v*=0 player) does not weakly dominate the TBV strategy. Rosenthal finds several NE for his 3x3 symmetric game, and all-D need not be the only stable equilibrium. For certain parameter configurations, there is an interior NE that is stable under some (but not all) monotone dynamics. Unfortunately, no such stable equilibrium would exist under our maintained assumption that vengeance is costly and cannot reduce the sting of the sucker payoff.

Huck and Oechssler (1996) deal with the problem in a richer context than ours. In the "ultimatum game" they study, players interact in small groups and have two roles, each played half the time. In one role ("responder") they can pursue a costly vengeance strategy. Since there are only two possible offers, shading of punishments is not possible. With finite populations (or infinite populations interacting in small groups), punishments may increase the individual's relative fitness although it lowers absolute fitness. As the dynamics in their model are solely driven by relative fitness, the vengeful trait survives. However, there is no continuous evolvable trait in their model, which would be analogous to our vengeance parameter, *v*.

The solution we have proposed to the viability problem is related to the two-level model for the evolution of cooperation as exposited in Sober and Wilson (1998) and S. Frank (1998). These authors note that, using a tautology known as the Price equation (G. R. Price, 1970)[19], one can demonstrate the possibility that a socially beneficial but dominated strategy (call it C) might survive in evolutionary equilibrium when group interactions are important. The idea in their analysis is that groups with a high proportion of C players have higher average fitness and thus grow faster than groups with a smaller proportion, and this effect may more than offset C's decline in relative prevalence within each particular group. The necessary conditions for C to survive (it can never eliminate D but may be able to coexist in equilibrium) are rather stringent. Besides the obvious condition that the group effect favoring C must be stronger than the individual effect favoring the dominant strategy D, it must also be the case that the groups dissolve and remix sufficiently often, and that the new groups have sufficiently variable proportions of C and D players. These special conditions may be met for some parasites, but seem quite implausible as a genetic explanation of human cooperation. Indeed, Sober and Wilson

[19] The Price equation uses the definition of covariance to decompose the change in prevalence of a trait into two components, e.g., the direct effect from individual fitness and an indirect effect incorporating the spillovers within the group.

invoke memetic evolution and discuss the importance of cultural norms for rewarding cooperative behavior and punishing uncooperative behavior. They avoid the viability problem by assuming in essence that *c* is 0; see p151 for the most explicit discussion of this point.

Bowles and Gintis (1998) consider the genetic evolution of vengeance in the context of a voluntary contribution game. They assume a direct tie between two discrete traits, a preference for punishing shirkers (analogous to our *v*) and a preference for helping a team of cooperators. Their argument is a version of two-level selection as in Sober and Wilson and again is rather delicate. In an essay on the rise of the nation state in the last millennium, Bowles (1998) uses a version of the same model that allows for cultural and genetic coevolution.

Yet another way to avoid the viability problem is to assume that individuals with higher values of *v* encounter D play less frequently. R. Frank (1987) discusses this possibility informally and formally models the evolution of a visible altruistic (rather than vengeful) trait. It is not hard to show under some specifications of how the frequency of cooperators depends on *v* that there is a positive level of *v* that maximizes fitness. Indeed, if each individual's *v* were observable, then those with higher *v* might encounter D-play less frequently (as in R. Frank's 1988 discussion) and thus maintain equal or higher fitness. This "greenbeard" solution[20] ignores the evolutionary pressure for lower *v* individuals to mimic the visible signs of higher *v*, which we have discussed in the context of the Viceroy problem.

We have offered a somewhat more complex resolution of the viability problem because we believe that the relation between *v* and the frequency of encountering cooperators arises mainly at the group level rather than at the individual level. We have argued that within well-functioning groups, D behavior is rare and dealing with it is not an important source of fitness differences. Presumably D behavior is more frequently encountered with partners outside one's own group, and we believe that here group reputations are the key, not individual signals or individual reputations. We have also suggested how within-group mechanisms might control the Viceroy problem.

[20] This term is due to Dawkins (1976), and is a used as a fanciful but striking example of identifiability. A certain type of individual is identified by their green beards, and somehow no other type of individual is able to mimic them, even when it is strongly in their evolutionary interest.

Our approach has focused most directly on the problem of the evolution and persistence of vengefulness, and we believe that it provides some new insights. Nevertheless, our discussion has finessed many important questions. Here are two methodological questions that we have not addressed in this paper.

- Other-regarding preferences may involve a host of contingencies besides whether Other belongs to Self's own group and whether he is a culprit. What theoretical discipline, as well as empirical evidence, can keep such models sharp and tractable? Intuitively, the requisite preferences must aid fitness in a variety of situations, and the answer to this question may require identifying canonical games that best capture human experience.
- Introducing a group structure on interactions and allowing groups a very low cost punishment strategy creates a huge set of possible evolutionary equilibria, larger even than in the "folk theorem." What selection criteria can be brought to bear on the model to narrow down the set of equilibria? In Friedman and Singh (2003a), we have introduced the concept of Evolutionary Perfect Bayesian Equilibrium, which may offer one approach to answering this question.

Finally, we provide some broader perspective on our approach to modeling the arising and persistence of vengefulness. We have used the existence of well-functioning norms within small groups to support the long-run use of vengeful behavior in across-group interactions. The analogy we can offer is to a trellis or scaffolding, where either structure supports the growth or erection of something else. The difference between a trellis and scaffolding is that the latter is temporary, whereas the former is permanent. In that sense, group traits or norms in our model act as a trellis. Without them, the kind of behavior that we posit would erode, as, over time, individuals would find it beneficial to shade their vengefulness. Some aspects of within group interactions, however, have the characteristics of scaffolding – in particular, in overcoming the threshold problem because a small amount of vengefulness increases the range of discount factors for which cooperation works in repeated settings. Once the threshold is crossed, other factors sustain the level $v > c$. Of course, the repeat interactions can still play a role in enforcing the norms that matter for sustaining vengefulness. We would like to suggest that this perspective, of one set of traits, whether cultural or biological, providing direct support for another trait to develop, is a useful idea in general discussions of coevolution. In particular, distinguishing between trellises and scaffoldings can be helpful in understanding the relationship between present and past.

Bibliography

Abreu, Dilip and Sethi, Rajiv. “Evolutionary Stability in a Reputational Model of Bargaining.” Working Paper, Princeton University and Barnard College, March 2003, http://www.columbia.edu/ rs328/evolution.pdf

Akerlof, G. & Kranton, R. (2000). Economics and identity. *Quarterly Journal of Economics*, **115**, 715-753.

Akerlof, George (1983), *An Economist’s Book of Tales*, NY: Cambridge University Press.

Alchian, Armen. “Uncertainty, Evolution and Economic Theory.” *Journal of Political Economy*. 1950, 58, pp. 211-221.

Alexander, R. D. (1987). *The Biology of Moral Systems*, New York: Aldine de Gruyer.

Andreoni, James and John H. Miller (1996), “Giving According to GARP: An Experimental Study of Rationality and Altruism,” University of Wisconsin working paper.

Becker, Gary S. “Irrational Behavior and Economic Theory.” *Journal of Political Economy*. 1962, 70, pp. 1-13.

Becker, Gary S. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press, 1976.

Bergstrom, Theodore C. “Evolution of Social Behavior: Individual and Group Selection.” Journal of Economic Perspectives, 2002, 16:2, pp. 67-88.

Binmore, K. & Samuelson, L. (1999). Evolutionary drift and equilibrium selection. *Review of Economic Studies*, **66,** 363-393.

Black-Michaud, J. (1975). *Cohesive Force: Feud in the Mediterranean*, Oxford: Basil Blackwell.

Blackmore, S. (1999). *The Meme Machine*, Oxford: Oxford University Press.

Blackmore, S. (2000). The power of memes. *Scientific American*, October, 64-73.

Bolton, Gary E. and Ockenfels, Axel. “ERC: A Theory of Equity, Reciprocity and Competition.” *American Economic Review*, March 2000, 90(1), pp. 166-93

Bowles, S. & Gintis, H. (1998). The evolution of strong reciprocity. University of Massachusetts, Amherst working paper.

Bowles, S. (1998). Cultural group selection and human social structure: the effects of segmentation, egalitarianism and conformism. University of Massachusetts, Amherst working paper.

Boyd R. & Richerson P. J. (1990). Group selection among alternative evolutionarily stable strategies. *Journal of Theoretical Biology*, **145**, 331-342.

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

Boyd R. & Richerson P. J. (1998). The evolution of human ultra-sociality. In: Eibl-Eibisfeldt I, Salter F. K. (Eds.) *Indoctrinability, Ideology and Warfare: Evolutionary Perspectives*. Berghahn Books, New York

Boyd, R. & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, R., Bowles S., Gintis, H. & Richerson P. (2001) The evolution of punishment of noncooperators through intergroup conflict, manuscript presented at Santa Fe Institute workshop, January.

Catanzaro, R. (1992). *Men of Respect: A Social History of the Sicilian Mafia.* New York: The Free Press.

Charness, Gary and Rabin, Matthew. “Social Preferences: Some Simple Tests and a New Model.” Discussion paper, University of California at Berkeley, 2001.

Cipolla, Carlo. The Basic Laws of Human Stupidity. Bologna: The Mad Millers, 1976.

Cox, James C. and Friedman, Daniel. “A Tractable Model of Reciprocity and Fairness.” Manuscript, University of California at Santa Cruz, 2002.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*, 2 vols. New York: Appleton.

David K. Levine (1998) “Modeling Altruism and Spitefulness in Experiments,” *Review of Economic Dynamics* 1, 593-622.

Davis, J. (1980). *Antropologia della Societa Mediterranee: Un’analisi Comparata*, Turin: Rosenberg & Sellier.

Dawkins, R. (1976). *The Selfish Gene.* New York: Oxford University Press.

Dawkins, R. (1982). *The Extended Phenotype: The Gene as the Unit of Selection*. San Francisco: Freeman.

Dekel, Eddie, Ely, Jeffrey C. and Yilankaya, Okan. “The Evolution of Preferences.” Working Paper, Northwestern University (1998) http://www.kellogg.nwu.edu/research/math/Je_Ely/working/observe.pdf

Dufwenberg, Martin and Kirchsteiger, Georg. “A Theory of Sequential Reciprocity.” Discussion paper, CentER for Economic Research, Tilburg University, 1999.

Dupré, J. (1987). *The Latest on the Best: Essays on Evolution and Optimality*. Cambridge, MA: MIT Press.

Durham, W. H. (1991). *Coevolution: Genes, Culture, and Human Diversity*. Stanford, CA: Stanford University Press.

Durham, William H. (1991), *Coevolution : genes, culture, and human diversity*, Stanford, Calif.: Stanford University Press.

Ely, Jeffrey C. and Yilankaya, Okan. "Nash Equilibrium and the Evolution of Preferences." Journal of Economic Theory, 97, pp. 255-272, 2001.

Eshel, I. (1983). Evolutionary and continuous stability. *Journal of Theoretical Biology*, **103**, 99-111.

Falk, Armin and Fischbacher, Urs. "Distributional Consequences and Intentions in a Model of Reciprocity." Annales d'Economique et de Statistique, 63-64 (Special Issue), July-December 2001.

Farb, P. (1978). *Man's Rise to Civilization: The Cultural Ascent of the Indians of North America*, New York: Penguin.

Fehr, E. & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, **14, 3**, 159-182.

Fehr, Ernst and Schmidt, Klaus M. "A Theory of Fairness, Competition, and Cooperation." Quarterly Journal of Economics, August 1999, 114(3), pp. 817-68.

Fehr, Ernst and Simon Gaechter(1998), "Cooperation and Punishment," University of Zurich manuscript, September.

Fehr, E. and Henrich, J. (2003), Is Strong Reciprocity a Maladaptation? Forthcoming in *Genetic and Culture Evolution of Cooperation* edited by Peter Hammerstein. MIT Press.

Frank, R. (1987). If *Homo Economicus* could choose his own utility function, would he want one with a conscience? *American Economic Review,* **77**, 593-604.

Frank, R. (1988). *Passions within Reason*: *The Strategic Role of the Emotions*, New York: WW Norton.

Frank, Steven (1998), *Foundations of Social Evolution*, Princeton NJ: Princeton University Press.

Friedman, D. & Yellin, J. (1997). Evolving landscapes for population games. UC Santa Cruz manuscript.

Friedman, D. (1991). Evolutionary games in economics. *Econometrica*, **59**, 637-666.

Friedman, Daniel and Joel Yellin (1997), “Evolving Landscapes for Population Games,” UC Santa Cruz manuscript.

Friedman, Daniel and Nirvikar Singh (1999), “On the Viability of Vengeance,” UC Santa Cruz manuscript, May. http://econ.ucsc.edu/~dan/

Friedman, Daniel and Nirvikar Singh (2001), “Evolution and Negative Reciprocity,” in Y. Aruka, ed., *Evolutionary Controversies in Economics,* Tokyo: North-Holland, forthcoming 2001.

Friedman, Daniel and Singh, Nirvikar. “Vengeful Preferences.” Paper presented at the UC Davis conference on Preferences and Social Settings, May 18-19, 2001.

Friedman, D. and Singh, N. (2003a) Negative Reciprocity: The Coevolution of Memes and Genes. Working Paper, UC Santa Cruz.

Friedman, D. and Singh, N. (2003b) Equilibrium Vengeance. Working Paper, UC Santa Cruz.

Friedman, Milton. “The Methodology of Positive Economics.” In *Essays in Positive Economics*. Chicago: University of Chicago Press, 1953.

Fudenberg, D. & Tirole, J. (1991). *Game Theory*, Cambridge, MA: MIT Press.

Fudenberg, Drew, and Maskin, Eric. “The Folk Theorem in Repeated Games with Discounting or with Incomplete Information.” Econometrica, 1986, 54:3, pp. 533-554.

G¨uth, Werner and Kliemt, Hartmut and Peleg, Bezalel. “Co-evolution of Preferences and Information in Simple Games of Trust.” Manuscript, Humboldt University Berlin, 2001.

G¨uth, Werner and Kliemt, Hartmut. “Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes.” *Metroeconomica*, 1994, 45:2, pp. 155-187.

G¨uth, Werner and Yaari, Menachem. “An Evolutionary Approach to Explaining Reciprocal Behavior,” in U. Witt, ed., Explaining Process and Change-Approaches to Evolutionary Economics. Ann Arbor, The University of Michigan Press, 1992.

Galaty, J.G. & Bonte, P., eds. (1991). *Herders, Warriors and Traders: Pastoralism in Africa*, Boulder, CO: Westview Press.

Geanakopolis, John , Pearce, David and Stacchetti, Ennio. “Psychological Games and Sequential Rationality.” *Games and Economic Behavior*, 1989, 1, pp. 60-79.

Gilmore, D.D. (1991). *Manhood in the Making: Cultural Concepts of Masculinity*, New Haven: Yale University Press.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, **206**, 169-179.

Haldane, J.B.S. (1955). Population genetics. *New Biology*, **18**, 34-51.

Hamilton, W.D. (1964). The evolution of social behavior. *Journal of Theoretical Biology*, **7**, 1-52.

Heckathorn, D. (1996). The dynamics and dilemmas of collective action. *American Sociological Review*, 61, 250-277.

Henrich, J. and Boyd, R. (2001). Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, 208, 79-89.

Hirshleifer, J. (1987). On the emotions as guarantors or threats and promises. In: J. Dupré (ed.) *The Latest on the Best: Essays in Evolution and Optimality.* Cambridge, MA: MIT Press.

Huck, S. & Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, **28**, 13-24.

Jacobsen, Hans Jørgen, Jensen, Mogens and Sloth, Birgitte. "Evolutionary Learning in Signalling Games." Games and Economic Behavior, 2001, 34:1, pp. 34-63.

Kaufman, S. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*, NY: Oxford U Press.

Kockesen, Levent, Ok, Efe A. and Sethi, Rajiv. "The Strategic Advantage of Negatively Interdependent Preferences" *Journal of Economic Theory*, June 2000, Vol. 92, No. 2, pp. 274-299.

Leimar, O. and Hammerstein, P. (2000). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London* B, **268**, 745-753.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics,* **1**, 593-622.

Lowie, R. H. (1954). *Indians of the Plain*. New York: McGraw-Hill.

MacDonald K. B. (1994). *A People That Shall Dwell Alone: Judaism as a Group Evolutionary Strategy*. Westport, CT: Praeger.

Maynard Smith, John, and Price, George R. "The Logic of Animal Conflict." *Nature*, 1973, 246, pp. 15-18.

Mullainathan, Sendhil and Thaler, Richard "Behavioral Economics." MIT Working Paper 00-27, September 2000. To appear in textitInternational Encyclopedia of the Social and Behavioral Sciences

Nelson, R.R. and Winter, S.G. (1982). *An evolutionary theory of economic change*. Cambridge, MA: Belknap Press of Harvard University Press.

N¨oldeke, Georg and Samuelson, Larry. “A Dynamic Model of Equilibrium Selection in Signaling Markets.” Journal of Economic Theory, 1997, 73, pp. 118-156.

Nisbett, R.E. & Cohen, D. (1996). *Culture of Honor: the Psychology of Violence in the South*, Boulder, CO: Westview Press.

Nowak, M. A. & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, **393**, 573-577.

O’Kelley, C.G. & Carney, L.S. (1986), *Women and Men in Society*, New York: D. Van Nostrand Co.

Peristiany, J.G. ed., (1965). *Honor and Shame: The Values of Mediterranean Society*, London: Weidenfeld and Nicolson.

Pettigrew, J. (1975). *Robber Noblemen: A Study of the Political System of the Sikh Jats*. London: Routledge & Kegan Paul.

Possajennikov, Alex. (2002a) ”Two-Speed Evolution of Strategies and Preferences in Symmetric Games”, Discussion Paper 02-03, National Research Center 504 ”Rationality Concepts, Decision Behavior, and Economic Modeling”, University of Mannheim, January

Possajennikov, Alex. (2002b) Cooperative Prisoners and Aggressive Chickens: Evolution of Strategies and Preferences in 2x2 Games.” Discussion Paper 02-04, National Research Center 504 ”Rationality Concepts, Decision Behavior, and Economic Modeling” University of Mannheim, January

Price, G. R. (1970). Selection and covariance. *Nature,* **227**(5257, August 1), 520-521.

Rabin, Mathew (1993), “Incorporating Fairness into Game Theory and Economics,” *American Economic Review* 88:5, 1281-1302.

Richerson, Peter J and Robert Boyd, “The Evolution of Ultrasociality,” in I Eibl-Eibesfeldt and F. K Salter, eds, *Indoctrinability, Ideology and Warfare*, NY: Berghahn Books.

Rilling, James K., Gutman, David A., Zeh, Thorsten R., Pagnoni, Guiseppe, Berns, Gregory S., and Kitts, Clinton D. “A Neural Basis for Cooperation.” *Neuron*, 2002, 35, pp. 395-405.

Robson, Arthur J.. “Evolution and Human Nature.” Journal of Economic Perspectives, 2002, 16:2, pp. 89-106.

Rosenthal, R. W. (1996). Trust and social efficiencies. Boston University manuscript.

Rubin, Paul H. and Paul, C.W. “An Evolutionary Model of Taste for Risk.” *Economic Inquiry*, 1979, 17, pp. 585-596.

Samuelson, Larry and Swinkels, Jeroen. “Information and the Evolution of the Utility Function.” Mimeo, University of Wisconsin, 2001.

Samuelson, Larry. “Introduction to the Evolution of Preferences.” *Journal of Economic Theory*, 2001, 97, pp. 225-230.

Sethi, R. & Somanathan, E. (1996). The evolution of social norms in common property resource use. *American Economic Review*, **86**, 766-788.

Sethi, R. and Somanathan, E. (2001). Preference evolution and reciprocity. *Journal of Economic Theory*, 97, 273-297.

Sethi, R. & Somanathan, E. (2003). Understanding reciprocity. *Journal of Economic Behavior and Organization*, **50**, 1-27.

Sobel, Joel. “Social Preferences and Reciprocity.” Mimeo, University of California at San Diego, 2000.

Sober, E. & Wilson, D.S. (1998). *Onto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sugden, R. (1986). *The Economics of Rights, Co-operation and Welfare*, New York: B. Blackwell.

Trivers, Robert L. (1971), “The Evolution of Reciprocal Altruism,” *Quarterly Review of Biology* 46, 35-57.

van Winden, Frans. “Emotional Hazard Exemplified by Taxation-induced Anger” *Kyklos*, 2001, 54, pp. 491-506.

Weibull, J.W. (1995). *Evolutionary Game Theory.* Cambridge, MA: MIT Press.

Weingart, P., Boyd, R., Durham, W. H. & Richerson, P. J. (1997). Units of culture, types of transmission. In Weingart, P., Mitchell, S. D., Richerson, P. J., & Maasen, S. (Eds.) *Human By Nature: Between Biology and the Social Sciences*. Lawrence Erlbaum, Mahwah, NJ. .

Wittman, Donald. “Why Democracies Produce Efficient Results.” *Journal of Political Economy*, 1989, 97(6), pp. 1395-1424.

Wright, Sewall. “Adaption and Selection,” in L. Jepsen, G.G. Simpson, and E. Mayr eds., *Genetics, Paleontology, and Evolution*. Princeton, N.J.: Princeton University Press, 1949.

Figure 1: Payoffs to Self and Other

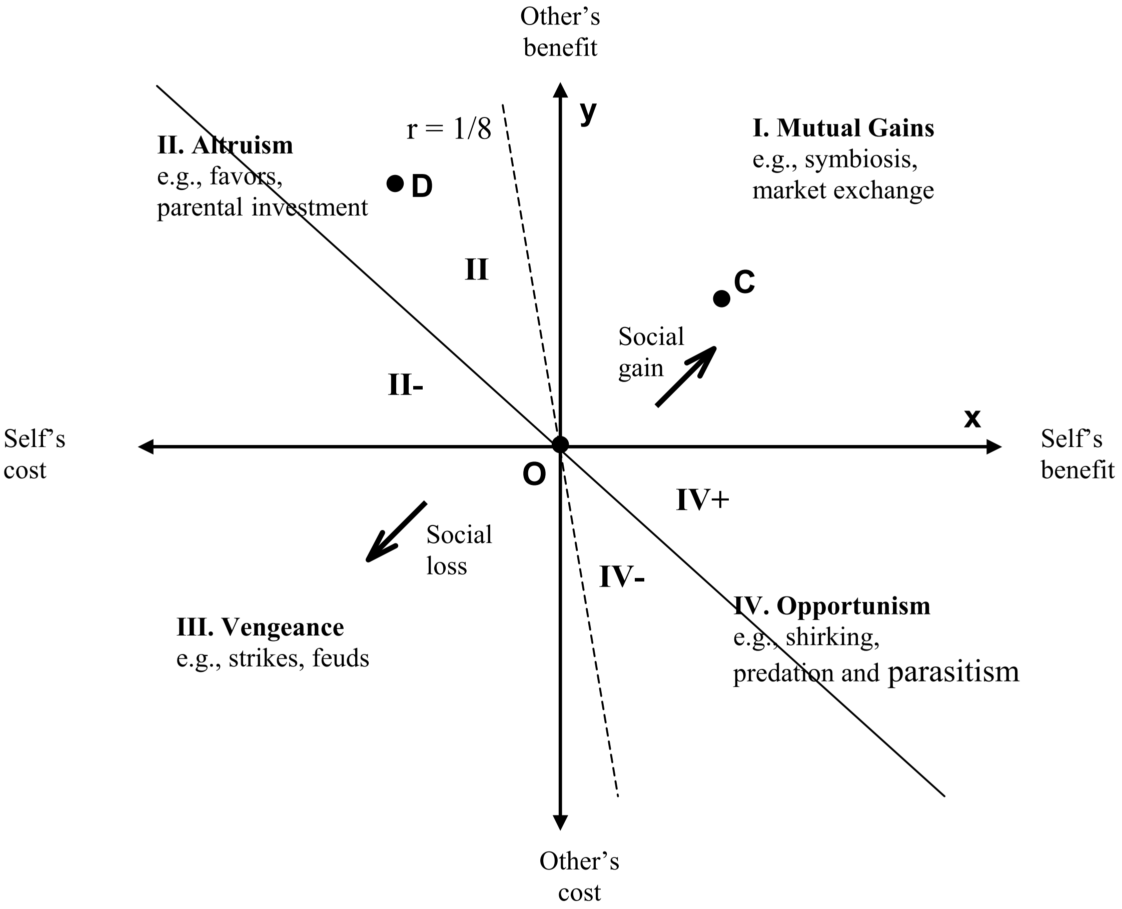
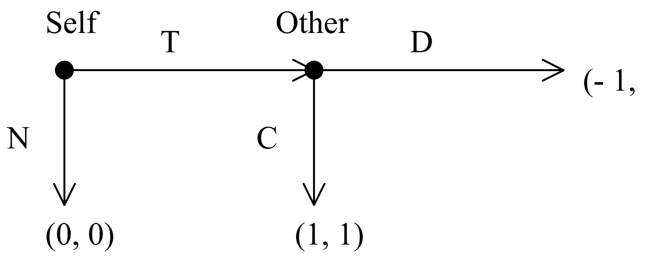
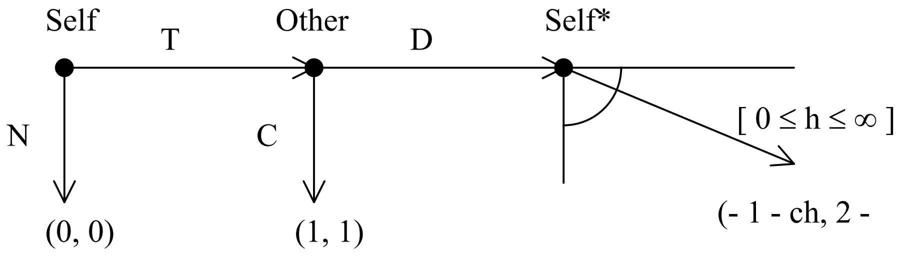


Figure 2: Fitness Payoffs

A. Basic Trust Game

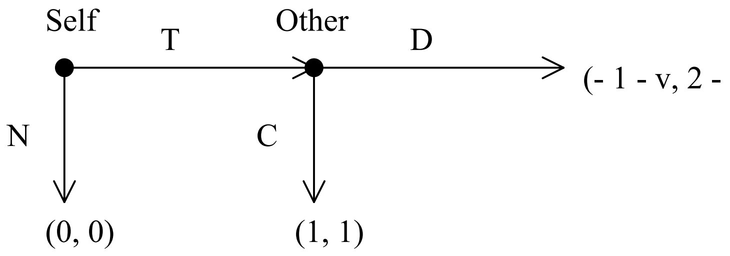


B. Trust with a Vengeance Technology



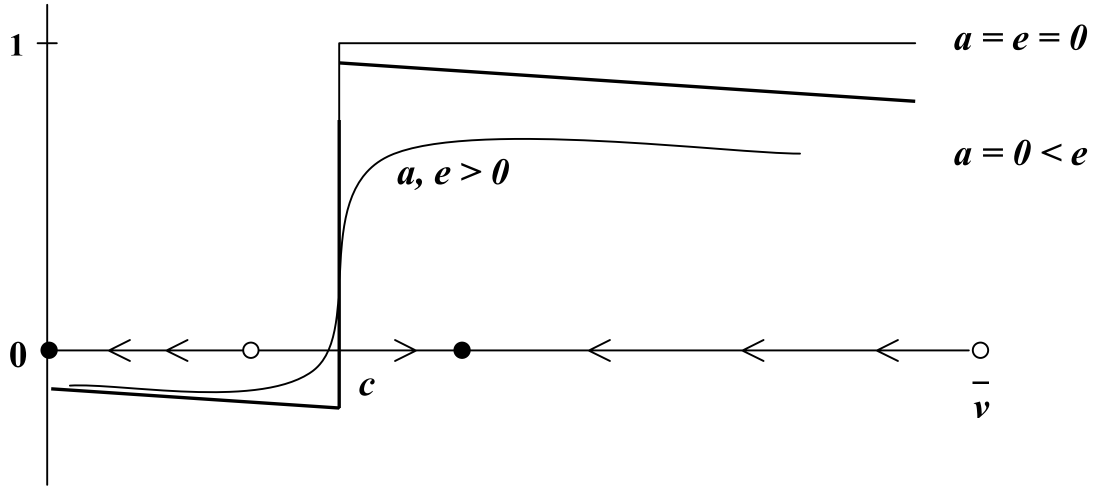
*Utility payoff to Self is v *ln*h - 1 – ch

C. Trust with a Vengeance (Reduced*)



*Self's last move on branch D inflicts harm h=v/c at cost v.

Figure 3: Fitness *W* as a Function of Vengefulness *v*



Note: For *a* = *e* = 0, the fitness function is a unit step function at *v* = *c*. Up to first order in behavioral noise amplitude *e*, the fitness function for *a* = 0 has slope –*e* on the first segment and –2*e* on the second segment. For signal noise amplitude *a* > 0, the fitness function is the convolution of the *a*=0 fitness function with the signal noise density function. It has a local maximum at *v*=0 and a global maximum near *v*=*c*+ *a* (solid dots) and a minimum near *v*=*c*- *a* (open circle).