Equilibrium Vengeance

Daniel Friedman and Nirvikar Singh
Economics Department
University of California, Santa Cruz

April 2006

Abstract

Using a simple social dilemma game in extensive form, we illustrate the efficiency-enhancing role of the vengeance motive. Incorporating behavioral noise and observational noise leads to seven continuous families of (short run) Perfect Bayesian equilibria (PBE) that involve both vengeful and non-vengeful types. We then show that a new long run evolutionary equilibrium concept, Evolutionary Perfect Bayesian Equilibrium (EPBE), shrinks the equilibrium set to two points. In one EPBE, only the non-vengeful type survives and there are no mutual gains. In the other EPBE, both types survive and reap mutual gains.

[*] We are grateful to Matt McGinty for patient research assistance, and we benefitted from helpful comments of seminar participants at UC Berkeley, UC Riverside, UC Santa Cruz, USC and Yale, and readers of earlier versions of the paper. In particular, we thank Joshua Aizenman, Steve Goldman, Steffen Huck, Bentley MacLeod, Steve Morris, Matt Rabin, Donald Wittman, Huibin Yan, and Daniel Zizzo.

[†] Please email comments to dan@ucsc.edu.

1 Introduction

Craving vengeance is a powerful human motive: when some culprit harms you or your loved ones, you may choose to incur a substantial personal cost to harm him in return. There can be major economic and social consequences, positive and negative. Economic theory has not yet fully come to grips with such motives. In this paper we model vengeance as an emotional state dependent utility component (ESDUC) and investigate its efficiency impact and its viability.

A taste for vengeance, the desire to "get even," is so much a part of daily life (and the evening news) that it is easy to miss the evolutionary puzzle. We shall argue that indulging one's taste for vengeance in general reduces one's material payoff or fitness. Absent countervailing forces, the meek (less vengeful people) should have inherited the earth long ago, because they had higher fitness. Why then does vengeance persist?

To investigate the question, we introduce a new equilibrium concept,[1] evolutionary perfect Bayesian equilibrium (EPBE), that seems germane in a wide variety of applications. EPBE extends the equal profit condition of competitive markets into games of incomplete information with possible entry, exit and/or switching among multiple player types. Our paper uses EPBE to show how vengeance can persist despite its apparent fitness handicap.

Vengeance is closely tied to several vexing issues, methodological and substantive. Therefore we begin in Section 2 with a preliminary discussion on the nature of social dilemmas, the meaning of positive and negative reciprocity, why both are important to economists, and various modeling approaches. Section 3 presents the basic social dilemma as a simple extensive form game, and shows how vengeful preferences can dramatically improve equilibrium efficiency. It also spotlights the evolutionary problem when an individual's vengefulness cannot be perfectly known in advance and when behavioral errors are possible.

Section 4 derives seven continuous families of perfect Bayesian equilibria (PBE): two pooling equilibria, one separating equilibrium, two mixed equilibria and two hybrids. The PBE are short-run in that the nature and proportions of all player types are fixed. Section 5 examines the long-run in which the nature and proportions of types can evolve. We define EPBE and show that in our game it refines the equilibrium set from seven families down to two points: a unique EPBE that supports social gains (characterized in Proposition 2, our central result), and a trivial, inefficient EPBE (also in Proposition 2). A concluding section discusses generalizations and emergent issues. The Appendix collects the mathematical details.

[1] As noted in concluding discussion and in the Appendix, Abreu and Sethi (2003) independently use essentially the same concept in a particular bargaining model.

2 Preliminaries

An action has a social dimension when it affects non-actors as well as the actor. Figure 1 lays out the possibilities in terms of the net material benefit ($x > 0$) or cost ($x < 0$) to the actor, denoted "Self," and the net material benefit ($y > 0$) or cost ($y < 0$) to counterparties, denoted "Other". Economists think most often about the mutual gains quadrant I, where actions simultaneously benefit Self and Other. Such symbiotic actions increase social efficiency.

Quadrant IV is the well-studied opportunistic region, where Self benefits at Other's expense; the biological terms are parasitism and predation. The flip side is the altruism quadrant II, where Self bears a personal cost in order to benefit Other. Quadrant III is especially interesting to us. Cipolla (1976) refers to actions producing such outcomes as stupidity, but vengeance often will be a better explanation.

Social dilemmas arise from the fact that evolution directly supports behavior that benefits Self, i.e., outcomes $x > 0$ in quadrants IV (or I) but not $x < 0$ in II (or III), while in contrast, efficiency requires outcomes above the diagonal $[x + y = 0]$.[2] Social creatures (such as humans) thrive on devices that support outcomes in the half-quadrant II+ and discourage outcomes in IV-. Such devices somehow internalize Other's costs and benefits.

————fig 1 about here————

2.1 Efficiency-enhancing devices

Biologists emphasize the role of genetic relatedness.[3] If Other is related to Self to degree $r > 0$, then a positive fraction of Other's payoffs are internalized via "inclusive fitness" (Hamilton, 1964) and evolution favors outcomes above the line $[x + ry = 0]$. For example, the unusual genetics of insect order hymenoptera produce r up to $\frac{3}{4}$ between sisters, so it is no surprise that most social insects (including ants and bees) belong to this order and that the workers are sisters. For humans and most other species, r is only $\frac{1}{2}$ for full siblings and for parent and child, is $\frac{1}{8}$ for first cousins, and goes to zero exponentially for more distant relations. On average r is 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.

[2] More precisely, Self's iso-fitness curves are the vertical lines $x = C$ while iso-efficiency curves are diagonal lines $x + y = C$. The status quo point (0, 0) ensures that $C \geq 0$ is feasible.

[3] Bergstrom (2002) and Robson (2002) provide excellent summaries of recent biological insights into economic behavior.

Economists emphasize devices based on repeated interaction, as in the "folk theorem" (e.g., Fudenberg and Maskin, 1986). Suppose that Other returns the benefit ("positive reciprocity") with probability and delay together summarized in discount factor $\delta \in [0, 1)$. Then that fraction of other's payoffs are internalized (Trivers, 1971) and evolution favors behavior producing outcomes above the line $[x+\delta y = 0$This internalization 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, frequent and ongoing.[4] But humans specialize in exploiting once-off opportunities with a variety of different partners, in which case δ is small, as in the same steep dashed line. Other devices are needed to explain cooperation in such situations.

Here we will emphasize devices based on other-regarding preferences. For example, suppose Self gets a utility increment of ry from his or her action,[5] in addition to the material benefit x. Hence Self partially internalizes the material externality, and undertakes behavior that is above the line $[x + ry = 0]$. Friendly preferences, $r \in [0, 1]$, thus can explain the same range of behavior as genetic relatedness and repeated interaction. 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? This question motivates the present paper.

2.2 Modeling other regarding preferences

Two main modeling approaches can be distinguished in the recent literature. The distributional preferences approach[6] begins with a standard selfish utility function and adds additional terms capturing Self's response to how own payoff compares to Other's payoffs. The psychological games approach captures reciprocity by postulating that my preferences regarding your payoff depend on

[4] See Sethi and Somanathan (2003) for an extended discussion and survey of this device from the perspective of evolutionary game theory.

[5] Rilling et al (2002) present recent physiological evidence for such increments, based on fMRI brain scans of subjects playing prisoner's dilemma.

[6] This approach is exemplified in the Fehr and Schmidt (1999) inequality aversion model, the Bolton and Ockenfels (2000) mean preferring model, and the Charness and Rabin (2001) social maximin model.

my beliefs about your intentions.[7]

We favor a simple and direct approach, inspired by the pioneering work of Hirshleifer (1987) and Frank (1988). Model reciprocal preferences as state dependent: my attitude towards your payoffs depends on my emotional state, e.g., friendly or vengeful, and your behavior systematically alters my emotional state. Cox, Friedman and Gjerstad (2005) show that this emotional state dependent other-regarding utility component (ESDUC) approach is quite flexible and tractable.[8] Fortunately, a very simple rule suffices for present purposes: you become vengeful towards those who betray your trust, and otherwise have standard selfish preferences.

To be convincing, a model of other regarding preferences must account for the empirical data and also should pass the following theoretical test: people with the hypothesized preferences receive at least as much material payoff (or evolutionary fitness) as people with alternative preferences. This test is sometimes referred to as indirect evolution (Güth and Yaari, 1992) because evolution operates on preference parameters that determine behavior rather than directly on behavior.[9] Our task is to show that people whose utility functions contain a vengeful component will achieve in social interactions at least as much material payoff as other people whose utility functions contain only their own material payoff. Equally important, we want to show that a greater or lesser degree of vengefulness will not lead to higher material payoffs. Thus we respond to the first challenge raised by Samuelson (2001). The point is important here because many previous models of negative reciprocity are susceptible to unraveling: slightly lesser degrees of vengefulness have higher fitness. Our PBE and EPBE models also respond to Samuelson's other challenge, to consider issues of preference observability; this is made precise in the next section.

[7] Building on the Geanakoplos, Pearce and Stacchetti (1989) model, Rabin (1993) constructs reciprocity equilibria for two player normal form games, and Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (1998) extend this approach. Levine (1998) improves tractability by replacing beliefs about others' intentions by estimates of others' type.

[8] They present a fairly general specification of a utility bonus as a function of the kindness or unkindness of Other's choice as well as distributional and other status concerns. A psychological theory of how emotional states change (e.g., van Winden, 2001) rounds out this approach. In the present paper, Other has only two possible choices so a single parameter suffices.

[9] The idea goes back at least to Becker (1976) and Rubin and Paul (1979), and can be seen in many recent papers such as Huck and Oechssler (1999), Dekel, Ely and Yilankaya (1998), Ely and Yilankaya (2001), Kockesen, Ok and Sethi (2000), Possajennikov (2002), and Samuelson and Swinkels (2001). Many of these papers focus on positive reciprocity rather than negative reciprocity, or vengeance.

3 The Underlying Game

The first step in analyzing social preferences is to model explicitly the underlying social dilemma. We use a simple extensive form version of the prisoner's dilemma, or the holdup problem, also known as the Trust game (Güth and Kliemt, 1994). As shown in Panel A of Figure 2, Player 1 (Self) can opt out (N) and ensure payoffs normalized to zero for both players. Alternatively Self can trust (T) player 2 (Other) to cooperate (C), giving both players payoffs normalized to 1 and (assuming equal welfare weights) a social gain of 2. There is a social dilemma because 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. In the Appendix, we show how these specific payoff values can be generalized. The basic game has a unique Nash equilibrium found by backward induction: Self chooses N because Other would choose D if given the opportunity, and social gains are zero.

To this underlying game we add a punishment technology and a punishment motive as shown in Panel B. Self now has the last move and can inflict harm (payoff loss) h on Other at personal cost ch. The marginal cost parameter c captures the technological opportunities for punishing others.

————fig 2 about here————

Self's punishment motive is given by state dependent preferences. 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. The functional forms for punishment technology and motivation are convenient (we will see shortly that v parameterizes the incurred cost), but 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 as $U = x + vI_D \ln h$, that is, own material payoff x plus the relevant ESDUC. When facing a "culprit" (I_D =1), Self chooses the reduction h in Other's payoff so as 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 subgame perfect Nash equilibrium (T, C) yielding the efficient outcome (1, 1). The threat of vengeance rationalizes Other's cooperation and Self's trust.

3.1 Can vengeful preferences evolve?

Vengeance thus may have a pro-social role, but is it viable? Huck and Oechssler (1999), among others, show that vengeance can survive in small groups, where a vengeful person can impair others' fitness more than his own. In the present paper we are concerned with large populations, and here existing literature suggests an equivocal answer: Yes if Self's vengefulness is observable by Others, but No if it is not.[10]

To answer the question properly (Samuelson, 2001), we must consider intermediate cases, which we refer to as noisy perceptions. For the moment, assume Other perceives Self's vengeance level as $u = v + y$ when the true vengeance level is v. The error y has scale (e.g., standard deviation) $\sigma \geq 0$. We will soon see that moderate positive σ can help stabilize a unique positive level of vengeance.

Behavioral noise is also crucial. Self may intend to choose N but may twist an ankle and find himself depending on Other's cooperative behavior, and Other may intend to choose C but oversleeps or gets tied up in traffic. Such considerations can be summarized in a tremble rate $e \geq 0$. Larger values of e tend to raise Self's cost of vengefulness and reduce fitness.

A preliminary analysis of viability proceeds as follows. Fix noise levels $e \geq 0$ and $\sigma \geq 0$, and fix the marginal punishment cost $c > 0$. Assume that for a given distribution of v within the population, the choices of Self and Other adjust rapidly towards (short run) Nash equilibrium. The task is to compute Self's expected fitness or material payoff $w(v; \sigma, e)$ for each value of v at the relevant short run equilibrium.

First consider the case $\sigma = e = 0$, where v is perfectly perceived and behavior is noiseless. Recall that in this case the short run equilibrium (N, D) with payoff $w = 0$ prevails for $v \leq c$, and (T, C) with $w = 1$ prevails for $v \geq c$. Thus $w(v; 0, 0)$ is the unit step function at $v = c$, as in Figure 3.

————fig 3 about here————

With behavioral but no perceptual noise, $e > 0 = \sigma$, more vengeful types incur a greater cost when punishment is called for. Figure 3 shows that now Self's fitness function slopes downward at approximate rate $-e$. Finally, with perceptual noise also present, $\sigma > 0$, the sharp step at $v = c$ is smoothed out. The underlying calculations are collected in the Appendix.

[10] In particular, note that Other's inference from Self's move alone (typical in models of Bayesian equilibrium) is not enough for the kind of equilibrium we characterize; a direct perception of Self's type is needed.

4 Perfect Bayesian Equilibrium

Figure 3 shows two local fitness maxima for Self, one at $v = 0$ and the other at $v = v_H > c$, when σ and e are both small and positive. The fitness function w defines a one-dimensional landscape in which evolution pushes the evolving trait v uphill along the fitness gradient.[11] The figure therefore suggests that we will end up with some fraction x of the Self population with vengeance near $v_H > c$ and the remaining $(1-x)$ with vengeance near $v = 0$. The fractions represent the arbitrary portions of the population initially above and below the fitness minimum (near $c - \sigma$).

The implication is that we can focus on the game of incomplete information with two fixed types of Self players and with noisy perceptions. With two types one can streamline the analysis by focusing on the misperception probabilities rather than the entire error distribution, although this sacrifices some generality. Therefore we define perception as a binary variable s, with $s = 1$ denoting the perception that Self is vengeful, and $s = 0$ denoting the perception that Self is not vengeful. It is convenient (but not essential) to assume equal misperception probabilities and write $a = \Pr[s = 0|v = v_H] = \Pr[s = 1|v = 0]$.

————fig 4 about here————

Figure 4 shows the game tree. Nature chooses Self's true preference parameter as $v = 0$ (unvengeful) with probability $1 - x$, or as $v = v_H > c$ (vengeful) with probability x. Nature also independently chooses Other's perception as correct ($s = 0$ for $v = 0$, or $s = 1$ for $v = v_H$) with probability $1 - a$, or incorrect with probability $a \in [0, 1/2)$. Self knows her own preference but not the realized perception, and Other knows the perception but not the true preference.

Self's "pure" strategy set is denoted {NN, NT, TN and TT}, where XY means the unvengeful type tries to play X and the vengeful type tries to play Y. To spell this out, the space of mixed strategies is the unit square with corners at the pure strategies when there are no trembles. With trembles $e \geq 0$, Self's strategy space shrinks to the smaller square $[e, 1 - e] \times [e, 1 - e]$, and a corner strategy such as NT means that that N and T are actually played with probability $1 - e$ by respectively the unvengeful and vengeful type Self. Similarly, Other's "pure" strategy set is {DD, DC, CD and CC}, where now XY stands for the strategy 'play X if $s = 0$ and play Y if $s = 1$,.' Here "play" means to actually play with maximal probability $1 - e$. Thus Other's strategy space is also $[e, 1 - e] \times [e, 1 - e]$. The payoffs shown in Figure 3 are the same as in the reduced Trust game of Figure 1C.

The relevant equilibrium concept is perfect Bayesian equilibrium, PBE (e.g., Fudenberg and Tirole, 1991, chapter 8), suitably phrased to deal with large populations and explicit trembles.

[11] See for example Wright (1949), Eshel (1983) and Kauffman (1993).

PBE requires all players to optimize given beliefs, and requires that beliefs are Bayesian posterior probabilities obtained from perceptions, observed actions, and prior information on the type proportions.

What sort of PBE might exist? The first candidate is a separating equilibrium, call it SEP, in which Other plays DC and Self plays NT. Other prominent candidates are GP, the "good pooling" PBE in which Self plays TT and Other plays CC, and the "bad pooling " equilibrium BP = (NN, DD). In testing for any of these equilibria, the key conditions arise from Other's decision problem after a noisy perception. Other compares the expectation of the D payoff $2 - v/c$ to the C payoff 1. This comparison immediately leads to the rule: play D if $E(v|s) \leq c$, or play C if $E(v|s) \geq c$. To illustrate, consider an $s = 0$ perception when Self plays NT. The perception is erroneous precisely when a v_H type actually chooses T (i.e., doesn't tremble) and Other misperceives, which happens with probability $x(1-e)a$. The perception is correct precisely when a $v = 0$ type trembles to T and is correctly perceived, which happens with probability $(1-x)e(1-a)$. A straightforward Bayesian calculation now shows that the critical posterior expectation $E(v|s = 0) = c$ corresponds to prior probability (or population fraction) $x^s = 1/(1 + (\frac{a}{1-a})(\frac{1-e}{e})(\frac{v_H - c}{c}))$. Hence the rule states that Other should play D when observing $s = 0$ if $x \leq x^s$. Using the log odds function $L(y) = \ln(\frac{1-y}{y})$, this necessary condition for SEP can be rewritten $L(x) \geq L(x^s) = -L(a) + L(e) + L(c/v_H)$.

————table 1 about here————

Using Table 1, the reader can perform very similar calculations for other cases ($s = 1$ perceptions and Self strategies TT and NN) to obtain bounds on Other's best responses in terms of the population fractions x or their log odds. Combining them with straightforward computations of Self's best responses leads to necessary and sufficient conditions for the existence of the three pure strategy PBEs.

Straightforward computations show that Other strategy CD is dominated and that Self strategy TN is never a best response to Other's undominated strategies, so only strategies on the Northwest frontier of the strategy spaces need be considered in candidate PBEs. As shown in Figure 5, SEP and BP exist over overlapping ranges in the prevalence x of vengeful types, and there is a gap between these and the range where GP exists.

—————fig 5 about here——

There are also mixed PBEs.[12] The best response correspondences show that there is some mix $q^* \in [0, 1]$ of DC and CC that makes Self indifferent between TT and NT. We have a candidate mixed PBE if there is also some mix $t^*(x) \in [0, 1]$ of TT and NT that makes Other indifferent between DC and CC. It turns out that the profile $(t^*(x)TT + (1 - t^*(x))NT, q^*DC + (1 - q^*)CC)$

[12] We are indebted to Steve Morris for urging us to investigate these.

is a PBE, call it the Good Mix (GM), precisely when x lies in the gap.

Are there any other mixed PBEs? The same logic points to one other possibility. Consider $(u^*(x)NT + (1-u^*(x))NN, r^*DD + (1-r^*)DC)$, where $r^* \in [0,1]$ makes Self indifferent between NN and NT, and $u^*(x) \in [0,1]$ makes Other indifferent between DC and DD. This Bad Mix (BM), as we shall call it, turns out to be a PBE whenever x lies in the range overlap for the BP and the SEP PBEs.

The best response correspondences permit no other PBEs over a nontrivial range of x. They do produce other PBEs at two isolated points. When $L(x) = L(c/v_H) - L(a)$, both CC and DC are best responses to TT, and TT is a best response to qDC+ $(1-q)$CC as long as $q \in [0, q^*]$. Hence at this point, there is a continuum of PBEs, call them Good Hybrids, that vary only in Other's mixing probability q. Finally, where $L(x) = L(c/v_H) + L(e) + L(a)$, we have the Bad Hybrids (BH) (NT, rDD+ (1-r)DC) for $r \in [r^*, 1]$. Proposition 1 characterizes all the PBEs.

Proposition 1. Given perceptions with error rate a and choices with tremble rate e, and given types $v = 0$ and $v = v_H > c$ constituting respectively Self population fractions $(1-x)$ and $x \in (0,1)$, assume that $0 < a, e < 1/2$ and $\alpha = a + e - 2ae \leq 1/(2 + v_H)$. The complete set of PBE consists of:

1. the GP family (TT, CC) for $L(x) \leq L(c/v_H) - L(a)$;

2. the SEP family (NT, DC) for $L(c/v_H) + L(e) - L(a) \leq L(x) \leq L(c/v_H) + L(e) + L(a)$;

3. the BP family (NN, DD) for $L(x) \geq L(c/v_H) + L(a)$;

4. the GM family ($t^*(x)$TT+(1-$t^*(x)$)NT, q^*DC+(1-q^*)CC) for $L(c/v_H) - L(a) \leq L(x) \leq L(c/v_H) + L(e) - L(a)$;

5. the BM family ($u^*(x)$NT+(1-$u^*(x)$)NN, r^*DD+(1-r^*)DC) for $L(c/v_H) + L(a) \leq L(x) \leq L(c/v_H) + L(e) + L(a)$;

6. the GH family (TT, qDC+(1-q)CC), for $q \in [0, q^*]$, at the point where $L(x) = L(c/v_H) - L(a)$; and

7. the BH family (NT, rDD+(1-r)DC) for $r \in [r^*, 1]$, at the point where $L(x) = L(c/v_H) + L(e) + L(a)$.

The proof, in the Appendix, includes formulas for q^*, r^*, t^* and u^*.

A numerical example will help fix ideas. Set the marginal punishment cost at $c = 0.5$ and the vengeful type's preferred punishment expenditure at $v_H = 2.0$. Set the tremble rate at $e = 0.05$ and the misperception rate at $a = 0.10$. As shown in Figure 5, for sufficiently small proportions of the vengeful type ($L(x) \geq 3.30$ or $x \leq 0.036$) we have a Bad Pooling equilibrium: both types of Self try to opt out and Other tries to defect regardless of perception. For an overlapping range of vengeful type proportions ($L(x) \in [1.85, 6.24]$ or $x \in [0.002, 0.136]$) we have the Separating equilibrium. In the overlap $x \in [0.036, 0.136]$, there is also the Bad Mix PBE. No pure strategy PBE exists (just the Good Mix PBE, for which $q^* = 5/9$ and $t^*(x) \approx .35\frac{x}{1-x} - .06$) for higher values of x until we reach $x = 0.75$, after which point we have the Good Pooling equilibrium. The Good Hybrid equilibrium exists at x=0.75 for $q \in [0, 5/9]$, and the Bad Hybrid equilibrium exists at x=0.002 for $r \in [70/81, 1]$.

5 Evolutionary Perfect Bayesian Equilibrium

The numerical example spotlights an evolutionary problem. In the separating PBE, the vengeful type has higher fitness (0.418) than the unvengeful type (-0.036). Therefore, by the basic principle of evolution, the fraction x of vengeful types should increase. But the separating PBE disappears when x gets above .136. The same is true for the GM equilibrium: the vengeful types have fitness 0.665 while the unvengeful types have fitness 0, so again x should increase past the point (here 0.75) where the equilibrium disappears. However, when $x > .75$, we have only the GP equilibrium. Now the unvengeful type is fitter (0.855) than the vengeful type (0.760), so x should decrease until it falls below .75 and the GP equilibrium disappears. None of these equilibria seems stable in the long run.

————table 2 about here————

The evolutionary problem is not due to an unfortunate parameter choice in the numerical example. In the separating PBE, the vengeful type always achieves positive fitness; otherwise she would not try to play T. The unvengeful type always has negative fitness in this equilibrium because, with observational error rate $a < 1/2$, the payoff -1 is more frequent than +1. (See Table 2 for the general fitness expressions.) Hence evolutionary forces will always increase x in the separating PBE. In the GP PBE, the vengeful type always has lower fitness because of the extra cost $(1-e)v_H e$ of reacting to Other's trembles, so evolutionary forces will decrease x. Unvengeful types in the GM equilibrium always have fitness zero, and vengeful types always have nonnegative fitness. Once again, x will tend to increase until the equilibrium disappears. It seems that evolution undermines perfect Bayesian equilibrium.

5.1 Equal Fitness Principle

The problem is not due just to the peculiarities of our noisy trust game. Games of incomplete information generally have multiple types, and numerous mechanisms tend to increase the prevalence of high payoff types relative to low payoff types. For example, in an industry where firms with high quality products compete with those with low quality, one expects the market share of the less profitable type of firms to decrease over time because such firms expand less rapidly or exit, or switch types. As another example, a type of worker with lower full compensation (earnings, benefits and perks net of effort cost and opportunity cost) should become less prevalent due to earlier retirements, lower accession rates, etc.

The point is that payoffs should be equal across surviving types in long run equilibrium. In this context, PBE (or any standard refinement) is a short run equilibrium concept, while in the long run the types and their relative prevalence should adjust so that only those types with highest payoff remain. This is precisely the "survival of the fittest" principle of evolutionary theory. It is also the textbook distinction between short run and long run competitive equilibrium. The Appendix contains a formal definition; here we write out a verbal definition for long run equilibrium in extensive form games of incomplete information.

Definition. An evolutionary perfect Bayesian equilibrium (EPBE) is a PBE distribution over type-contingent strategy profiles such that in each population all types in the support of the distribution achieve equal and maximal expected fitness.

We now develop the EPBE concept specifically for our noisy trust game with endogenous values for the vengeful type v_H, its prevalence x and the perceptual error rate a. In EPBE, v_H maximizes fitness over an appropriate space of types, which we shall take to be the closed interval $[0, v^{\max}]$. The idea is that within broad limits, social (and perhaps genetic) forces shape Self's emotional response to violation of trust. We assume $v^{\max} > 0$ is large enough not to be a binding constraint; see Friedman and Singh (2004a) for a supporting discussion. In general one considers a distribution or measure over the space of types, but (for reasons discussed above in connection with Figure 2) the relevant distributions in the noisy trust game have support on at most two points, $v = 0$ and $v = v_H < v^{\max}$.

A fitness maximizing value $v_H > 0$ will be characterized by a marginal balance between two opposing effects. When vengefulness increases,

Perception Effect: Other is more likely to perceive $v > c$ (or $s = 1$), and hence is more likely to choose C, enhancing Self's fitness. However,

Cost Effect: when Other chooses D (either intentionally or via a tremble), Self will incur greater cost to punish him, reducing Self's fitness.

The cost effect can be derived from model elements used in the PBE analysis, but the perception effect cannot. Extremely vengeful types should be easier than slightly vengeful types to distinguish from $v = 0$ types, so the misperception probability a must be endogenized. For convenience we simply postulate $a = A(v)$,[13] where A is a smooth, positive and decreasing function, with $A(v) \to 0$ as $v \to \infty$ and $A(0) = 1/2$. Thus the types cannot be distinguished in the limit as the vengeful type becomes completely unvengeful, and can be distinguished perfectly in the limit as the vengefulness becomes extreme.

Crisp results require a parametric form for the perception technology A. Our choice is a simple Gaussian function with precision parameter $k > 0$,

$$A(v) = 0.5 \exp(-kv^2), \text{so} \;\; A' = -2kvA. \tag{1}$$

Besides the perception technology A (or precision parameter $k > 0$), we retain only two exogenous parameters: the marginal punishment cost parameter $c > 0$ and the tremble rate $e \in [0, 1/2)$. We continue to assume error symmetry for simplicity.

Characterizing EBPE for the noisy trust game comes down to the following conditions. Given the exogenous parameters, find endogenous values for a, v_H and x such that

1. There is a PBE strategy profile for the exogenous parameters c and e and the endogenous values a, v_H, and x.

2. The misperception rate is $a = A(v_H)$.

3. The preference parameter v_H maximizes Self's expected fitness given Other's PBE strategy. Formally,
$$v_H = \arg \max_{v \in [c, v^{\max}]} \{E_q W^S(v|A(v), e)\}, \tag{2}$$
where $E_q W^S$ is the maximal expected fitness Self can attain in the constrained strategy set $[e, 1-e]$, given Other's q-mixed strategy. The cost effect is captured in the argument v and the perception effect is captured in the conditioning variable $A(v)$.

4. If $0 < x < 1$ then the equal fitness principle implies that the unvengeful type Self achieves the same maximal expected fitness as the vengeful type, given Other's PBE strategy. Formally,
$$E_q W^S(0|A(0), e) = E_q W^S(v_H|A(v_H), e). \tag{3}$$

[13] In principle, one could derive A from the underlying distribution of perception errors y. In practice, estimating the error distribution is unlikely to be as useful as estimating A directly.

That is, the strategy mix q employed by Other must equalize payoffs between unvengeful and vengeful Selfs.

5. If $0 < q < 1$ then the equal fitness principle requires that both surviving types of Other achieve equal fitness, i.e.,

$$E_x W^O(CC|a,e) = E_x W^O(DC|a,e). \tag{4}$$

One must also check that the extinct types of Other (DD and CD, and also DC when $q = 0$ and CC when $q = 1$) achieve no higher fitness.

5.2 Results

What sort of PBE might survive the EPBE refinement? The discussion at the beginning of this section showed that the equal fitness principle fails for the SEP and GP families of PBE; specifically condition 4 fails except at GP with $x = 1$, where condition 3 fails. The discussion also showed that the GM family is an unlikely habitat for EPBEs; the Appendix rules it out.

Clearly the BP family contains a trivial EPBE. The family exists when vengeful types are so rare that Other always plays D, and so both types of Self play N. In this case the less vengeful are always fitter because trembles hurt them less. Hence the vengeful types become extinct, i.e., $x \to 0$, so the only possible BP candidate for EPBE is at the extreme, $x = 0$. It indeed is an EPBE: condition 1 holds because we are already working with a PBE, condition 3 holds because $v = 0$ uniquely maximizes Self's fitness, and conditions 2, 4 and 5 are moot. In this EPBE (except for double trembles) there are no mutual gains.

When might there be an efficient EPBE, one that supports mutual gains in the noisy trust game? The BM and BH families are unlikely habitats, again ruled out in the Appendix. The remaining family, Good Hybrid (GH), seems more promising because it allows both vengeful and unvengeful Selfs to achieve positive fitness, sometimes higher for the vengeful and sometimes higher for unvengful. The GH strategy profiles are (TT, qDC + (1-q)CC).

Our main result is that such an efficient EPBE does exist and is unique over a wide range of the exogenous parameters. The upper bound on the tremble rate is an increasing function $\hat{e}(k)$ of the precision parameter k, derived in the Appendix. This bound is approximately 0.23 (i.e., players might tremble a bit more than once in five tries) when $k = 0.5$ as in the unit Normal distribution, and it is about 0.13 for $k = 0.1$.

Proposition 2. Given marginal punishment cost $c \in (0, 1)$, behavioral error rate $e \in (0, \hat{e}(k))$, and perception technology (1) with precision parameter $k \in (0, 0.6)$, there is a unique efficient (Good Hybrid) EPBE whose characteristics (v_H, a, q, x) depend smoothly on the exogenous parameters.

There is only one other EPBE in the noisy trust game: the trivial (Bad Pooling) EPBE with proportion $x = 0$ of vengeful types. It exists for all perception technologies, all marginal punishment costs $c > 0$, and all behavioral error rates $e \in (0, 1/2)$.

The parameter k is bounded above by $\bar{k} \approx 0.612$; at higher values, the second order condition for v_H fails. A finite value of $v^{\max}$ creates a lower bound on k; for example $v < v^{\max} = 10$ implies $k > 0.028$. The proposition restricts parameter c to its natural interval (0,1). Higher values of c (for which Self's fitness reduction is larger than Other's) can tighten the upper bound on k due to the constraint $v_H > c$. For example, when $c = 2$ the upper bound is near $k = 0.3$.

The proof appears in the Appendix. It is constructive, and proceeds by writing explicit versions of the last three equations, solving them in terms of the exogenous parameters, and checking the relevant side conditions. It turns out that the equilibrium values $v_H = v^*(k)$ and $a = a^*(k)$ depend on k but are independent of e and c, while $q = Q(e, k)$ is independent of c, and $x = X(c, k)$ is independent of e.

Some of the comparative statics for the efficient EPBE are intuitive and others take a little explanation. The Appendix shows that $v^*(k)$ decreases in k. That is, as suggested by the perception effect, the equilibrium level of vengeance declines as perceptions become more precise. Perhaps surprisingly, a^* is increasing in k, that is, the equilibrium observational error rate goes up as the precision increases. It turns out that the indirect effect via $v^*(k)$ dominates the direct effect of k.

How about the probability with which Other attends to perceptions? $Q(e, k)$ increases in the tremble rate e as a consequence of the Self's indifference condition (3), and decreases in the precision of perceptions, k. Finally, the equilibrium fraction $X(c, k)$ of vengeful Selfs increases in the cost of punishment c as a consequence of the Other's indifference condition (4). However, the precision parameter k can have either a positive or negative effect on X depending on the level of c.

6 Discussion

We may summarize the argument as follows. Economists need to come to grips with human motives such as vengeance. Since vengeance generally reduces own material payoff or fitness, its persistence is an evolutionary puzzle. We therefore construct a model in which a taste for vengeance survives in a long run evolutionary equilibrium. The model uses ESDUCs, or emotional state dependent utility components, to represent such motives. The presence of ESDUCs is the proximate answer to the question of why individuals may want to harm (or help) others. However, the deeper questions of why certain ESDUCs exist and how they survive requires an analysis of their indirect fitness consequences. Studying vengeance is just one (interesting and complicated) application of the indirect evolutionary approach.

Our answer to the evolutionary puzzle proceeds in three stages. First, we construct a simple but representative situation in which ESDUCs matter, viz., a noisy version of the Trust game. Second, we compute all perfect Bayesian equilibria (PBE). We note that different types of individuals (vengeful or not) have different fitness in most PBE, leaving room for evolutionary pressures to operate. The third stage, therefore, is to introduce a new long run equilibrium concept called evolutionary PBE, which allows adjustment in the proportion of vengeful types, as well as the intensity of their vengefulness. We characterize the unique efficient EPBE for a wide domain of parameter values.

The conclusions are fairly robust within the context of the noisy Trust game. The argument can accommodate more general specifications of the payoffs, the perception technology, the punishment technology and preferences, and asymmetric perception errors. The Appendix shows that the expressions become much messier but the qualitative results are unchanged.

At least three open questions remain for the noisy Trust game. First, are the EPBE dynamically stable? The answer may depend the specific form of adjustment dynamics.[14] Friedman and Singh (2004a) suggests that short run dynamics enforcing PBE are entirely cultural (e.g., imitation or belief learning), and that the longer run dynamics enforcing EPBE also are mostly cultural (e.g. family moral codes) with some genetic components (e.g., capacity for anger). A relatively uncontroversial form of group selection (Wright's shifting balance) may promote convergence to the efficient EPBE. Adjustment dynamics surely are an important area for future work.

A second question concerns the trivial EPBE: how can one get a critical mass to escape it? Put more simply, in the context of the basic Trust game in Figure 1, how can one get $v > c$ starting from $v = 0$? Friedman and Singh (2004b) suggests a possible answer to this threshold problem. Subthreshold $v < c$ is not adaptive in a large population, but in small groups one can show that it works together with the discount factor δ to increase fitness. Thus positive values of v could get started in smaller groups and eventually become advantageous in larger groups.

Third, which remaining parameters can be endogenized? Keeping punishment technology c constant (or doing comparative statics exercises for c) seems to make sense. The tremble rate parameter e trades off trivially against the endogenous probability q that Other attends to the perception, as can be seen from equations (7) and (10) in the Appendix. However, there is every reason to take seriously the evolution of perception technology. A mutant Self with true vengeance

[14] Our current conjecture is that for all sensible dynamics the trivial (and inefficient) EPBE will have an open basin of attraction, and that the efficient EPBE will be stable in the same sense for some dynamics and for other sensible dynamics will be neutrally stable (e.g., will have a one dimensional stable manifold and a two dimensional center manifold).

parameter $v = 0$ who could somehow mimic the vengeful type would receive a major fitness boost. Friedman and Singh (2004b) refer to this possibility as the Viceroy problem, a reference to Monarch butterflies that correspond to vengeful types and their mimics known as Viceroys. That paper sketches an elaborate solution to the problem that involves interactions within and across small groups, but the issue remains open for large unstructured populations.

How might the ideas extend to more general classes of games? After all, people play many different social games, not just the noisy Trust game.[15] For example, consider the famous Ultimatum game. The first mover proposes a division of a fixed pie. A second mover with $v = 0$ will accept any proposal that gives him a positive payoff, but in most experiments the second mover often rejects small offers, giving both players zero payoff. Cox et al (2004) estimate parameters that translate to $v > 0$, but they do not consider equilibrium. It is reasonable to conjecture that a noisy version of the Ultimatum game supports two EPBE: a trivial EPBE with only $v = 0$ and greedy proposals, and an equitable EPBE with a mix of vengeful and unvengeful second movers and with more generous proposals.

We do not claim existence and uniqueness of nontrivial EPBE in great generality. In our Trust game (and also, it would seem, in a noisy Ultimatum game) the payoff ordering of vengeful and unvengeful types differs in different PBE, and this was the key to obtaining the nontrivial EPBE. We suspect that only trivial EPBE can exist when the same type in a given population has the highest payoff in every PBE, or when there are not enough margins for evolutionary adjustment. As for non-uniqueness, Abreu and Sethi (2003) obtain a continuum of EPBE in a bargaining model with wide classes of behavioral types. On the other hand, Friedman and Singh (2004a) study a simultaneous move social dilemma and obtain an efficient equilibrium, implicitly an EPBE with only one particular vengeful type. The questions of EPBE existence, uniqueness and efficiency remain open for general classes of games.

To conclude, the present paper combines two ideas, each of which we believe has widespread applicability independent of the other. Emotional state dependent utility components (ESDUCs) offer a tractable and flexible way to model other-regarding preferences, and can address many of the leading issues in behavioral economics. In particular, the vengeful components emphasized in the present paper may help give new insights into ”irrational” conflicts ranging from employment relations to international struggles. Friendly components likewise may give insight into behavior within the family and firm, and into the dynamics of charitable giving and social capital.

The second idea is evolutionary Perfect Bayesian equilibrium (EPBE). We wrote a general

[15] Our methods apply directly to any stable mix of games, and comparative statics apply to small one-time shifts in the mix. Large or continuing shifts in the mix would require a dynamic analysis.

verbal definition and worked it out explicitly for a particular (and not especially simple) game of incomplete information. The Appendix concludes with a more general definition and remarks. We believe that EPBE is an appropriate characterization of long run behavior when there are multiple potential "types" and some opportunity for entry, exit and/or switching among types. It endogenizes the set of types and their proportions, key variables that otherwise must be specified arbitrarily. Many games of incomplete information could be reconsidered in this light.

7 Appendix. Mathematical Details.

7.1 Proof of Proposition 1

Let $q^* = 1/(2-2a) \in (1/2, 1)$ and define $t^*(x) \in (0,1)$ as the solution to

$$(\frac{1-c/v_H}{c/v_H})(\frac{x}{1-x}) = t(\frac{1-a}{a}) + (1-t)(\frac{1-a}{a})(\frac{e}{1-e}). \tag{5}$$

Also, let $r^* = [1+v_H - e(2+v_H)]/[(1-a)(1-2e)(2+v_H)] = q^*(\frac{1+v_H}{2+v_H} - e)/(\frac{1}{2} - e) \in (q^*, 1)$ and define $u^*(x) \in (0,1)$ as the solution to

$$(\frac{c/v_H}{1-c/v_H})(\frac{1-x}{x}) = (1-u)(\frac{1-a}{a}) + u(\frac{1-a}{a})(\frac{1-e}{e}). \tag{6}$$

To check for all PBE we map out the best response correspondences (building in Bayesian updating) and look for mutually consistent profiles. We proceed stepwise.

Step 1. Confirm that CD is dominated. Recall from the payoff structure that Other is indifferent between C and D iff $E(v|s) = c$, and strictly prefers C (D) if $E(v|s) > (<) c$. It follows that DD dominates CD if $c > E(v|s=0)$, and that CC dominates CD if $c < E(v|s=1)$. At least one of these two cases always holds since $E(v|s=1) > E(v|s=0)$, establishing the claim. ♢

Step 2. The only undominated Self strategies in Figure 6A are those on the NW frontier. Any other strategy Y can be written as a convex combination of CD and a NW frontier strategy, since the set is the convex hull of its four corners, and the other three corners are contained in the NW frontier. But step 1 shows that CD and hence Y is dominated. ♢

————Figure 6 about here————

Step 3. TN is not a best response to any undominated strategy of Self. Direct computation shows that the best responses are as in Figure 6C: TT along the portion of the N frontier (convex combos of CC and DC) east of q^*, NT around the NW corner, and NN for the portion of the W frontier south of r^*.

To spell it out, we show explicitly that NT is always the best response to DC, as is TT to CC, given the hypotheses $0 < a, e < 1/2$ and $\alpha = a + e - 2ae \leq 1/(2+v_H)$. Given that Other will play

DC, Self with $v = v_H$ will face D with probability $\alpha = (1-e)a + e(1-a) = a + e - 2ae$ when playing T. A simple calculation shows that Self's expected payoff is nonnegative (and therefore she will indeed try to play T) as long as $\alpha \leq 1/(2+v)$, which holds by hypothesis. Self with $v = 0$ will face D with probability $1-\alpha$ when playing T; and she will avoid doing so as long as $\alpha \leq 1/(2+v) = 1/2$, a redundant condition. Hence NT is indeed the best reply to DC. If instead Other plays CC, the condition ensuring that Self indeed wants to play T is the same as before, taking $a = 0$, so it holds *a fortiori*. Hence TT is a best response to CC.

The q^*-mix of DC and CC that makes Self indifferent between NT and TT is precisely the mix that gives unvengeful Self zero expected payoff when actually choosing T, because the actual N payoff is also (always) 0. This condition is $0 = E_q W^S(T|v = 0) \equiv 1(1-\gamma) - 1\gamma$, or $\gamma = 1/2$, where γ is the probability that Other actually chooses D when unvengeful Self chooses T. There are three ways this can happen: Other ignores s but trembles, with probability $\gamma_1 = (1-q)e$; Other incorrectly perceives $s = 1$ and trembles, with probability $\gamma_2 = qae$; and Other correctly perceives $s = 0$ and doesn't tremble, with probability $\gamma_3 = q(1-a)(1-e)$. So the condition is $1/2 = \gamma = \gamma_1 + \gamma_2 + \gamma_3 = e + q(1-2e)(1-a)$. The solution is indeed $q^* = 1/(2-2a)$, and is clearly unique. It follows that NT (resp. TT) is Self's best response to qDC + (1-q)CC for $q \in [0,1]$ larger (resp. smaller) than q^*.

By a very similar argument, one verifies that $r^* = q^*(\frac{1+v_H}{2+v_H} - e)/(\frac{1}{2} - e)$ makes vengeful Self indifferent between N and T in response to r-mixes of DD and DC. (The condition is that vengeful Self obtains payoff zero from T, or $0 = -(1+v)[(1-e) - r(1-a)(1-2e)] + [1-(1-e) + r(1-a)(1-2e)]$, with root r^*.) Again, it follows that NN (resp. NT) is Self's best response to rDD + (1-r)DC for $r \in [0,1]$ larger (resp. smaller) than r^*. $\diamondsuit$

Thus any mutual best response involves only the NW frontier strategies, TT-NT-NN for Self and CC-DC-DD for Other. The proof will be complete after we check all combinations for mutual consistency.

To streamline notation, let $L_1 = L(c/v_H) - L(a)$; $L_2 = L(c/v_H) - L(a) + L(e)$; $L_3 = L(c/v_H) + L(a)$; and $L_4 = L(c/v_H) + L(a) + L(e)$. We have $L_i < L_{i+1}$ because $L(a)$ and $L(e)$ are positive for $a, e < 1/2$.

Step 4. Suppose first that Other perceives $s = 0$ when Self plays NT. Recall that in this case the perception is erroneous with probability $x(1-e)a$ and is correct with probability $(1-x)e(1-a)$. Hence by Bayes Theorem $c = E(v|s = 0) = v_H \Pr[v = v_H|s = 0] + 0 = v_H[x(1-e)a/(x(1-e)a + (1-x)e(1-a))]$. Cross-multiply, divide both sides by $ca(1-e)(1-x)$ and collect terms to obtain $\frac{1-x}{x} = (\frac{a}{1-a})(\frac{1-e}{e})(\frac{1-c/v_H}{c/v_H})$. Recall $L(y) = \ln(\frac{1-y}{y})$ for $y \in (0,1)$, so $\ln(\frac{y}{1-y}) = -L(y)$. Hence Other is indifferent after seeing $s = 0$ when $L(x) = -L(a) + L(e) + L(c/v_H)$, and prefers D when the prior odds $L(x)$ that $v = v_H$ are longer.When Self plays NT, Other will correctly perceive $s = 1$ with probability $x(1-e)(1-a)$ and incorrectly perceive it (i.e., when $v = 0$) with probability $(1-x)ea$. Algebra similar to the $s = 0$ case shows that $L(x) \leq L(c/v_H) + L(e) + L(a)$ now motivates Other to play C. Note that $L(a)$ is positive since $a < 1/2$. Hence DC is Other's best response to NT over the relevant x-range. If Self plays TT and $s = 0$, the expression $(1-x)e(1-a)$ for NT is replaced by $(1-x)(1-e)(1-a)$ in the Bayesian algebra, and the $1-e$ factors cancel. The usual cross multiplication and simplification now shows that Other wants to play C even when $s = 0$ iff $L(x) \leq L(c/v_H) - L(a)$. That is, CC here is a best response to TT.

These computations are summarized in the first two lines of Figure 6D. The top line indicates that the unique best response to TT (or to NN!) is CC for $L(x) < L_1$, is DC for $L_1 < L(x) < L_3$, and is DD for $L(x) > L_3$; at L_1 the best responses are CC and DC (and convex combinations), and at L_3 the best responses are DD and DC (and convex combinations). Likewise, the second line indicates that the unique best responses to NT are CC for $L(x) < L_2$, DC for $L_2 < L(x) < L_4$, and DD for $L(x) > L_4$; at L_2 the best responses are all convex combinations of CC and DC, and at L_4 they are all convex combinations of DD and DC.$\diamondsuit$

Step 5. We now examine every Self strategy that could be part of a PBE, find all best responses by Other, and check for mutual consistency. Begin with TT. For $L(x) < L_1$, the unique best response is CC, and TT is its best response, so we see that the GP family (TT, CC) is a PBE, and that there is no other candidate in this case. For $L(x) = L_1$, step 4 told us that the best responses to TT are convex combinations of DC and CC. Step 3 told us that TT is a best response to the convex combination qDC+ (1-q)CC iff $q \in [0, q^*]$, and is never a best response to DD or DC (or convex combinations). Hence the only PBE where Self plays TT and $L(x) = L_1$ consist of the $q \in [0, q^*]$ mixes, i.e., the GH family. When $L(x) > L_1$, the best response to TT is DD or DC, to which TT is never a best response. Hence there are no other PBE profiles where Self plays TT.$\diamondsuit$

Step 6. Now consider strict mixes of TT and NT. By Figure 6D, the unique best response is CC for $L(x) < L_1$, but NT can't be a best response to CC so no such PBE is possible here. For $L(x) > L_2$ the best responses are DC and DD, neither of which admits TT as a best response, so again no such PBE is possible. But for any x such that $L_1 \leq L(x) \leq L_2$ we can construct a unique PBE, which takes the Good Mix form ($t^*(x)$TT+(1-$t^*(x)$)NT, q^*DC+(1-q^*)CC). The construction proceeds as follows.

Recall from step 3 that only $q^* = 1/(2-2a)$ mixes of DC and CC allow strict mixes of TT and NT as best responses. Hence it suffices to find $t^*(x) \in (0, 1)$ such that convex combinations of DC and CC are best responses to the the t^* mix of TT and NT. That is, $t^*(x)$ makes Other indifferent between N and T when $s = 0$. The condition is $c = E(v|s = 0) = v_H\beta$, where β is the posterior probability that Self is vengeful. Thus $\beta = x(1-e)a/\eta$, where η is the unconditional probability of Other seeing $s = 0$, which now can happen in three different ways. The first way (also represented in the numerator) is that a vengeful Self doesn't tremble but is misperceived: $\eta_1 = x(1-e)a$. The second way is that an unvengeful Self tries to play T, doesn't tremble, and is correctly perceived: $\eta_2 = t(1-x)(1-e)(1-a)$. The last way is that an unvengeful Self tries to play N, trembles, and is correctly perceived: $\eta_3 = (1-t)(1-x)e(1-a)$. Hence $\eta = \eta_1 + \eta_2 + \eta_3$ and the condition (after cross-multiplying) is $\eta_1 v_H/c = \eta_1 + \eta_2 + \eta_3$. Collecting the η_1 terms and dividing through by $a(1-x)(1-e)$ we obtain equation (5). Observe that the RHS of (5) is strictly increasing in t since $0 < e < 1/2$. When $t = 0$ in (5) we obtain $L(x) = L_2$, and when $t = 1$ we obtain $L(x) = L_1$. Hence for intermediate values of x, which satisfy the given inequalities $L_1 \leq L(x) \leq L_2$ we obtain from (5) a unique $t^* \in [0,1]$ that makes Other indifferent between C and D when $s = 0$.♢

Step 7. Now consider pure NT. The argument has the same structure as step 5. We confirm that for x such that $L_2 \leq L(x) < L_4$, the SEP family (NT, DC) is the only PBE in which Self plays NT. For $L(x) = L_4$, the best responses to NT are convex combinations of DC and DD, and NT is a best response to rDD + (1-r)DC iff $r \in [r^*, 1]$. NT is never a best response to convex combinations of CC and DC. Hence we pick up the BH family. When Other best responds to NT with DD (as will happen if $L(x) > L_4$), then Self's best response is not NT but rather NN. Likewise, when Other best responds to NT with CC (as will happen if $L(x) < L_2$), then Self's best response is not NT but rather TT. In neither case can we have a PBE of the desired form.♢

Step 8. Now consider strict mixes of NN and NT. The argument has the same structure as step 6. Self will play such a mix in mutual best response only if $L_3 \leq L(x) \leq L_4$ and Other plays the r^* mix of DC and DD. The mix of NN and NT that allows Other to mix DC and DD must satisfy $c = E(v|s=1) = v_H(\kappa_1 + \kappa_2)/(\kappa_1 + \kappa_2 + \kappa_3)$, where $\kappa_1 = xu(1-a)(1-e)$ for correctly perceived vengeful Self not trembling, $\kappa_2 = x(1-u)(1-a)e$ for correctly perceived vengeful Self trembling, $\kappa_3 = (1-x)ae)$ for incorrectly perceived unvengeful Self trembling. The expression can be rewritten to obtain (6). Hence we obtain the BM family and no other PBEs.♢

Step 9. Finally consider pure NN. For $L(x) > L_3$, DD is the unique best response, and NN is of course the best response to DD, so we see that the BP family (NN, DD) is a PBE, and that there is no other candidate in this case. For $L(x) = L_3$, the only candidates are the BP and the extreme BM with $u = 0$; both are already picked up. When $L(x) < L_3$, the best response to NN is CC or DC, for which NN is never a best response. Hence there are no other PBE profiles where Self plays NN. ♢

Step 10. We picked up the GP and GH families at step 5, the GM family at step 6, the SEP and BH families at step 7, the BM family at step 8, and the BP family at step 9. Since we now have looked at all possible equilibrium profiles, the proof is complete.◇

7.2 Proof of Proposition 2 and comparative statics

For convenience, the derivations of comparative statics are included in this proof. The Proposition applies to a parameter domain defined by the functions $R(k) = (kv(2+v) - \frac{1}{2})\exp(-kv^2)$ and $\hat{e}(k) = \frac{R(k)}{2-2a+2R(k)}$. These are functions of the exogenous parameter k because in equilibrium v and a are specific functions (derived below) of k only.

The first and most laborious step in the proof is to derive v_H for a given k. Recall that the Good Hybrid strategy profile is (TT, qDC+(1-q)CC), so the probabilities in Table 1 give the fitness function $E_qW^S(v) = (1-e)[q(1-\alpha-(1+v)\alpha) + (1-q)((1-e)-(1+v)e)] = (1-e)[1-(2+v)e - qa(2+v)(1-2e)]$. The first order condition (FOC) $0 = dE_qW^S/dv$ for the maximization problem (2) simplifies slightly to $0 = -e - qA'(2+v)(1-2e) - qa(1-2e)$ or, separating variables,

$$[\frac{e}{1-2e}]q^{-1} = -(2+v)A' - a. \tag{7}$$

The second order condition is $(2+v)A'' + 2A' \geq 0$. Substituting in the A' expressions from (1), the FOC is

$$[\frac{e}{1-2e}]q^{-1} = [2kv(2+v)-1]a = (kv(2+v) - \frac{1}{2})\exp(-kv^2) \tag{8}$$

and the SOC is

$$kv^3 + 2kv^2 - \frac{3}{2}v - 1 \geq 0. \tag{9}$$

Equation (3) says that vengeful and unvengeful type Selfs coexist in the EPBE because they have equal fitness. Recall that $E_qW^S(v) = (1-e)[1-(2+v)e - qa(2+v)(1-2e)]$. Recall also that we are looking for an EPBE in which even the unvengeful try to play T, so $E_qW^S(0) = (1-e)[q(\alpha-(1-\alpha)) + (1-q)((1-e)-e)] = (1-e)[1-2e-q(2-2a-4e+4ae)]$. Thus (3) reduces to $ve = q[2(1-2\alpha) - av(1-2e)] = q(1-2e)[2-a(4+v)]$. Separating variables again, we obtain

$$[\frac{e}{1-2e}]q^{-1} = (2-a(4+v))/v. \tag{10}$$

Note that (8) and (10) have the same left hand side. Equating the right hand sides, we get $2kv(2+v)a - a = (2-4a)/v - a$ or

$$kv^3 + 2kv^2 + 2 = 2\exp(kv^2) = 1/a. \tag{11}$$

This equation holds trivially for $v = 0$ and $a = 1/2$, but we now show that it also implicitly defines a candidate equilibrium level of vengefulness $v^*(k) > 0$.

Lemma 1. Equation (11) has a unique positive solution $v^*(k)$ for any positive k. The solution $v^*(k)$ decreases in k over the range where the second order condition (9) is valid.

Proof of Lemma. At $v = 0$ both sides of (11) are equal to 2, and have equal slopes of 0. The LHS has slope $4kv(1 + \frac{3}{4}v)$ and the RHS has slope $4kv\exp(kv^2) = 4kv(1 + kv^2 + ...)$. For small positive v (up to approximately $v = \frac{3}{4k}$) the LHS has steeper slope but the reverse is true for larger v (indeed, the slope ratio tends towards ∞). Hence RHS = LHS at some $v \approx \frac{3}{4k}$ (with this approximation being better for larger k and smaller v), so (11) indeed has a unique positive solution $v^*(k)$ for any positive k.

Implicitly differentiate (11) to get

$$v^{*\prime}(k) = -[v^3 + 2v^2 - 2v^2\exp(kv^2)]/[3kv^2 + 4kv - 4kv\exp(kv^2)]. \tag{12}$$

Use (11) to substitute for the exponential term and rearrange to obtain

$$-kv^{*\prime}(k)/v = [kv^2 + 2kv - 1]/[2kv^2 + 4kv - 3]. \tag{13}$$

The RHS of (13) is $[g + \frac{1}{2}]/[2g]$ for $g(k) = kv^2 + 2kv - \frac{3}{2}$. Rewrite the second order condition (9) as $g \geq 1/v$, and since $v > 0$, we have $g > 0$. Hence the RHS of (13) is positive. Since v and k are also positive, we conclude from (13) that $v^{*\prime}(k) < 0$ when the SOC holds. $\diamondsuit$

We now show that the SOC (9) holds over the indicated range of k and is independent of the other exogenous parameters.

Lemma 2. Let $v = v^*(k)$ and $g(k) = kv^2 + 2kv - \frac{3}{2}$, and define $S(k) \equiv vg$. Then the equation $S(k) = 1$ has a unique solution $k = \bar{k} \approx 0.612$, and the second order condition (9) holds as an equality iff $k = \bar{k}$, and holds as a strict inequality iff $k \in (0, \bar{k})$.

Proof of Lemma. Write (9) as $S(k) \geq 1$. We first show that S strictly decreases in an open set U containing $S^{-1}[1, \infty)$. By direct computation we get $S'(k) = v^3 + 2v^2 + (v')(3kv^2 + 4kv - \frac{3}{2})$. Use (13) and simplify to write the RHS in the form $[vM]/[2kg]$, where v and k are positive and g is positive in U. The messy factor reduces to $M = -[(kv^2 + \frac{1}{2})g + \frac{3}{4}]$, which is strictly negative in U. Hence S indeed strictly decreases in U.

Use $v = O(1/k)$ from the proof of Lemma 1 to conclude that $S \to \infty$ as $k \to 0$ and $S \to 0$ as $k \to \infty$. Hence by the intermediate value theorem there is some $k \geq \varepsilon > 0$ such that $S(k) = 1$; let $\bar{k}$ be the smallest such k. We have $S'(\bar{k}) < 0$ and by the definition of U and continuity we have $S'(k) < 0 \; \forall k > \bar{k}$ s.t. $S(k) \geq 1 - \epsilon$. It follows that S is strictly bounded above by $1 - \epsilon$ on $(k + \delta, \infty)$. Therefore $\bar{k}$ is the unique solution to $S(k) = 1$ and the SOC fails for $k > \bar{k}$. Numerical methods give $\bar{k} \approx 0.612$. $\diamondsuit$

Equations (8), (10) and (11) together with Lemmas 1 and 2 show that $v_H = v^*(k)$ and $v = 0$ indeed both maximize Self's fitness, and that $v_H = v^*(k)$ has the indicated comparative statics. We still must find corresponding values of a, q and x; check their comparative statics; and verify the EPBE conditions.

The misperception rate is simply $a = a^*(k) \equiv A(v^*(k))$. To check its comparative statics, insert $v^*(k)$ into $A(v) = 0.5 \exp(-kv^2)$ and differentiate to get $\frac{da^*}{dk} = -(2kvv' + v^2)A$. Use (13) to get $2kvv' + v^2 = v^2/(3 - 4kv - 2kv^2) = -v^2/(2g) < 0$. Hence $\frac{da^*}{dk} > 0$, so indeed a increases in the precision parameter k.

Other's mixing probability q appears on the left hand side of both (8) or (10). Use the right hand side of (8) with $v = v^*(k)$ to get the desired function of k only, $R(k) \equiv (kv(2 + v) - \frac{1}{2}) \exp(-kv^2)$. Note that $R(k)$ has the same sign as $2kv^2 + 4kv - 1 = 2g + 2$, which is positive over $(0, \bar{k}]$. It therefore makes sense to rewrite (8) as

$$q = Q(e, k) \equiv \frac{e}{(1 - 2e)R(k)} > 0. \tag{14}$$

Inspection of (14) shows that Q is increasing in e. To show that $Q(e, k)$ is decreasing in k, use (14) to write $Q = \frac{e}{(1-2e)} \frac{\exp(kv^2)}{4(1+g)}$, differentiate and simplify using (13). Eventually one obtains $\partial Q/\partial k = \frac{ve}{(1-2e)} \frac{\exp(kv^2)}{8g(1+g)} [1 - vg - 2g]$. All factors are positive except $[1 - vg - 2g]$, which is negative because $vg > 1$ by the SOC and $2g > 0$, so indeed $\partial Q/\partial k < 0$.

The fraction x of vengeful Selfs comes from (4), which is the same PBE condition that defined $L(x) = L_1 \equiv L(c/v_H) - L(a)$. Hence

$$x = X(c, k) = L^{-1}(L_1) = \frac{1 - a}{1 + (\frac{v_H}{c} - 2)a}. \tag{15}$$

Conditions already imposed, viz., $v_H > c > 0$ and $0 < a < 1/2$ (or simply the domain of L), ensure that $0 < x < 1$. Since a and v_H are independent of c, inspection of (15) reveals that x is increasing in c. Simulations show that x can be increasing or decreasing in k, depending on the value of c.

The construction of (v_H, a, q, x) guarantees the last four of the five EPBE conditions listed at the end of section 5.1. The only remaining condition, the first, is that (TT, qDC+(1-q)DD) is a PBE. By Proposition 1, we need only check that $q \leq q^* \equiv 1/(2 - 2a)$. Use (14) and rearrange to obtain

$$e \leq \frac{R(k)}{2 - 2a + 2R(k)} \equiv \hat{e}(k). \tag{16}$$

Hence the hypothesis $e \in (0, \hat{e}(k))$ is sufficient, and we have verified the existence of a unique EPBE in the GH family.

The second paragraph of section 5.2 already verified the inefficient EPBE, $v_H = 0, a = 1/2, x = 0$ with strategy profile (NN,DD). The verification works for any values $a \in [0, 1/2)$, $c > 0$ and $k > 0$ of the exogenous parameters, and shows that there are no other candidate EPBE in the BP family. The discussion in section 5 also eliminated the GP and SEP families. We have just shown that there is only one EPBE in the GH family.

Is there an EPBE in the BH family (NT, rDD+(1-r)DC) for some $r \in [r^*, 1]$? To investigate, first note that the vengeful Self's payoff here is $E_rW^S(v) = (1-e)[r(e-(1+v)(1-e)) + (1-r)((1-\alpha)-(1+v)\alpha)]$=$(1-e)[1-(1-e)(2+v)+(1-r)a(1-2e)(2+v)]$. Hence

$$(1-e)^{-1}dE_rW^S/dv = -(1-e) + (1-r)(1-2e)\xi(v), \tag{17}$$

where $\xi(v) \equiv a + (2+v)A'$. Note that $\xi(v) < 1/2$ because $a < 1/2$ and $A' < 0$. Hence (17) is negative for all $r \in [r^*, 1]$, indeed, for all $r \in [0, 1]$. Since $E_rW^S(v)$ is decreasing in v, EPBE condition 3 cannot be satisfied, eliminating the BH family.

How about the BM family? It also requires that NT be a best response to rDD+(1-r)DC), for $r = r^*$. Hence the same argument also eliminates the possibility of an EPBE in this family.

The GM family is the last possibility, and its close relation to the GH family allows us to rule it out. The EPBE equal payoff condition (10) and the definition of $R(k)$ imply $\frac{e}{1-2e} = qR(k)$. Imposing the GM condition $q = q^* = 1/(2-2a)$ and rearranging yields the following necessary condition for an EPBE in the GM family:

$$e = \frac{R(k)}{2-2a+2R(k)} \equiv \hat{e}(k). \tag{18}$$

Hence the equal payoff condition fails within the relevant parameter domain $e < \hat{e}(k)$. ♢

Remark. The last part of the proof uses the fact that an EPBE in the GM family would have to satisfy a zero payoff condition (for the unvengful type, hence for the vengeful type Self playing T) as well as the conditions imposed in the efficient (GH) EPBE. The proof shows these conditions are not compatible in the relevant open set of exogenous parameters, but leaves open the possibility that they are compatible on the boundary $e = \hat{e}$.

7.3 Notes on more general models.

The basic payoffs can be normalized for each population so that without loss of generality the payoffs following action N are (0,0), and those following T then C are (1,1). A general Trust game has three restrictions on the payoffs (ς, τ) following T then D, namely $\varsigma < 0$, $\tau > 1$ and $\varsigma + \tau < 2$. If the choice $(-1, 2)$ used in the text is replaced by more general (ς, τ) satisfying these restrictions, then in subsequent analysis one must replace the condition $v > c$ by $v > c(\tau - 1)$ for Other's choices, and replace the condition $Pr[C] > 0.5$ by $Pr[C] > \varsigma/(\varsigma - 1)$ for Self's choices. It is tedious but straightforward to check that the PBE families still exist and that the key orderings of Self's payoffs across PBE still hold. Likewise, this holds for differing Type I and type II misperception probabilities.

Generalizing the perception technology A requires additional considerations. The maintained assumption is that A is a smooth, positive and decreasing function, with $A(v) \to 0$ as $v \to \infty$ and $A(0) = 1/2$. For such a function, existence of $v^* > 0$ is established as follows. First, impose the second order condition noted in the proof of Proposition 2 above, (i) $(2+v)A'' + 2A' \geq 0$. Next, note that equations (7) and (10) still have the same left hand side, so we can equate the right hand sides and rearrange to get the condition ($\sharp$) $4A - 2 = (2v + v^2)A'$. Since $A(0) = 0.5$, condition ($\sharp$) holds for $v = 0$. Since $A(v) \to 0$ as $v \to \infty$, the left hand side asymptotes to -2. If (ii) $(2v + v^2)A' \to 0$ as $v \to \infty$, then the right hand side has a larger asymptote. Hence, if for some smaller value the right hand side is smaller, i.e., if (iii) there exists $v > 0$ such that $(2v + v^2)A' < 4A - 2$, then by continuity condition ($\sharp$) must hold for some $v^* > 0$. Hence $v^* > 0$ exists if (i), (ii) and (iii) hold,[16] and it is easy to see that all three are satisfied by a wide variety of functions besides the Gaussian. For example with simple exponential $A(v) = 0.5\exp(-kv)$, condition ($\sharp$) has two positive roots for $k > 0$ sufficiently small, and (i-iii) hold for the larger root when $k \leq 0.2$. The conditions do have some bite, however. The perception technology $A(v) = 1/(2 + kv)$ is admissible for all positive k and ($\sharp$) has a positive solution iff $k < 1/2$. However, (i) fails for $k < 1$ so no efficient EPBE exists for this technology. Our tentative interpretation is that perception effect is inadequate when the tail is too fat, i.e., when the asymptotic error rate is $O(1/v)$.

7.4 Preliminary computation of Self's fitness.

Figure 2 indicates that Self's fitness has a local maximum at $v = 0$, a global minimum near $c - \sigma$ and a global maximum near $c + \sigma$. An argument supporting this conclusion is as follows. Let $e \geq 0$ be the behavioral noise amplitude as in the text. Assume that it is small, in particular that $e < 1/(2 + v^{\max})$, where $v^{\max}$ is a finite upper bound on the vengeance parameter. To define the observational noise amplitude $\sigma \geq 0$, let z be a continuous random variable located at zero (i.e., mean=mode=median=0) with an otherwise arbitrary density function $h(z)$ and cdf $H(z)$. For example, z could have a uniform or a Normal distribution. Other perceives not Self's true vengefulness v, but rather a noisy version $u = v + \sigma z$.

Key to the analysis is the probability $P(v)$ that Other will try to play D against Self with true

[16] One still has to check that other variables determined in equilibrium are in their feasible ranges, but the implied restrictions were redundant in the examples we checked.

parameter v, or equivalently, that c will exceed Other's posterior expectation of v. Computation is straightforward when Other has a uniform prior distribution for v. In this case, Other's posterior expectation of v is simply u and so $P(v) = \Pr[u < c|v] \equiv \Pr[\sigma z < c - v] = H(\frac{c-v}{\sigma})$. Then $P'(v) = -\sigma^{-1}h(\frac{c-v}{\sigma}) < 0$; its minimum is attained at $-\sigma^{-1}h(0)$ when $v = c$. Hence the inverse Mills ratio $P(v)/|P'(v)|$ attains a positive minimum of approximately $\sigma/(2h(0))$ near $v = c$. When Other has prior on v with positive density in the relevant neighborhood but otherwise arbitrary, the computation is much messier. It still can be shown that $P(v)/|P'(v)|$ attains a positive minimum of approximately $\kappa\sigma$ near $v = c$. (Now κ depends on the prior density as well as on h.) The approximations hold exactly in the limit as $\sigma \to 0$.

The preceding computation helps characterize the fitness function $w^S(v|\sigma, e)$. The probability that Other will actually play D (not just try) is $\alpha(v) = e + (1 - 2e)P(v)$, and so Self will achieve fitness $\beta(v) = 1 - (2 + v)\alpha(v)$ if she actually plays T and fitness 0 otherwise. Self will try to play N when $\beta(v) < 0$ and will try to play T when $\beta(v) \geq 0$. Thus

$$\begin{aligned} w^S(v|\sigma, e) &= e\beta(v) && \text{if } \beta(v) < 0, \\ &= (1 - e)\beta(v) && \text{otherwise.} \end{aligned}$$

When σ is small, $P(0) \approx 1$ and $\beta(0) \approx -(1 - 2e) < 0$, while for v moderately above c, $P(v) \approx 0$ and $\beta(v) \approx 1 - (2 + v)e > 0$. Suppose that β has two regular critical points, one near $c - \sigma$ and the other near $c + \sigma$. Since $\beta'(0) = -(1 - e) < 0$ we see that β and w^S slope downward from 0 to the first critical point, upward between the critical points, and downward beyond the second critical point. It follows that β is zero only at one point between the two critical points, and hence w^S indeed has the shape indicated in Figure 2.

Thus it remains only to verify the critical points. They are given by the first order condition (FOC) $0 = \beta' = 2\alpha + (2 + v)\alpha'$. After straightforward algebraic manipulation the FOC can be rewritten as $P(v)/|P'(v)| = v/2 + (1 - 3e)/(1 - 2e)$. The Right Hand Side (RHS) of this last expression has slope $+1/2$ and $v = 0$ intercept a bit below 1. As noted above, the LHS (the inverse Mills ratio) has a unique minimum near c and (since $P' \to 0$ as v moves away from c in either direction) and increases without bound on either side. The minimum value is of order σ so for σ sufficiently small there are exactly two regular solutions to the FOC and the verification is complete.

Figure 3 sketches w^S using the indicated values of e and σ, a uniform prior and the unit triangular density function for z.

7.5 A more general definition of EPBE

Equations (2-4) define EPBE for a particular PBE of a particular game. We now propose a more general definition of EPBE that may provide additional insight. To better connect with standard literature we use notation in this subsection that is not entirely consistent with the rest of the present paper.

Take as given a finite set of player populations $i = 1, ..., I$. In the model, $I = 2$ and $i = 1$ is called Self and and $i = 2$ is called Other. For each population i there is given a set Θ_i of possible types. In the model, $\Theta_1 = [0, v^{\max}]$ and $\Theta_2 = \{0, 1\}$. Let P be the set of feasible joint population distributions (or priors) $p = (p_1, ..., p_I)$ over $\Theta = \Theta_1 \times ... \times \Theta_I$. For given $p \in P$, the support of p is the smallest closed set in Θ containing population profiles with total mass 1. The set $S_i(p) \subset \Theta_i$ of surviving types in population i is the projection of the support of p onto Θ_i. In the model, $S_1(p) = \{0, v_H\}$ and $p_1(v) = 1 - x$ for $v = 0$ and $= x$ for $v = v_H$ but otherwise $= 0$, while $S_2(p) = \Theta_2$ and $p_2(0|v = v_H) = A(v_H) = p_2(1|v = 0)$. Thus for given perception technology A the set P of feasible distributions is a 2-dimensional set parametrized by $x \in [0, 1]$ and $v_H \in [0, v^{\max}]$, and each feasible distribution p has discrete support consisting of the four corners of a rectangle contained in $[0, v^{\max}] \times [0, 1]$.

Define actions and (partial) histories and, as in Fudenberg and Tirole (1991, p 331), use these to define (type-contingent mixed behavior) strategies σ_i and beliefs μ_i. Define the payoff function $u_i(\theta_i|p, \sigma)$ as the expected payoff for type θ_i when the type distribution is p and the strategy profile is $\sigma = (\sigma_1, ..., \sigma_I)$. Note that the payoff function is defined for all $\theta_i \in \Theta_i$, not just for $\theta_i \in S_i(p)$, because σ_i specifies an action mixture at every information set for every type $\theta_i \in \Theta_i$. The model used notation such as $E_q W^S(v_H|A(v_H), e)$ for the function u_1, with q and e referring to parameters of the strategy σ.

For distribution $p \in P$ let $\mathbf{PBE}(p)$ denote the set of PBE of the game just described, i.e., the set of pairs (σ, μ) that satisfy Definition 8.2 of Fudenberg and Tirole (1991 pp. 331-333, 349). As noted earlier, the definition says that beliefs μ are derived via Bayes theorem from the prior p and the observed partial histories, and the strategy profile σ employs only expected utility maximizing actions at each information set.

Definition. An evolutionary perfect Bayesian equilibrium is a triple (σ, μ, p) such that

1. $p \in P$ and $(\sigma, \mu) \in \mathbf{PBE}(p)$, and

2. for each population $i = 1, ..., I$ and each $\theta_i \in S_i(p)$, the payoff $u_i(\theta_i|p, \sigma) \geq u_i(\tilde{\theta}_i|p, \sigma)$ for all $\tilde{\theta}_i \in \Theta_i$.

Item (2) is the equal, maximal payoff property: equilibrium payoffs of each surviving type achieves equal and maximal payoff in each population. We close with a series of remarks on the nature of EPBE.

- The original evolutionary equilibrium concept (Maynard Smith and Price, 1973), ESS, is a static concept that applies to symmetric bimatrix games. Similarly, EPBE is a static equilibrium concept for extensive form games of incomplete information that leaves implicit the evolutionary dynamics.

- EPBE appears to be new. In the context of a bargaining game with incomplete information, Abreu and Sethi (2003) independently introduce essentially the same concept, and use it to examine the long run persistence of certain "irrational" types of bargainers. To the best of our knowledge, other papers that consider evolution in games of incomplete information allow arbitrary distributions of types that generally have different fitnesses. For example, Nöldeke and Samuelson (1997) and Jacobsen et al (2001) fix the proportions of two seller types (high quality and low quality) and model the evolution of buyer beliefs regarding the costly signals sent by sellers. Such analysis apparently applies to short or medium run equilibrium before the more profitable types can increase their market share.

- We regard EPBE as an appropriate concept for long run equilibrium whenever (a) material payoffs such as income or evolutionary fitness can be compared across types, and (b) adjustment mechanisms can affect existing types and their prevalence. Earlier definitions of evolutionary equilibrium might be interpreted as long run equilibria when the types are determined by last minute circumstance, and evolutionary selection applies to complete type-contingent strategies rather than to the types themselves.

- Appealing features of EPBE are that it endogenizes crucial variables and selects among multiple equilibria. One often has a multiplicity of PBE that depend rather sensitively on arbitrary exogenous specifications of the types and their distribution. EPBE can greatly reduce the equilibrium set while endogenizing the set of types and their distribution. In our noisy trust model, EPBE collapses seven continuous families of PBE to just two points.

References

[1] **Abreu, Dilip and Sethi, Rajiv.** "Evolutionary Stability in a Reputational Model of Bargaining." *Games and Economic Behavior.* August 2003, 44(2), pp. 195-216.

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

[3] **Bergstrom, Theodore C.** "Evolution of Social Behavior: Individual and Group Selection." *Journal of Economic Perspectives*, 2002, 16:2, pp. 67-88.

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

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

[6] **Cipolla, Carlo.** *The Basic Laws of Human Stupidity.* Bologna: The Mad Millers, 1976.

[7] **Cox, James C., Friedman, Daniel, and Gjerstad, Steven.** "A Tractable Model of Reciprocity and Fairness." Manuscript, University of California at Santa Cruz, 2005. http://econ.ucsc.edu/faculty/dan/Tractable.pdf

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

[9] **Dufwenberg, Martin and Kirchsteiger**, **Georg.** "A Theory of Sequential Reciprocity" *Games and Economic Behavior* 47:2 (May 2004) pp.268-298.

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

[11] **Eshel, Ilan**. "Evolutionary and Continuous Stability." *Journal of Theoretical Biology*, 103, pp. 99-111, 1983.

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

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

[14] **Frank, Robert.** *Passions within Reason: The Strategic Role of the Emotions*, New York: WW Norton, 1988.

[15] **Friedman, Daniel and Singh, Nirvikar** (2004a), ‘ ‘Negative Reciprocity: The Coevolution of Memes and Genes,” *Evolution and Human Behavior*, 25(3), pp. 155-173.

[16] **Friedman, Daniel and Singh, Nirvikar.** (2004b). “Vengefulness Evolves in Small Groups,” pp. 28-54 in Steffen Huck, ed., *Advances in Understanding Strategic Behavior*, Palgrave.

[17] **Fudenberg, Drew and Tirole, Jean.** *Game Theory*, Cambridge, MA: MIT Press, 1991.

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

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

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

[21] **Güth, Werner and Kliemt, Hartmut and Peleg, Bezalel.** “Co-evolution of Preferences and Information in Simple Games of Trust.” Manuscript, Humboldt University Berlin, 2001.

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

[23] **Hamilton, William D.** “The evolution of social behavior.” *Journal of Theoretical Biology*, 1964, 7, pp. 1-5.

[24] **Hirshleifer, Jack.** “On the Emotions as Guarantors of Threats and Promises,” p307-326 in J. Dupre, ed, *The Latest on the Best: Essays in Evolution and Optimality,* MIT Press, 1987.

[25] **Huck, Steffen and Oechssler, Jorg.** “The indirect evolutionary approach to explaining fair allocations.” *Games and Economic Behavior*, 1999, 28, pp. 13-24.

[26] **Jacobsen, Hans Jørgen, Jensen, Mogens and Sloth, Birgitte.** “Evolutionary Learning in Signalling Games.” *Games and Economic Behavior*, 2001, 34:1, pp. 34-63.

[27] **Kaufman, Stuart.** *The Origins of Order: Self-Organization and Selection in Evolution*, NY: Oxford U Press, 1993.

[28] **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.

[29] **Levine, David K.** "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1998, 1, pp. 593-622.

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

[31] **Nöldeke, Georg and Samuelson, Larry.** "A Dynamic Model of Equilibrium Selection in Signaling Markets." *Journal of Economic Theory*, 1997, 73, pp. 118-156.

[32] **Possajennikov, Alex.** (2002) 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

[33] **Rabin, Matthew.** "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 1993, 83, pp. 1281-1302.

[34] **Robson, Arthur J..** "Evolution and Human Nature." *Journal of Economic Perspectives*, 2002, 16:2, pp. 89-106.

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

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

[37] **Samuelson, Larry and Swinkels, Jeroen.** "Information and the Evolution of the Utility Function." Mimeo, University of Wisconsin, 2001.

[38] **Samuelson, Larry.** "Introduction to the Evolution of Preferences." *Journal of Economic Theory*, 2001, 97, pp. 225-230.

[39] **Sethi, Rajiv and Somanathan, E.** "Understanding reciprocity." *Journal of Economic Behavior and Organization*, 2003, 50, pp. 1-27.

[40] **Sobel, Joel.** "Social Preferences and Reciprocity." Mimeo, University of California at San Diego, 2000.

[41] **Trivers, Robert.** "The Evolution of Reciprocal Altruism." *Quarterly Review of Biology*, 1971, 46, pp. 35-58.

[42] **van Winden, Frans.** "Emotional Hazard Exemplified by Taxation-induced Anger" *Kyklos*, 2001, 54, pp. 491-506.

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