Title: **Evolution and Negative Reciprocity[1]***

Authors: **Daniel Friedman[2] & Nirvikar Singh*[3]**

Address: **Department of Economics**
Social Sciences 1
University of California at Santa Cruz
Santa Cruz CA 95064
USA

Address for Proofs: **Professor Daniel Friedman**
Department of Economics
Social Sciences 1
University of California at Santa Cruz
Santa Cruz CA 95064
USA

Fax Number: **(831) 459-5900**

E-mail Address: **dan@cats.ucsc.edu (Daniel Friedman)**
boxjenk@cats.ucsc.edu (Nirvikar Singh)

Key words: **Negative reciprocity, vengeance, coevolution, memes**

[2] Department of Economics, University of California at Santa Cruz
[3] Department of Economics, University of California at Santa Cruz

* Department of Economics, University of California, Santa Cruz
Summary

We offer a theoretical explanation of negative reciprocity or vengeance, the human desire to harm those who have harmed us. Our model shows how negative reciprocity can be sustained by the coevolution[4] of genes that determine the capacity for vengeance and group memes (e.g., social norms) that regulate its expression. The model begins with a standard free rider game that captures, simply and directly, a personal cost incurred to reap social gains. The model shows that a taste for vengeance realigns incentives and supports a socially efficient equilibrium, but that by itself the taste for vengeance is not evolutionarily viable. We then show how groups of individuals can use low-power sanctions (or simply status changes) to enforce a particular norm on the proper degree of vengeance. The main result is that actual behavior typically will fall short of the norm, but selection across groups will adjust the norm so that actual behavior maximizes the fitness of group members.

---

[4] See Dawkins (1976), Boyd and Richerson (1985), and Durham (1991).

1. Introduction

Vengeance is a powerful human motive. We become angry when someone wrongs us, and often try to harm the culprit in return, even at some personal cost. Vengeance deters opportunistic behavior[5] that otherwise might undermine positive reciprocity, direct or indirect[6], and thereby supports social cooperation. On the other hand, misplaced vengeance sometimes leads tragic feuds and even genocide, as in Rwanda or the Balkans.[7]

The existence of vengeance is empirically obvious (and confirmed in controlled experiments)[8], but theoretically mysterious, because vengeance is not individually rational: it is weakly dominated by otherwise similar behavior that shirks on the personal cost. Therefore it is a theoretical puzzle how vengeance ever established itself in the repertoire of human motives, and how it sustains itself. Until the puzzle is solved, theory will offer no guidance on how to regulate vengeance to maximize its social value and to minimize its devastation.

2. Vengeance Solves Social Dilemmas

[5] See, for example, Fehr and Gaechter (1999).
[6] See, for example, Nowak and Sigmund (1998).
[7] See, for example, Hagen (1999).
[8] See Fehr and Gaechter (1999).

We begin our analysis the simplest possible social dilemma, written as a symmetric 2-player game (Table 1). Entries in Table 1 denote fitness payoffs to a player choosing the row (C = cooperate or D = defect when the partner chooses the column (again C or D). Thus the personal cost (fitness reduction) to choosing C rather than D is 1 and the social gain (own plus partner's payoff) is also 1. The game has a unique equilibrium: each player chooses the dominant strategy D and achieves fitness 0, thus missing the potential gains of 1 for each player.

[Table 1 about here]

We now add a punishment technology and a punishment motive, parameterized by its incurred cost v. We hypothesize that a player can inflict harm (fitness loss) h on the other player at personal fitness cost ch. The marginal cost c in (0,1) is a constant parameter that captures the technological opportunities for punishing others.

Also, inflicting harm h yields the player a utility bonus of $v \ln h$ (but no fitness bonus) when he is the victim of the sucker payoff (receiving -1, while the culprit receives +2) and no bonus in other circumstances. Thus the

motive is not spite[9] but rather is vengeance for damage personally experienced. The motivational parameter v is subject to evolutionary forces and captures an individual's temperament, e.g., his susceptibility to anger[10].

[Table 2 about here]

The objective function when victim of a sucker payoff now is v ln h – ch - 2. Then $h^*=v/c$ is the utility-maximizing degree of inflicted damage and $ch^*=v$ the incurred cost. The game now is as in Table 2. For $v > c$, the strategy D is not dominant. 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 fitness is higher for D and play converges to the inefficient (fitness 0) all-D equilibrium, as in the basic game. But for $s>s^*$ the expected fitness is higher for C and play converges to the efficient all-C equilibrium.

Hence, for $v > c$ we have a coordination game with two locally stable pure Nash equilibria and an unstable mixed Nash equilibrium at $s^*<1$

---

[9] See, for example, Levine (1998).
[10] See Frank (1988) for a detailed discussion of this idea.

(Figure 1). Thus the threat of vengeance can deter defection and support fully cooperative, socially efficient behavior (C,C) as a Nash equilibrium.

[Figure 1 about here]

3. **The Viability Problem**

The vengeance motive v itself is subject to evolutionary forces, perhaps slower than those determining the prevalence s of cooperation, but real forces nonetheless. The expected fitness of a cooperator is $W(C|\ s,\ v) = 2s-1-v(1-s)$, which is a strictly decreasing function of v for any fixed $s<1$. Only when there are no culprits ($s=1$) is the expected fitness independent of v. Thus the fitness of player v is weakly dominated by that of player v' whenever $0<v'<v$. Assuming that players occasionally encounter culprits, the vengeance preference parameter v will be driven towards 0 under any plausible evolutionary dynamics. We have a second-order free rider problem, and it seems that vengeance is not viable.

Standard solutions do not work well for this viability problem. Of course, the problem is attenuated for social creatures that form groups of

closely related individuals, such as slime molds (relatedness near $r=1$) or ants ($r=2/3$). But we are interested in humans, whose groups typically consist of individuals who are not that closely related (say $r=0.05$). If each individual's v were observable, then those with higher v might encounter D-play less frequently (as in Robert Frank's 1988 discussion) and thus maintain equal or higher fitness. This "greenbeard" solution[11] ignores the evolutionary pressure for lower v individuals to mimic the visible signs of higher v. As noted below, building a reputation is a bit complicated in our setting. Less standard solutions are discussed at greater length in our related work[12] and include a continuing stream of mutants (Sethi and Somanathan, 1996), ruling out intermediate values of v (Axelrod, 1986, Huck and Oechssler, 1999), and moralistic strategies of extended negative reciprocity (Boyd and Richerson, 1992).

4. Groups and Memes

During most of our evolutionary history, humans, like other social primates, presumably lived in small groups of individuals. In constructing our model, we assume that the typical person interacts every day with other

---

[11] This term is due to Boyd and Richerson (1998), and is a fanciful but striking example of identifiability.
[12] See Friedman and Singh (1999).

members of his or her group and also often interacts with others outside the group, but repeat encounters with any particular outsider are sporadic. In this setting, cooperation within the group can be maintained by many forms of reciprocation (positive or negative, direct or indirect) but cooperation outside the group is problematic: see Section 7 for discussion.

All known groups of humans maintain memes that prescribe appropriate behavior towards fellow group members and typically prescribe different appropriate behavior towards individuals outside the group[13]. The term "meme" as used here seems to have been introduced by Dawkins (1976). He defines it (Dawkins, 1982) as "the unit of information that is conveyed from one brain to another during cultural transmission." Durham (1991, pp. 188-90) illustrates the idea of memes with examples as diverse as marriage rules (which we would typically call social norms), and labeling boundaries for colors (at which wavelength of light does the color term change in different societies). Blackmore (1999) focuses on imitation as a mechanism for cultural transmission of memes but, as is implicit in our analysis, transmission mechanisms are richer than simple imitation.

---

[13] Sober and Wilson (1998), provide a catalogue of illustrations of this idea: we use some of their illustrations below.

The analysis below focuses on competition among memes that prescribe behavior towards culprits outside the group and towards group members who deviate from that prescription. These memes determine the group's reputation and therefore the fitness its members receive in encounters outside the group.

The success of the meme, as with any other adaptive unit, is measured by its ability to displace alternatives, i.e., by its fitness. There are many distinct mechanisms by which one meme may displace another, ranging from warfare to fashion, but for the most part these mechanisms align with the most fundamental mechanism, enhanced individual fitness. Without necessarily accepting assertions[14] that misalignments always are minor and temporary, our analysis will assume that a meme prescribing a particular degree of vengeance is fitter than existing alternatives when it brings higher individual fitness to group members on average.[15]

Groups affect individual fitness in several ways. As already noted, they provide gains from internal cooperation and (depending on the

---

[14] Wilson (1975) is a prominent proponent of this view.

[15] Note that we are putting aside the issue of redistribution within the group in this formulation. There is no alternative meme within the group: selection is therefore essentially at the group level, because all in the group have the same meme. For further discussion and numerical examples, see Sober and Wilson (1998).

reputation they carry) some gains from external cooperation. They also regulate access to scarce resources such as favorable home sites, stored food and marriage partners. Status or prestige within the group affects access: Sober and Wilson (1998) extensively document these kinds of patterns of norms. For example, here is their illustration from Jewish society in thirteenth-century Spain (itself taken from MacDonald, 1994):

> A law was then in place that a man who did not pay his taxes would have a blot placed on his genealogy...[A] marriage could not take place unless the genealogies of both the bride and the groom were consulted and approved. A blot on a person's genealogy could prevent a marriage. Thus, failure to pay taxes could have severe costs...Similar laws discouraged individuals from converting to other religions or even associating with members of other groups. (p. 145)

Another striking example provided by Sober and Wilson comes from Levine's (1965) study of Ethiopian culture: "The Amhara peasant who avenged an insult is 'following an ethic of cardinal importance'" (p. 182). Sober and Wilson discuss several other variants in norms across different societies.[16]

---

[16] Their source is the Human Relations Area File, a database developed by anthropologists to facilitate comparisons across cultures. It contains coded and indexed ethnographies of hundreds of different cultures.

5. Model Elements

The model has two parameters describing relevant memes:

- $v^n \geq 0$ is the group's normative vengeance level, the prescribed cost group members are supposed to incur when punishing outgroup culprits; and

- $a \geq 0$ is the group's tolerance of deviations from that norm. A deviation $x = v^n - v$ incurs a status loss $x^2/(2a)$ and, of course, corresponding status gains by others[17].

The relevant genes are summarized in one parameter:

- $v^{max} \geq 0$ is the maximum possible taste for vengeance that any meme could induce, given the individual's capacity for anger and his malleability.

Individual characteristics are described by:

- $v \in [0, v^{max}]$ is the actual vengeance cost an individual prefers; we assume it is learned from personal experience within the group, through fitness-maximizing best responses.

- $\bar{v} \in [0, v^{max}]$ is the group average degree of actual vengeance cost incurred in punishing outgroup culprits.

---

[17] This assumption can be relaxed, so that there is a net loss to the group. This generalization is outlined in the paragraph following the derivation of Result 1, in Section 7.

Finally, the impact of group reputation is captured in:

- $f(\bar{v}) = \exp(-\bar{v}/b)$ is the frequency with which an individual encounters culprits (who are all outgroup members, in equilibrium). It decreases in the group's reputation, summarized in $\bar{v}$, and increases in the positive parameter b representing the hostility of the environment.

6. Results

Result 1: The level of vengeance that maximizes fitness of group members is $v^o = \max\{0, b\text{-}2\}$. This level would be advantageous for the group to induce in its members, given the hostility b of the environment. The intuition is straightforward (see Methods). Under present assumptions, sanctions against deviators are zero sum within the group, so the group optimum simply trades off the increased fitness cost of higher v against the increased benefit, encountering fewer culprits. Section 7 shows that this tradeoff implies an optimum at v^0.

Of course, this result describes optimal rather than actual behavior and ignores the second-order free rider problem that no individual captures much of the benefit, which is dispersed throughout the group. To predict actual

behavior, we now consider a fixed meme summarized by v^n and a and examine how individual preferences will adapt.

Result 2: Individual adaptation drives v and $\bar{v}$ toward the individual optimum $v^* = v^n - a$, truncated to $[0, v^{max}]$. For given group characteristics $\bar{v}$ and v^n, the individual fitness cost is proportional to the sum of the direct cost v and the loss $(v^n - v)^2/(2a)$ from deviating from the group norm. The sum is minimized at $v^* = v^n - a$. Individual fitness is single peaked at v^*, so adaptive dynamics push the individual's parameter towards this optimum. The optimum is attained as long as the value is within the feasible range; otherwise v^* is truncated below at 0 and above at v^{max}. Assuming that individual adaptation is faster than either memetic or genetic evolution, v^* is a good approximation of actual preferences of each individual and an even better approximation of their average.

What is the relation between the group optimum v^o and the individual optimum v^*? The group meme, embodied in the parameters a and v^n, is subject to selective pressures in the medium run, and group average fitness is single-peaked. Any group whose memes bring actual behavior (near v^* by Result 2) closer to true optimum (v^o by Result 1) has a selective advantage.

Hence memetic evolution will drive actual behavior towards the optimum, as long as it is feasible.[18] But if v^o is infeasible (because the environment is so hostile that $v^o > v^{max}$), then there is a genetic selective advantage to increasing v^{max}.[19] Hence we have the following conclusion.

Result 3: Coevolution of memes and genes drives actual behavior v^* toward the group optimum v^o; in equilibrium $v^n = \max \{0, a+b -2\}$.

Thus the actual vengeance level is socially efficient in evolutionary equilibrium. Note that the memes, v^n, that support this efficient behavior are not the optimum value v^o but rather exaggerated versions, $v^n = v^o + a$.

On longer time scales there can be shifts in the environment b and in the punishment technology c. These shifts affect the encounter function f and hence the group optimum v^o. Our main conclusion implies that memes

---

[18] We do not formally model the dynamics of this process, but it is a straightforward exercise. The idea is that groups with memes that maximize group average fitness increase their population share at the expense of other groups. Sober and Wilson (1998) provide an example that they call a 'smoking gun' for cultural group selection in action. Drawing on the work of Kelly (1985), they summarize how the Nuer, an African pastoral group, rapidly expanded at the expense of the neighboring Dinka group prior to European colonization (pp. 186-191).

[19] Roughly, the mechanism would be as follows. Individuals with higher levels of v^{max} would stand out in the group, and have higher status. This will affect reproductive success. Obviously increases in v^{max} take place on a longer time scale.

(and, when necessary, genes) will adjust under selective pressure so that individual behavior $v*$ will track the new group optimum.

7.

Details and Generalizations

Here we explain how conclusions generalizing results 1-3 arise from less restrictive specifications than those used in the text, and spell out some details of the arguments. For concreteness we assume that the underlying social dilemma is still as in Table 1.

Even without imposing a group structure, one can write the expected fitness advantage to cooperating $A(p, u) = W(C)-W(D)$ as a function of the individual's probability estimate p that his partner will choose C and his expectation u of her vengeance parameter. One can derive p and u from a general specification of noisy observables, as discussed in Friedman and Singh (1999). Figure 2 shows how the fitter choice, C or D, depends on the sign of A, and how the choice shifts with the individual's own parameter v. One can derive[20] a smooth, decreasing encounter function $f(\bar{v})$ from the assumption that everyone chooses according to Figure 2. One can also

---

[20] See Friedman and Singh (1999).

develop a theory of group size and the reliability of group reputation from similar considerations.[21] The text shortcut these matters by implicitly obtaining p and u from a convenient but arbitrary encounter function $f(\bar{v})$ together with a given (high) level s of within group cooperation.

[Figure 2 about here]

The crucial feature of the group is that it imposes an expected fitness loss $\rho(x)$ when an individual deviates $x = v^n - v$ from the group norm v^n. The loss function ρ is assumed smooth, convex (with slope >1 for x sufficiently large) and minimized at 0, but it need not be quadratic or even symmetric.

7.1 Result 3 Remarks The text uses a direct argument from earlier results to establish the striking result that memes solve the second-order free rider problem and align actual behavior with optimal behavior. Here we highlight two underlying assumptions. The first is that there is a hierarchy of time scales, so that individual v (hence also $\bar{v}$) adapts most rapidly, then the memes v^n and/or a, then the genes v^{max}, and finally the environmental and technological parameters b and c. It does seem that v adapts rapidly to social

---

[21] See Friedman and Singh (1999).

memes; e.g., according to stories in the media, children raised in Belfast and Lebanon brought to the US for the summer have no problem adapting within a few months to the low v US norm, and then adapting back when they return.

It is natural to think of memes as evolving faster than genes, but that is not necessary for our result. Examples of coevolution on overlapping time scales include sickle-cell anemia in yam-growing areas, and lactose tolerance in herding communities[22]. The optimality result clearly fails when the environment or punishment technology changes faster than the memes (or, when v^{max} is too low, faster than the genes).

The other key assumption is that there is no fitness conflict between memes and genes, so memes are selected that bring higher fitness to group members. This assumption holds for memes transmitted vertically (e.g., from parents to children) because they share a common fate with genes and hence their fitnesses tend to coincide. However, horizontal meme transmission (e.g., from one teenager to another) need not respect genetic fitness. Here our assumption (and Result 3) may or may not hold. Our

---

[22] These examples are developed in Durham (1991).

reading of Sober and Wilson's wide-ranging discussion and other anthropological evidence supports both vertical and horizontal transmission of v^n and a, but we have not seen any examples of gene and meme conflict.

7.2 Result 2 Derivation With probability $f(\bar{v})$ a v–cooperator encounters a defector and receives fitness loss $(1+v+\rho(v^n-v))$, the sucker payoff plus the cost of wreaking vengeance plus the group status loss from departing from the norm. The individual receives a fitness gain of 1 in an encounter with a cooperator. Thus the individual's expected fitness is $W(v \mid \bar{v}, v^n) = 1(1 - f(\bar{v})) - (1+v+\rho(v^n-v)) f(\bar{v}) + R = 1 - f(\bar{v})(2+v+\rho(v^n-v)) + R$, where R is the base-level fitness including the (positive) effect on one's status from other group members' deviations from the norm v^n. The fitness function does not account for the possibility that the individual may sometimes play D, but this omission is harmless in terms of analyzing the adjustment of v.

Since the current choice of v has negligible effect on $\bar{v}$, the derivative of W is, up to the positive multiplicative constant $f(\bar{v})$, simply $\rho'(v^n-v) - 1$. The assumptions on ρ ensure that this derivative is decreasing in v and equal to zero at a unique $v^* > v^n$. Hence the single-peaked property holds and the argument in the text shows that individual adaptation pushes actual v

towards $v*$. With the quadratic specification for ρ we have the first order condition $0 = \rho'(v^n - v) - 1 = (v^n - v)/a - 1$, so in this case $v* = v^n - a$, truncated to $[0, v^{max}]$, as claimed.

7.3 Result 1 Derivation The text assumes that the group imposes ρ purely through status changes. In this case, R cancels the mean contribution of ρ, so the group average fitness is $W^g(\bar{v}) = 1(1 - f(\bar{v})) - (1 + \bar{v})f(\bar{v}) = 1 - f(\bar{v})(2 + \bar{v})$. The group optimum v^o maximizes this expression on $(0, v^{max}]$. Using $f(v) = \exp(-v/b)$, the first order condition reduces to $2 + v = -f/f' = b$, so v^o is $b - 2$, truncated to $(0, v^{max}]$ as claimed.

More generally, imposing a loss on a deviator may reduce the group's overall fitness by some fraction $t \in [0, 1]$; e.g., some potential gains to cooperation may not be realized. Group average fitness becomes $W^g(\bar{v}) = 1 - f(\bar{v})(2 + \bar{v} + t\rho(v^n - \bar{v}))$. With f and ρ as specified in the main text, direct computation yields $v^o = a(-t/2) + b(1 - t) - 2$. If $t = 0$, we have the case just analyzed, where v^o depends only on the environmental hostility parameter b. Higher t decreases the optimal level of vengeance and introduces positive dependence of v^o on the tolerance parameter a and the environmental hostility parameter b.

8. Discussion

Vengefulness, or a taste for negative reciprocity, is a crucial part of the human emotional repertoire. We model its importance in sustaining cooperative behavior, but highlight an intrinsic free-rider problem: the fitness benefits of vengeance are dispersed throughout the entire group but the fitness costs are borne personally. Evolutionary forces tend to unravel people's willingness to bear the personal cost of punishing culprits. The countervailing force that sustains vengeance is a group norm together with low-powered (and low-cost) group enforcement of the norm. Such memes coevolve with personal tastes and capacities so as to produce the optimal level of vengeance.

The underlying interaction in our model is the simplest possible social dilemma, but our methods extend easily to more complex interactions. It seems straightforward to redefine a culprit as one who harms any group member, not necessarily oneself. More generally, if the interaction took the form of a common pool resource or public goods game, a player would be considered a culprit to the extent that his contribution falls short by an amount e from the efficient level (or a normative level). The utility bonus

then could take the form $ve \ln h$, in which case the total harm to the culprit would be Ve/c, where V is the sum of the other participants' vengeance parameters. Results parallel to 1-3 above seem to follow.

Positive reciprocity could be analyzed directly in a similar fashion: preferences could include a utility gain (but no fitness gain) for rewarding a partner's cooperation, and a social norm could impose a fitness loss on deviators. However, since culprits are rare and cooperators are ubiquitous in successful society, the fitness cost of the rewards is excessive when relying entirely on positive reciprocity to sustain cooperation. Negative reciprocity greatly reduces the burden.

We can speculate how our model applies in different societies. The application to hunter-gatherer bands or villagers is clearest; here parameter b reflects directly the uncooperative tendencies of people from neighboring groups, and c reflects the opportunities to identify, track down and inflict harm on them. In more highly structured societies, vengeance is often exacted by delegated specialists (the criminal justice system). The marginal cost c of vengeance is lower, but is still positive, so the model remains valid.

Here an important shortcoming is that the model takes as exogenous the institutions that determine the vengeance technology.

What are the empirical implications of our model? Laboratory experiments can distinguish a taste for negative reciprocity from the egalitarian preferences hypothesized recently,[23] and generally favor our hypothesis.[24] New laboratory experiments can be conducted with a basic game similar to that in Table 1, together with a punishment technology similar to that used in the paper, $h = cv$. The punishment cost parameter c can be controlled at several levels. The impact of c on the frequency of punishment choices reveals the distribution of vengeance preferences v in a given subject population. Pilot experiment results, reported in Brownlow *et al* (2000), indicate that many UCSC undergraduate subjects have large positive values of this preference parameter.

Our model also has clear comparative statics predictions that in principle are testable with anthropological data: norms of vengeance and vengeful behavior should vary systematically with the hostility of the environment, the technology for harming culprits, and the technology for

---

[23] See, for example, Fehr and Gaechter (1999), Bolton and Ockenfels, (1999), and Fehr and Schmidt (1999).

[24] See also, Falk, Fehr, and Fischbacher (1999).

enforcing group norms. If the model is on the right track, one can hope that dysfunctional vengeful behavior (as in the Balkans) might improve in coming decades as the relevant memes evolve.

Table 1: Fitness with No Vengeance

Table 2: Fitness with Vengeance

References

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

Axelrod, R. An evolutionary approach to norms.R ean. American Political Science Review 80: 1095-1111

Blackmore, S (1999) The Meme Machine. Oxford, Oxford University Press (1986)

Bolton GE, Ockenfels A (1999) ERC: A theory of equity, reciprocity and fairness. Smeal College of Business Administration, Pennsylvania State University, http://silmaril.smeal.psu.edu: 80/lems/ERC.pdf

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

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

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

Brownlow A, Gullapalli R, Hoskins B, Ross M (2000) Game Theory: Introducing Vengeance – A Laboratory Experiment. UC Santa Cruz, Master's class term project

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

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

a.Stanford University Press, , CA

Falk A, Fehr E, Fischbacher U (1999) On the nature of fair behavior. Institute for Empirical Economic Research, University of Zurich, http://www.unizh.ch/iew/wp

Fehr E, Gaechter S (1999) Cooperation and punishment in public goods experiments. Institute for Empirical Economic Research, University of Zurich, http://www.unizh.ch/iew/wp

Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114: 817-868

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

Friedman D. & Singh, N. On the viability of vengeance. Department of Economics, University of California, Santa Cruz, http://zzyx.ucsc.edu/Econ/workpapers.html (1999)., N .http://zzyx.ucsc.edu/Econ/workpapers.html

Hagen WW (1999) The Balkans' lethal nationalisms. Foreign Affairs 78: 52-64

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

Kelly R (1985*)* The Nuer Conquest: The Structure and Development of an Expansionist System. University of Michigan Press, Ann Arbor

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

Levine DN (1965) Wax and Gold: Tradition and Innovation in Ethiopian Culture. University of Chicago Press, Chicago

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

Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. Nature 393: 573-577

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

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

Wilson EO (1975) Sociobiology: The New Synthesis. Belknap Press of Harvard University Press, Cambridge, MA

(v =0)	C	D
C	1	-1
D	2	0

Table 1: Fitness with No Vengeance

Table 2: Fitness with Vengeance

For $v > c$, the strategy D is not dominant. 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 fitness is higher for D and play converges to the inefficient (fitness 0) all-D equilibrium, as in the basic game. But for $s>s^*$ the expected fitness is higher for C and play converges to the effficient all-C equilibrium.

Figure 1: The Advantage of Cooperating.

The fitness advantage $A(s)=W(C)-W(D)=(v/c -1)s - (v+1)(1-s)$ is graphed as a function of the population fraction s playing C for two values of the vengeance parameter v. The graph of A rotates counterclockwise as v increases. The solid dot at s=1 represents a socially efficient, fully cooperative Nash equilibrium.
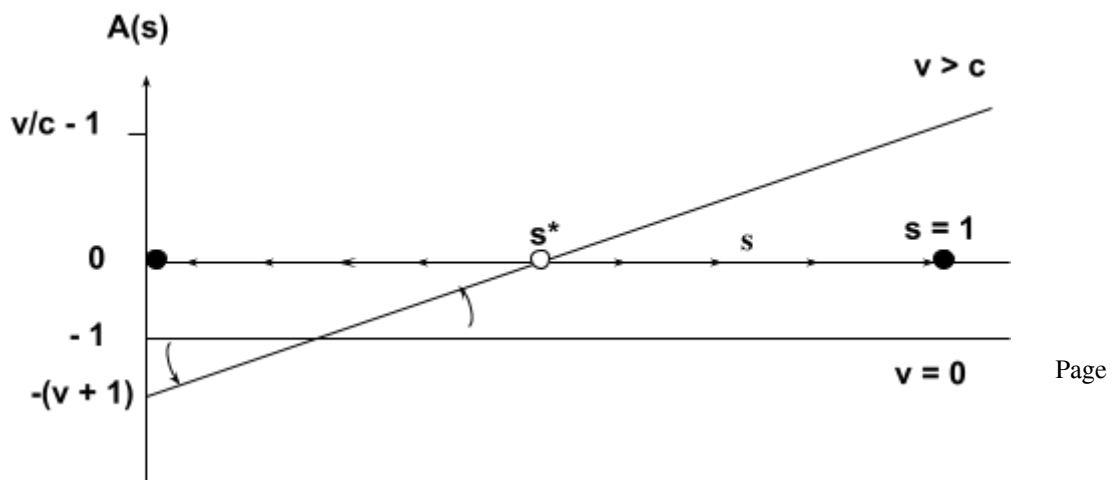
Figure 1: The Advantage of Cooperating.

Figure 2: The Decision Rule.

The appropriate choice of C or D is given by the sign of the advantage function A(p,u), where p is the probability that the partner will choose C and u is an unbiased estimate of her vengeance parameter. The A=0 locus shifts up with increases in thedecision maker's direct (v) or full ($\alpha = v + \rho(v^n - v)$)) vengeance cost.

Figure 2: The Decision Rule.