UC Santa Cruz
Working Paper Series

Title
Vengefulness Evolves in Small Groups

Permalink
https://escholarship.org/uc/item/0xp29105

Authors
Friedman, Daniel
Singh, Nirvikar

Publication Date
2004-04-01

Vengefulness Evolves in Small Groups

Daniel Friedman and Nirvikar Singh

Department of Economics

University of California, Santa Cruz

January 2004

Abstract

We discuss how small group interactions overcome evolutionary problems that might otherwise erode vengefulness as a preference trait. The basic viability problem is that the fitness benefits of vengeance often do not cover its personal cost. Even when a sufficiently high level of vengefulness brings increased fitness, at lower levels, vengefulness has a negative fitness gradient. This leads to the threshold problem: how can vengefulness become established in the first place? If it somehow becomes established at a high level, vengefulness creates an attractive niche for cheap imitators, those who look like highly vengeful types but do not bear the costs. This is the mimicry problem, and unchecked it could eliminate vengeful traits. We show how within-group social norms can solve these problems even when encounters with outsiders are also important.

1. Introduction

After a century of neglect, economists in the last decade or two began to write extensively about social preferences. The vast majority of the articles so far have focused on altruism or positive reciprocity. Only a few examine the dark side, negative reciprocity or vengefulness. When some culprit harms you (or family or friends), you may choose to incur a substantial personal cost to harm him in return. Vengeance deserves serious study because it has major economic and social consequences, positive and negative. For example, workers' negative reciprocity at the Decatur plant threatened to bring down Firestone Tires (Krueger and Mas, 2004); terrorists often explain their actions as revenge against the oppressor; and successful corporate cultures somehow forestall petty acts of vengeance and other sorts of dysfunctional office politics.

A taste for vengeance, the desire to "get even," is so much a part of daily life 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, vengefulness would have died out long ago, or would never have appeared in the first place.

Why then does vengeance exist? Economists' natural response is to think of vengeance as the punishment phase of a repeated game strategy that supports altruism. The models supporting this view are now taught to all Economics PhD students and many undergraduates, and for good reason. Yet they hardly capture the whole story. The standard models have no place for the powerful emotions surrounding vengeance, and their predictions do not match up especially well with everyday experience. One often sees vengeance when the discount factor is too small to support rational punishment (e.g., in once-off encounters with strangers), and often the rational punishment fails to appear (e.g., when a culprit apologizes sincerely).

The present paper explores a different class of models. We consider repeated interactions in the context of small groups that enforce social norms. The norms are modeled not as traits of individual group members, but rather as traits of the group itself. We show that such group traits naturally support efficient levels of the taste for vengeance when encounters outside the group are also important. However, the model discloses two further problems. The threshold problem asks how vengeance can evolve from low values where it has a negative fitness gradient. The

mimicry problem asks why cheap imitators do not evolve who look like highly vengeful types but do not bear the costs of actually wreaking vengeance. We argue that small group interactions can overcome both problems.

The next section sets the stage with a simple illustration of the 'fundamental social dilemma:' evolution supports behavior that is individually beneficial but socially costly. We mention the standard devices for resolving the dilemma—genetic relatedness and repeated interactions—but focus on the more recent device of social preferences under the indirect evolution approach, as pioneered by Güth and Yaari (1992). Section 3 lays out the issues in more detail. It presents a simple Trust game, very similar to that analyzed by Güth and various coauthors, and uses that game to lay out the social dilemma, and the threshold and mimicry problems.

Sections 4 and 5 are the heart of our analysis. We explain the role of group traits, their relation to individual fitness, the time scales governing their evolution, and how they can overcome the threshold and mimicry problems. Section 5 presents a more formal argument that the group traits adjust behavior in small groups towards the socially optimal level. Section 6 offers an extended discussion of how our approach relates to existing literature, and Section 7 concludes with remarks on remaining open issues.

2. Vengefulness as an Evolutionary Puzzle

Figure 1 illustrates 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".[1] Social dilemmas arise from the fact the Self's fitness gradient is the x-axis, while in contrast, the social efficiency gradient is along the 45 degree line. Social creatures (such as humans) thrive on cooperation, by which we mean devices that support efficient altruistic outcomes in II+ and that discourage inefficient opportunistic outcomes in IV-. Such cooperation arises from devices that somehow internalize Other's costs and benefits.

Quadrant III is anomalous; indeed, Cipolla (1976) refers to such behavior as "stupidity." Behavior producing quadrant III outcomes harms both Self and Other, contrary to efficiency as

---

[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.

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 cooperation. But first we mention two other, and better-known, devices that can serve the same purpose.

Genetic Relatedness

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 iso-fitness lines take the form $[x + ry = k]$. For example, the unusual genetics of insect order *hymenoptera* lead to $r=3/4$ between full 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.

Repeated Interactions

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 that produces outcomes on higher iso-fitness lines $[x + \delta y = k]$.[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.

---

[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.

Other Regarding Preferences and Indirect Evolution

Our focus is on other-regarding preferences. For example, suppose Self gets a utility increment of ry. Then Self partially internalizes the material externality, and will choose behavior that attains higher indifference curves $[x + ry = k]$. 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. 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?

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 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, Güth 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 Güth 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 (2000), Possajennikov (2002a, 2002b), and Samuelson and Swinkels (2001). Many of these papers focus on positive reciprocity rather than on negative reciprocity, or vengeance. For example, the key issue in Güth, Kliemt and Peleg (2001) is the cost of observing Other's true preferences for positive reciprocity (or altruism; in their game the two cannot be distinguished).

---

[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.

3. Modeling Issues

We discuss the leading approaches to modeling social preferences, and then lay out a canonical Trust game. Using this game, we present the evolutionary problems of viability, threshold and mimicry.

Social Preferences

Two main approaches can be distinguished in the recent literature. The distributional 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. Otherwise put, I am altruistic when I am ahead and spiteful when I am behind you, irrespective of what you might have done to put me ahead or behind.

The other main approach is to model reciprocity in equilibrium. Building on the Geanakoplos, Pearce and Stacchetti (1989) model of psychological games, Rabin (1993) constructs a model of reciprocity in two player normal form games, extended by Dufwenberg and Kirchsteiger (1998), as well as Falk and Fischbacher (1998), to somewhat more general games. The basic idea is that my preferences regarding your payoff depend 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 usually intractable. 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.

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 (i.e., positive and negative reciprocity), while noting that distributional preferences may also be part of the picture.

The Trust Game

The first step in developing these ideas is to model the underlying social dilemma explicitly. 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 known as the Trust game, introduced in Güth and Kliemt (1994) and Romer (1995).

Panel A of Figure 2 presents the basic 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. (Of course one can pick more general parameterizations of the game, but these simple numbers suffice for our purposes.)

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 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.[4] 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). The threat of vengeance rationalizes Other's cooperation and Self's trust.

The Viability Problem

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 and observational 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 an ankle and find himself depending on Other's cooperative behavior. Likewise, Other may intend to choose C but

---

[4] 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. Güth has a series of papers with various coauthors that develop the evolutionary implications of such friendly (or positively reciprocal) preferences.

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. Thus assume that Other's perception of v includes an observational error with amplitude $a \geq 0$.

The key task 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. 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 $v=c$. One can show (Friedman and Singh, 2003b) 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. In this case, a high level of vengefulness ($v > c + a$) brings high fitness and thus is viable.

The Threshold Problem

How will vengeful traits evolve in the Self population? It is inappropriate to assume standard replicator dynamics or monotone dynamics for a continuous trait like v.[5] Stochastic dynamics such as noisy fictitious play, or the Kandori-Mailath-Rob (1993)/Young (1993) dynamic also apply to traits with discrete alternatives, such as eye color, but have no natural application to traits with many ordered levels. Biological theorists from Wright (1949) through Eshel (1983) and Kaufman (1993) have routinely modeled continuous traits in terms of a fitness landscape in which evolution pushes the evolving trait v uphill. That is, selective pressure tends to increase (decrease) the level of the trait when higher (lower) levels are fitter. The underlying idea—that change is usually local and large jumps are rare—would seem to apply to a preference trait like vengefulness as well as to standard biological traits like height or foot speed (Friedman and Yellin, 1997).

Applying landscape dynamics to the fitness landscape in Figure 3, we see that 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 should lead to

---

[5] Oechssler and Riedel (2000) deal with evolutionary dynamics in continuous games, and point out some difficulties with evolutionary stability if the strategy space is continuous. In our case, it is the preference trait that is continuous, so again the situation we analyze is somewhat different.

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 $W \approx - e$.

There is a serious problem for the more vengeful type: how could it evolve from low values given the negative fitness gradient? How would a positive fraction of the subpopulation ever achieve levels above c-a in the first place? We refer to this as the threshold problem, and will outline a solution in the next section.

The Mimicry Problem

Putting aside the threshold problem for the moment, assume that there are indeed two stable types, a vengeful type with v near $c + a$, and an unvengeful type with v near 0. The observational error amplitude, a, is small so Other usually identifies Self's true type correctly. But the error amplitude itself is subject to evolutionary forces, creating what we shall call the mimicry or Viceroy problem.

An instructive example is a game played by butterflies and insect-eating birds. A butterfly can hide from birds (analogous to strategy 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 almost identical to the Monarch's, 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 cannot 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 Müllerian

mimicry) where Viceroys also evolve unpalatability. The field evidence for all three equilibria seems inconclusive.[6]

The mimicry or Viceroy problem surely arises in the extended Trust 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, absent some mechanism to control check mimicry. We shall discuss possible mechanisms in the next section.

4. Group Interactions and Group Traits

We do not know any way to overcome the threshold problem and the Viceroy problem within the context of unstructured interactions in a large population. Group interactions suggest an appealing solution to the threshold problem. Much of this section explicates the idea of group traits, which help solve the basic viability problem as well as the mimicry problem.

A Solution to the Threshold Problem

Standard game theory 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 $2 - v/c$ and continuation value 0 if he chooses D, to immediate payoff 1 and continuation value $\delta + \delta^2 + \delta^3 + \ldots = \delta/(1-\delta)$ if he chooses C. Simple calculations reveal that it is advantageous to choose C if $\delta > \frac{1}{2}$ in the $v=0$ case, and if $\delta > (c-v)/(2c-v)$ 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.)

---

[6] See Kapan (2001) and the references therein.

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.

Interactions with Outsiders and the Mimicry Problem

At first it seems that similar considerations also solve the mimicry or 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.

The real problem arises 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.[7] Assume also that individuals can fairly reliably assess any individual's group affiliation and know the reputation of the group.[8] Then we have a free rider problem with respect to group reputation. Each individual would benefit from

---

[7] Henrich and Fehr (2003) forcefully argue that this is the usual situation for contemporary hunter-gatherer groups as well as for our Paleolithic ancestors.

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 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 within 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.[9]

Three different responses to the Viceroy problem now present themselves. 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

---

[8] 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 $v^e = \lambda \bar{v} + (1-\lambda)E\bar{v} + \varepsilon$, an idiosyncratic error plus the weighted average of the partner's group average $\bar{v}$ and overall population average $E\bar{v}$, 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.

[9] M.N. Srinivas termed this process "Sanskritization" – see, for example, Srinivas (2002).

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.[10]

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 possible response 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.

Group Traits and Individual Fitness

We need to discuss the relevant traits before working out any of these responses in detail. 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

---

[10] See Akerlof (1983) for several seminal essays that model enforcement in this context, as well as Henrich and Boyd (2001) for a more recent contribution. In such cases, the costs of enforcement are a major concern. Fines are an enforcement mechanism whose cost to a group is near zero (or perhaps negative). Elinor Ostrom, in a communication with the first author, offered the example of cow jails in Nepal. Cows grazing in the wrong places are "jailed" and the owner has to pay a fine. Until he does so, the community gets the cow's milk. Hence the enforcement cost is negative, i.e., the community (apart from the owner) gets a small net benefit from punishing the norm violator.

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). Recent experiments by Weber and Camerer (2003) demonstrate that some facets of organizational culture are created by organization members but survive changes in individual membership of organizations.[11]

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, 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 the next Section) 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.

---

[11]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.

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. The model in the next section uses a single parameter t to combine the sensitivity of fitness to status with the sensitivity of status to behavior.

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 (e.g., Young, 1993), 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 for 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 through 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.

The relevant group traits here are prescriptions for responding to culprits and imitators from other groups, 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 (or imitators) 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 $t>0$.

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 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.

5. Results

We will now sketch how efficient norms of vengeance might evolve in our setting. We use the extended Trust game with observational and behavioral errors, as in Figure 2, and assume that Self and Other belong to different groups. For reasons discussed in connection with Figure 3, we assume a two-point distribution of types for Self: they can either have vengeance parameter 0 or *v* > *c*. We study a separating Perfect Bayesian Equilibrium (PBE). 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, there is a pooling equilibrium with only trust and cooperation.

Table 1: Fitness and Probabilities in Separating PBE

		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	$e(1 - \alpha)$

<u>Notes:</u> 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 $\alpha = a(1 - e) + e(1 - a) = e + a - 2ae$.

The fitness payoffs and probabilities in separating equilibrium are summarized in Table 1. 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, the group trait. 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

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

The short run dynamics push the individual's vengeance parameter toward the value that maximizes individual expected fitness at the given social norm. A simple calculation yields the first order condition $\rho'(v - v^n) = 1$. In the quadratic case the condition reduces $v = v^n - t$. Thus in short run (hence also in medium and long run) equilibrium, groups enjoin an exaggerated version of the optimal v, but the individually optimal v prevails. That optimum is as in Fig 3, since group reputations have only small observational error.

Note the comparative statics: the punishment technology for out-group interactions is parameterized by 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 punishment in medium and 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. But this group average evolves in the medium run. To see how, first note that status losses represented by the function $\rho(x; t)$ net out to 0 for the group, and so average fitness is

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

The subtlety for medium run dynamics is that the observational error amplitude is negatively related to the level of the group average vengeance parameter. That is, $a = A(\bar{v})$ for some decreasing function A. For the functional form $A(v) = 0.5\exp(-v/b)$,[12] and many other specifications, $W^g(\bar{v})$ is single peaked at some optimal level v^o of the group average vengeance parameter $\bar{v}$.

The group optimum v^o is characterized by first order condition $A'(v^o)(2 + v^o) + A(v^o) = 0$. This expression can be solved explicitly for the given parametric versions of A and ρ to yield the simple expressions $v^o = b - 2$ and $v^n = t + b - 2$. In general, evolution in the medium run pushes the group trait (v^n, ρ) so as to increase $W^g(\bar{v})$. Absent corner solutions or multiple peaks in the group fitness function,[13] the group trait will evolve so that it supports an optimal level of vengefulness in interactions outside the group.

We close this section with some caveats. Our result is partial equilibrium in that it takes as given the behavioral error rate e, the observational error function A, the marginal punishment cost c, and a sufficiently large upper bound v^{max} on vengefulness to avoid corner solutions. We have already pointed out, in section 4, that the upper bound is not a problem in the long run. Also, we have not worked out how the entire distribution of vengeance parameters over different groups might evolve. 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).

---

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

[13] Our papers discussed at the end of section 6 show that corner solutions disappear in the long run equilibrium and that multiple peaks do not arise in the interesting cases.

6. Related Literature

Ours is not the only analysis of vengefulness. Elster (1989) was perhaps the first to highlight vengeance as a problematic economic issue, and to suggest the importance of social norms in overcoming this problem. Since then several authors have encountered the viability problem in one form or another, and have found ways to finesse it.

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*.

Sethi and Somanathan (1996) offer two attempts to get around the viability problem. First, they define stability to include neutral stability, not requiring convergence back to an equilibrium point following a small perturbation (i.e., they do not require local asymptotic

stability). In their model there is a continuum of neutrally stable equilibria with no culprits. Following a perturbation (a small invasion of culprits) the state moves along the continuum away from a vertex. Eventually, following sufficiently many such perturbations, the state leaves the equilibrium set and ultimately converges back to the all-D equilibrium. Thus, from a long-run evolutionary viewpoint, their other equilibria really are not stable, and their vengeful strategies are not viable. Implicitly recognizing the problem, Sethi and Somanathan refer in an appendix to a second approach, due to Binmore and Samuelson (1999), in which the evolutionary dynamic is perturbed by a continuing stream of mutants in fixed positive proportions. The perturbed dynamic has a single asymptotically stable equilibrium point instead of the continuum of neutrally stable equilibria, but it has a very shallow basin of attraction and is supported by an arbitrary convention on the composition of mutants.

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)[14], 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. Richerson and Boyd (1998) point out that genetic group selection in humans is implausible due to relatively rapid cross-group gene flow rates. Indeed, Sober and Wilson devote much of their book to discussing cultural norms for rewarding cooperative behavior and punishing uncooperative behavior. They

---

[14] 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.

avoid the viability problem by assuming in essence that c is 0 (see p.151 of their book 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. 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). The former paper shares some ideas with Bowles and Gintis (1998), in a model of team production with mutual monitoring. A sufficient proportion of 'strong reciprocators', who gain subjective payoffs from punishing shirkers, leads to a more cooperative outcome. Sethi –Somanathan uses a variant of reciprocal preferences, which place negative weight on the payoffs of materialists (those with conventional selfish preferences) and positive weight on the payoffs of sufficiently altruistic individuals. Such preferences do better evolutionarily than purely altruistic or spiteful preferences.

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.[15] This "greenbeard" solution[16] of course ignores the mimicry problem.

---

[15] 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.

Henrich and Boyd (2001) argue that the negative gradient or threshold problem can be overcome within groups if more popular behavior tends to be imitated, even when this conformity effect is very weak. Groups with this trait would achieve better internal cooperation, and displace other groups. 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., second 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 the 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.)

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. Nowak and Sigmund model (and simulate numerically) indirect reciprocity as "image scoring", in which an individual's score increases when he or she helps someone who needs it, and decreases when such help is not offered. This process is thought of as taking place in social groups that are small enough to allow members to track everyone else's scores. Leimar

---

[16] 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.

and Hammerstein (2000) have suggested that image scoring by itself is not individually rational, and offered alternative simulations that call its evolutionary robustness into question.

Finally, we should note how the current paper fits with our own earlier work. Much of the material comes from our 1999 working paper. The underlying game there and in our 2001a and 2003a papers, however, is a simultaneous move Prisoners' dilemma rather than the Trust game. Our 2003b paper is based on the Trust game, but considers only interactions in a large population with no group structure, and obtains equilibria with no optimality properties parallel to those obtained here. It considers alternative assumptions on the observational error function $A(v)$ and focuses on a Gaussian rather than exponential form, and analyzes the resulting second order conditions it detail. The 2003a paper allows for somewhat more complex group interactions and perceptions than in the present paper, and includes a more detailed discussion of reputation and status issues, and of relevant biological constraints on vengeance. As in the present paper, it obtains an optimality result based on the marginal logic of trading off the cost of individual retaliation and the impact on status within the group. None of our previous papers treats the threshold and mimicry problems.

7. Discussion

We have argued that small group interactions play a crucial role in the evolution of vengeful preferences. The most relevant and problematic interactions across groups (a) are not frequent enough to support the use of repeated game or related mechanisms for reciprocity, yet (b) are important enough in the aggregate to affect fitness. We showed how small groups can, at low cost to the group, enforce specific norms of vengeance on their members. Status is key: those who depart further from the group norm suffer greater reductions in status, which ultimately decreases their fitness. Individual adherence to group norms, while imperfect, can be strong enough in evolutionary equilibrium to sustain cooperative outcomes in inter-group encounters. Thus small groups can overcome the basic viability problem for vengeance.

Earlier we presented a simple argument on how small groups overcome the threshold problem. Within small groups, even a small degree of vengefulness can help support repeated game equilibria. We also discussed how status-mediated group enforcement can also discipline mimicry by outsiders as well as by group members.

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? Indirect evolution dictates that 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? Friedman and Singh (2003b) introduced the concept of Evolutionary Perfect Bayesian Equilibrium, which is 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.

References

Akerlof, George, and Rachel Kranton (2000), Economics and Identity. *Quarterly Journal of Economics*, 115, pp. 715-753.

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

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

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

Binmore, Kenneth, and Larry Samuelson (1999), Evolutionary drift and equilibrium selection. *Review of Economic Studies*, 66, pp. 363-393.

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

Bowles, Samuel, and Herbert Gintis (1998), The evolution of strong reciprocity. Department of Economics, University of Massachusetts, Amherst working paper.

Bowles, Samuel, and Herbert Gintis (2001), Social Capital and Community Governance, Department of Economics, University of Massachusetts, Amherst working paper.

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

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

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

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

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

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

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

Elster, Jon (1989), Social Norms and Economic Theory, *Journal of Economic Perspectives*, 3 (4): 99-117

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

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

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

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

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

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

Frank, Robert (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, Daniel (1991), Evolutionary games in economics. *Econometrica*, 59, pp. 637-666.

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.

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

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

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

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

Güth, Werner and Kliemt, Hartmut and Peleg, Bezalel. (2001) "Co-evolution of Preferences and Information in Simple Games of Trust." Manuscript, Humboldt University Berlin.

Güth, Werner and Kliemt, Hartmut. (1994). "Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes." *Metroeconomica*, , 45:2, pp. 155-187.

Güth, Werner and Yaari, Menachem. (1992) "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.

Geanakoplos, John, Pearce, David and Stacchetti, Ennio. (1989) "Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1, pp. 60-79.

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

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

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.

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

Kandori, M., G. J. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. E*conometrica*, 61, pp. 29-56, 1993.

Kapan, Durrell D. (2001), Three-Butterfly System Provides A Field Test Of Müllerian Mimicry *Nature* 409, Pp. 338 – 340.

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.

Krueger, Alan B., and Alexandre Mas (2004), "Strikes, Scabs and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires," forthcoming, *Journal of Political Economy*.

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

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

Nelson, Richard R. and Stanley G. Winter (1982), *An Evolutionary Theory of Economic Change*. Cambridge, MA: Belknap Press of Harvard University Press.

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.

Oechssler, Jorg and Frank Riedel (2000), On the Dynamic Foundation of Evolutionary Stability in Continuous Models, ftp://ftp.wipol.uni-bonn.de/pub/RePEc/bon/bonedp/bgse7_2000.pdf.

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.

Romer, Paul, (1995), "Preferences, Promises, and the Politics of Entitlement," in *Individual and Social Responsibility: Child Care, Education, Medical Care, and Long-Term Care in America,* Victor R. Fuchs (ed.), Chicago: University of Chicago Press.

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.

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.

Srinivas, Mysore N. (2002), *Collected Essays*. New Delhi: Oxford University Press.

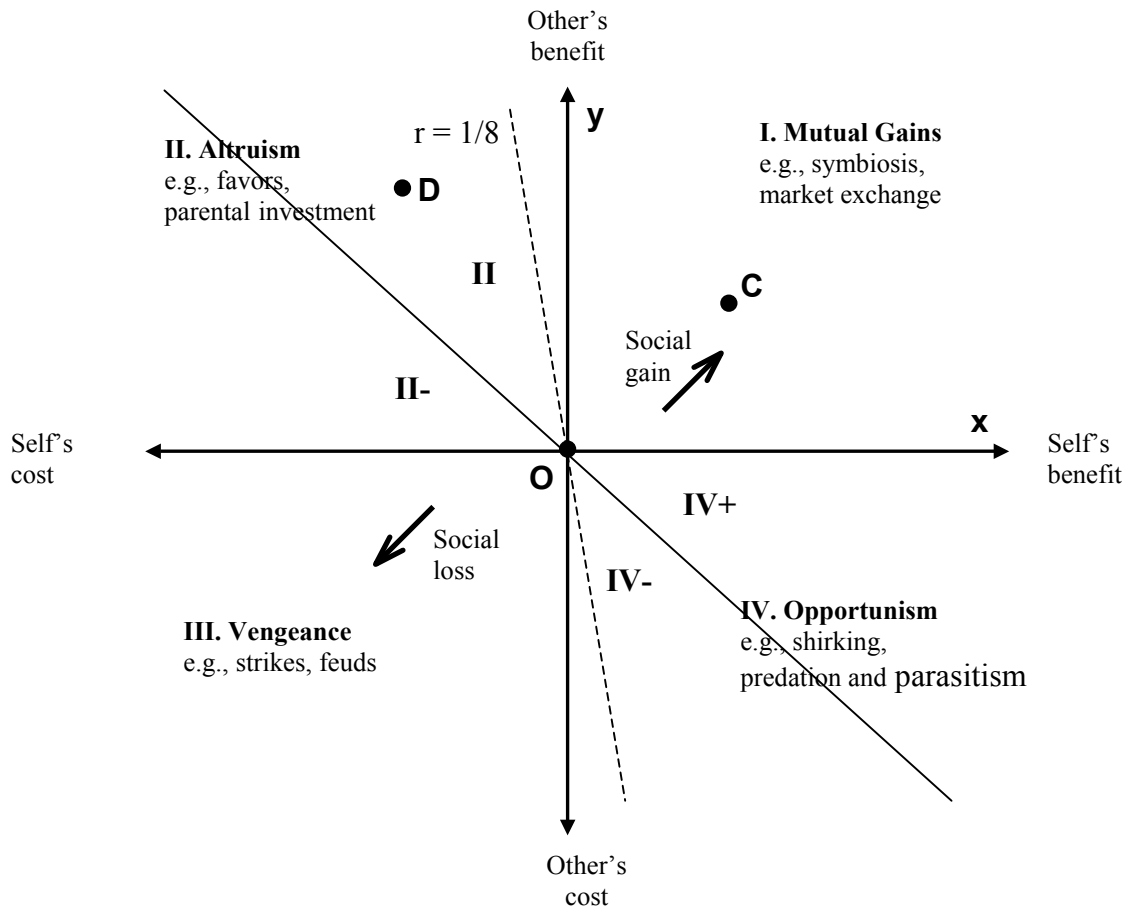Sugden, R. (1986). *The Economics of Rights, Co-operation and Welfare*, New York: B. Blackwell.

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

Weber, Roberto A., and Colin F Camerer (2003), Cultural Conflict and Merger Failure: An Experimental Approach, *Management Science*, 49, 4, 400-415.

Wright, Sewall (1949), "Adaption and Selection," in L. Jepsen, G.G. Simpson, and E. Mayr eds., *Genetics, Paleontology, and Evolution*. Princeton, N.J.: Princeton University Press, pp. 365-389.
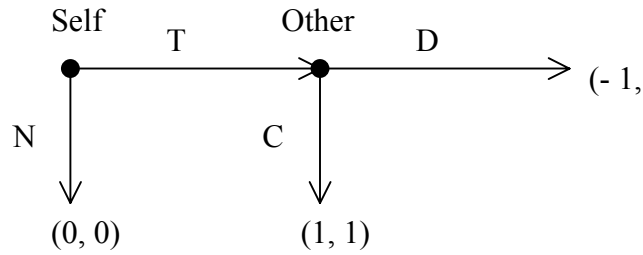
Young, H. Peyton (1993), "The Evolution of Conventions," *Econometrica*, 61, 57-84.
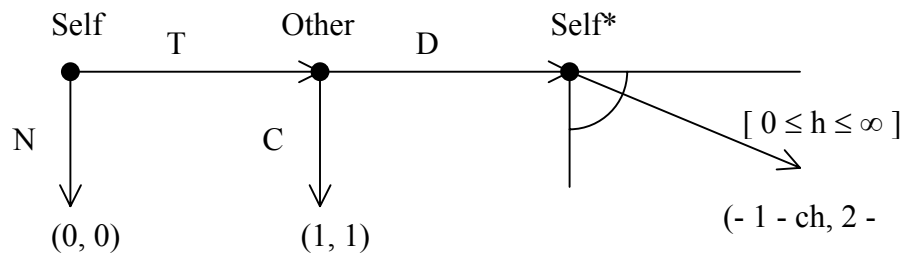
Figure 1: Payoffs to Self and Other

Other's
benefit

y

r = 1/8

II. Altruism
e.g., favors,
parental investment

● **D**

II

I. Mutual Gains
e.g., symbiosis,
market exchange

● **C**

II-

Social
gain

x

Self's
cost

O

Self's
benefit

Social
loss

IV+

IV-

III. Vengeance
e.g., strikes, feuds

IV. Opportunism
e.g., shirking,
predation and parasitism

Other's
cost

Figure 2: Fitness Payoffs

A. Basic Trust Game

Self   T   Other   D   (- 1,

N   C

(0, 0)   (1, 1)

B. Trust with a Vengeance Technology

Self   T   Other   D   Self*

$[\ 0 \leq h \leq \infty\ ]$

N   C

(0, 0)   (1, 1)

(- 1 - ch, 2 -

*Utility payoff to Self is v lnh - 1 – ch

C. Trust with a Vengeance (Reduced*)

Self   T   Other   D   (- 1 - v, 2 -

N   C

(0, 0)   (1, 1)
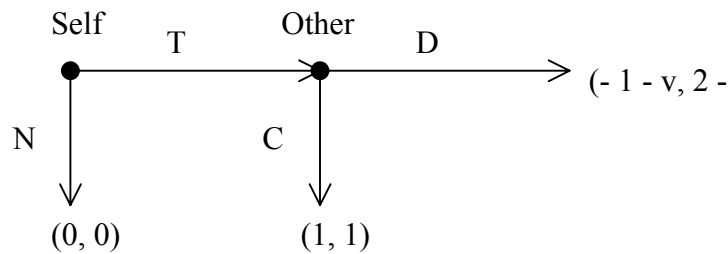
*Self's last move on branch D inflicts harm h=v/c at cost v.

Figure 3: Fitness _W_ as a Function of Vengefulness _v_
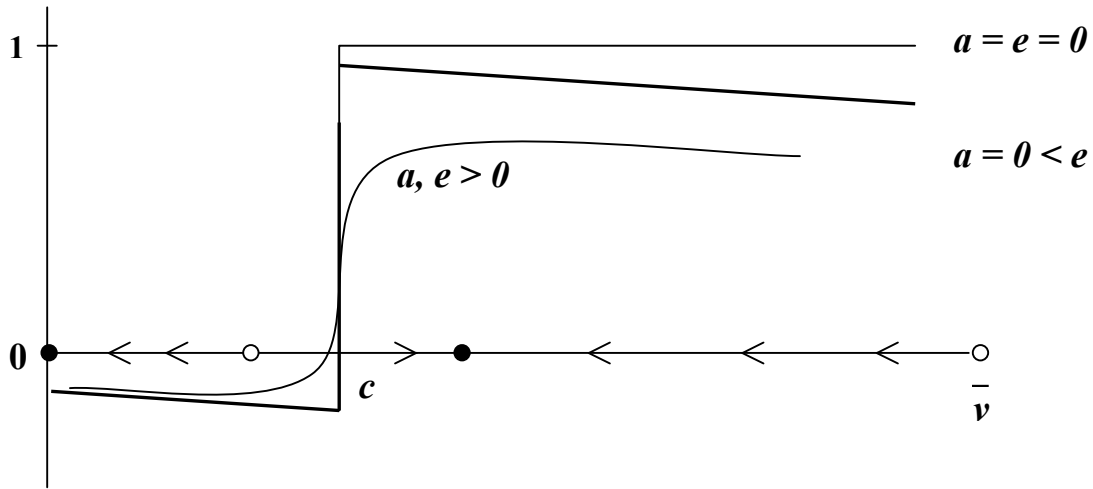
Note: For _a_ = _e_ = 0, the fitness function is a unit step function at _v_ = _c_. Up to first order in behavioral noise amplitude _e_, the fitness function for _a_ = 0 has slope –_e_ on the first segment and –2_e_ on the second segment. For signal noise amplitude _a_ > 0, the fitness function is the 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).