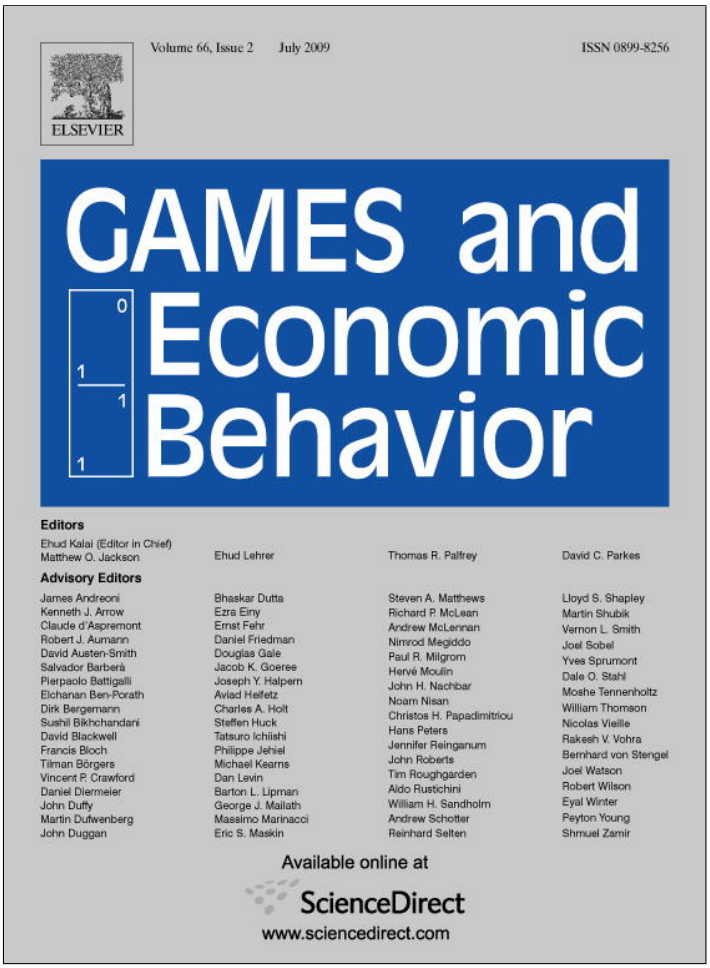


Games and Economic Behavior 66 (2009) 813–829

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Equilibrium vengeance ☆

Daniel Friedman, Nirvikar Singh *

Economics Department, University of California, Santa Cruz, CA, USA

ARTICLE INFO

Article history:
Received 25 June 2007
Available online 30 October 2008

JEL classification:
C73
Z13

Keywords:
Reciprocity
Vengeance
Evolutionary perfect Bayesian equilibrium
Social dilemmas

ABSTRACT

The efficiency-enhancing role of the vengeance motive is illustrated in a simple social dilemma game in extensive form. Incorporating behavioral noise and observational noise in random interactions in large groups leads to seven continuous families of (short run) Perfect Bayesian equilibria (PBE) that involve both vengeful and non-vengeful types. 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.

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 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. Our contribution to this literature is to demonstrate the

☆ We are grateful to Matt McGinty and Jean Paul Rabanal for excellent research assistance, and we benefited 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, Martin Dufwenberg, Steve Goldman, Steffen Huck, Bentley MacLeod, Steve Morris, Matt Rabin, Donald Wittman, Huibin Yan, Daniel Zizzo, and three anonymous referees of this journal.

* Corresponding author.
E-mail address: boxjenk@ucsc.edu (N. Singh).

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

0899-8256/$ – see front matter
doi:10.1016/j.geb.2008.10.006

viability of a taste for vengeance even when people interact in unstructured large groups, and the degree of vengefulness is continuously variable and imperfectly observed.

Section 3 presents the basic social dilemma as a simple extensive form game, and shows how vengeful preferences can dramatically improve equilibrium efficiency. It spotlights the evolutionary problem when an individual's vengefulness cannot be perfectly known in advance and when behavioral errors are possible. Finally, it argues for a simplification of the analysis: of all possible distributions in continuous type space, it suffices to look only at those supported on just two points.

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. Appendix A collects the mathematical details.

2. Preliminaries

Consider an actor, denoted "Self," whose choices also affect counterparties, collectively denoted "Other." Sometimes actions are available that simultaneously benefit Self and Other, increasing social efficiency. Alternatively, available actions may be opportunistic, benefitting Self at Other's expense, or may be altruistic, requiring Self to bear a personal cost in order to benefit Other. Of course, some actions may harm both Self and Other, as in costly vengeance. Social dilemmas arise from the fact that evolution directly supports behavior that benefits Self, while in contrast, efficiency involves net gains to Other as well as Self. Social creatures (such as humans) thrive on devices that encourage efficient altruism and discourage inefficient opportunism by somehow internalizing Other's costs and benefits.[2]

2.1. Efficiency-enhancing devices

Biologists emphasize the device of genetic relatedness,[3] but many human interactions involve non-closely related individuals. Economists emphasize devices based on repeated interaction, as in the "folk theorem" (e.g., Fudenberg and Maskin, 1986). These can support socially efficient behavior 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.

Here we will emphasize devices based on other-regarding preferences. For example, suppose that in addition to any personal material benefit, Self gets a utility increment from actions that materially benefit others.[5] Hence Self partially internalizes any material externality. Such friendly preferences 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 lesser internalization will tend to make more personally advantageous choices, gain higher fitness, and displace the more friendly types. Vengeful preferences rescue friendly preferences if Others punish Self for being insufficiently friendly.[6] But punishment is 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[7] 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 my beliefs about your intentions.[8]

We favor an alternative 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 et al. (2007) show that a 3 or 4 parameter model

[2] See Friedman and Singh (2004b) for a more detailed discussion.

[3] Bergstrom (2002) and Robson (2002) provide excellent summaries of this approach. Henrich (2004) and subsequent articles in the special issue of Journal of Economic Behavior and Organization (2004) dissect the biological and cultural basis of human cooperation. Henrich also notes the limited scope of standard inclusive fitness and folk theorem arguments, and emphasizes the role of structured interactions, or modern group selection.

[4] See Sethi and Somanathan (2003) for an extended discussion and survey of this device from the perspective of evolutionary game theory. Our own approach may be seen as complementary to the analysis of repeated interactions, covering different circumstances.

[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 idea is developed in the altruistic punishments literature (e.g., Fehr and Gachter, 2002; Boyd et al., 2003; Dufwenberg et al., 2008). As emphasized in Henrich (2004), the analysis relies on structured populations and group selection.

[7] 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 (2002) social maximin model.

[8] Building on the Geanakopolis et al. (1989) model, Rabin (1993) constructs reciprocity equilibria for two player normal form games, and Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) adapt the idea to extensive form games. Levine (1998) provides a more standard game theoretic alternative by replacing beliefs about others' intentions with estimates of others' types.

incorporating this approach accounts well for existing laboratory data.[9] 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 referred to as indirect evolution (Güth and Yaari, 1992; precursors include Becker, 1976, and Rubin and Paul, 1979) because evolution operates on preference parameters that determine behavior rather than directly on behavior.

The evolutionary test is prominent in a number of recent papers,[10] several of which bear directly on present concerns. Huck and Oechssler (1999) show that vengeance can survive in small groups, where a vengeful person can impair others' fitness more than his own. Herold (2004) shows that positive as well as negative reciprocity (vengeance) preferences can survive in a "haystack" model, in which people interact in small groups that are occasionally remixed. Our concern, however, is with large unstructured populations.

In his introduction to the literature, Samuelson (2001) poses two challenges: can the hypothesized preferences emerge when there is a full range of alternatives, not just a handful, and can they emerge when they are not perfectly observable? The first challenge is especially acute here because many previous models of negative reciprocity are susceptible to unraveling: slightly lesser degrees of vengefulness have higher fitness. Heifetz et al. (2007) is an important exception. They show in great generality that, given observability, some distortion of selfish preferences (what they call 'dispositions') generically survives evolutionary pressure. They also consider imperfect observability, but only in a special case that assumes that Bayesian Nash Equilibria are locally unique and that perceptual errors are uniformly bounded, neither of which holds in the model we present below. Bohnet et al. (2001, Appendix A) sketch a model with imperfect type observability, but again it covers only a special case that lies outside our present concerns. Like us, Güth et al. (2000) examine observability in a trust game, but it takes a quite different form. They analyze the fraction of players with "moral" dispositions and the fraction adopting a costly but perfect observation technology. As Samuelson notes, evolution favors those who can appear more committed (i.e., vengeful in our context) than they really are, so perfect observability seems unrealistic.

To summarize, within existing literature on other regarding preferences, our paper is distinguished by its focus on: (a) vengeance, a contingent preference for costly punishment. The majority of papers focus on noncontingent preferences (or dispositions), or on positive reciprocity. (b) Evolution in large unstructured populations, a more challenging and general environment. (c) Imperfect observability, responding to Samuelson's second challenge. Players never know for sure how vengeful their partners might be. (d) A continuum of types. In response to Samuelson's first challenge, we want to show that a greater or lesser degree of vengefulness will not lead to higher material payoffs.

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 Fig. 1, 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 Appendix A, we show how these specific payoff values can be generalized. The basic game has a unique Nash equilibrium outcome and unique subgame perfect 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.

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 $U = y + vI_D \ln h$, that is, own material payoff y plus the relevant emotional state component. 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

[9] A psychological theory of how emotional states change (e.g., van Winden, 2001) rounds out this approach; see also Gintis (2002) and Sobel (2005).

[10] For example, Güth (1995), Dekel et al. (1998), Ely and Yilankaya (2001), Kockesen et al. (2000), Ok and Vega-Redondo (2001), Samuelson and Swinkels (2006), and Possajennikov (2002).

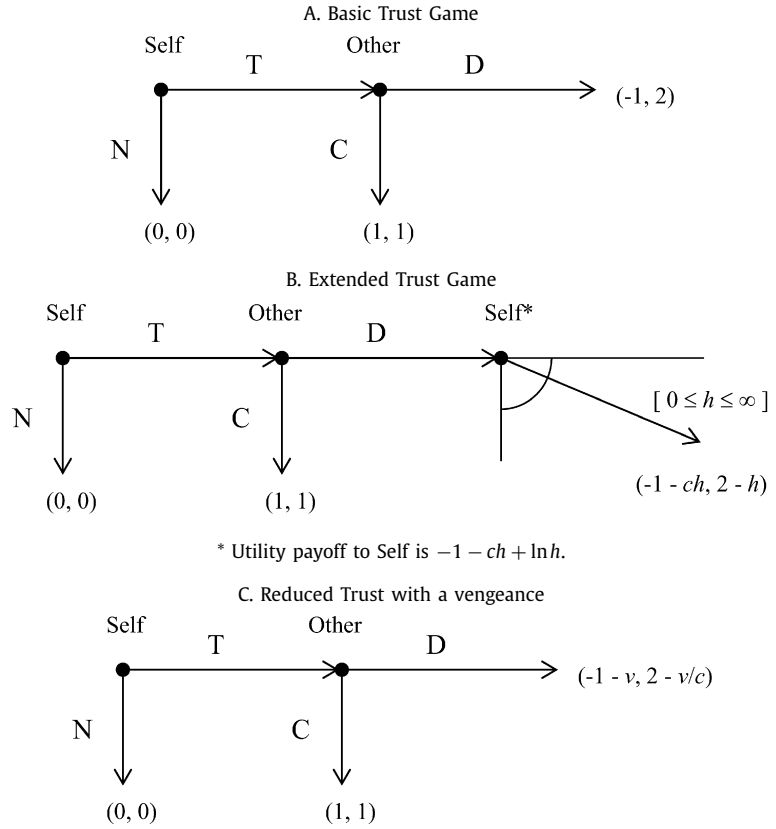


Fig. 1. Fitness payoffs.

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?

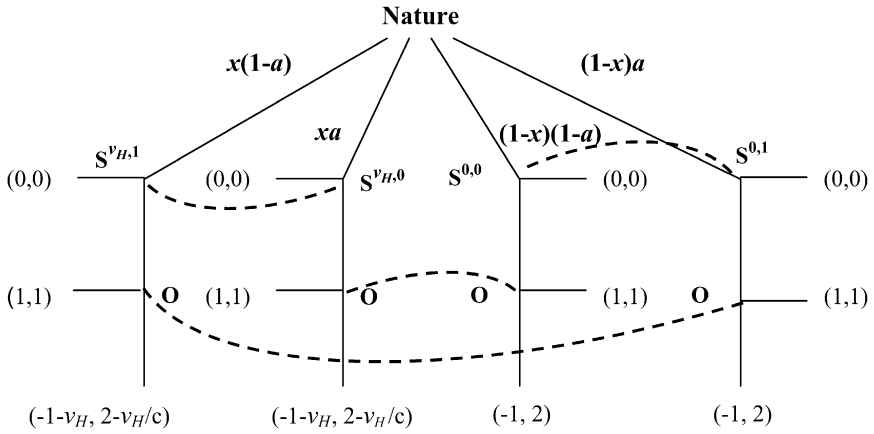
Vengeance thus may have a pro-social role, but is it viable? To answer the question properly (Samuelson, 2001), we must consider imperfect observability, which we refer to as noisy perceptions. Behavioral noise is also crucial in ensuring the evolutionary viability of vengeance, because it confounds the inference of type from behavior. 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 $e \geqslant 0$. Larger values of e tend to raise Self's cost of vengefulness and reduce fitness.

When there is behavioral noise as well as perceptual noise, while the vengeance parameter may take on a continuum of possible values, an argument based on fitness landscape dynamics (Friedman and Singh, 2007)[11] justifies the claim that equilibrium distributions will have support on just two points, one at $v = 0$ and the other at some specific $v_H > c$, fixed in the short run but variable in the long run. 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 $v = 0$. With just two types, we can streamline the short run analysis (at a slight loss of generality) by focusing on the misperception probabilities rather than the entire error distribution. 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 $a = \Pr[s = 0|v = v_H] = \Pr[s = 1|v = 0]$.

4. Perfect Bayesian equilibrium

Fig. 2 shows the game tree based on the assumptions made so far. 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

[11] As its title suggests, the present paper does not specify evolutionary dynamics. However, dynamical intuition will help motivate the formal definition of EPBE presented below. We therefore note that continuous movement up the fitness gradient has well established antecedents, e.g., Wright (1949), Eshel (1983) and Kaufman (1993). Such landscape dynamics (Friedman and Ostrov, 2008) are quite distinct conceptually and formally from mutations in a discrete type space and from replicator-type dynamics.



Note: *O* denotes Other; S^{ij} denotes Self with vengeance level i and perception j, as determined by Nature's move. The four branch labels are Nature's move probabilities.

Fig. 2. Game tree.

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 $e \geqslant 0$, 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 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 Fig. 2 are the same as in the reduced Trust game of Fig. 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. 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 the posterior expectation $E(v|s) \leqslant c$, or play C if $E(v|s) \geqslant c$. 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 $x \leqslant x^s$. Using the log odds function $L(y) = \ln(\frac{1-y}{y})$, this necessary condition for SEP can be rewritten $L(x) \geqslant L(x^s) = -L(a) + L(e) + L(c/v_H)$.

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.[12] As shown in Fig. 3, 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.

There are also mixed PBEs.[13] 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

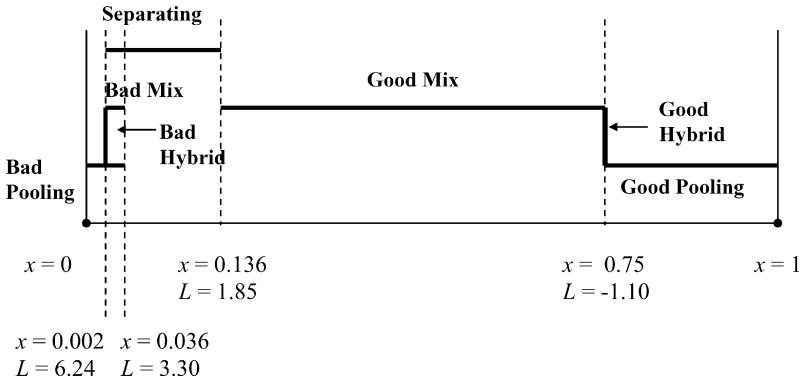
[12] Fig. 4 in Appendix A shows the implications geometrically: only strategies on the Northwest frontier of the strategy spaces need be considered in candidate PBEs.

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

Table 1
PBE probabilities.

	Choice	Fitness payoff	Equilibrium probability		
		Self, Other	(NT, DC) Separating	(TT, DC) Good Pooling	(NN, DD) Bad Pooling
$v = v_H$	(N, .)	0, 0	e	e	$1-e$
	(T, C)	1, 1	$(1-e)(1-\alpha)$	$(1-e)^2$	e^2
	(T, D)	$-(1+v)$, $2-v/c$	$(1-e)\alpha$	$(1-e)e$	$e(1-e)$
$v = 0$	(N, .)	0, 0	$1-e$	e	$1-e$
	(T, C)	1, 1	$e\alpha$	$(1-e)^2$	e^2
	(T, D)	-1, 2	$e(1-\alpha)$	$(1-e)e$	$e(1-e)$

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



Parameter values: $a = 0.1$, $e = 0.05$, $c = 0.5$, $v_H = 2$

Note: The vertical axis conflates q and r and so has no meaningful scale, but the vertical segments reflect the fact that the GH equilibrium coincides with GP at $q = 0$ and with GM at $q = q^*$, while the BH equilibrium coincides with BP at $r = 1$ and with BM at $r = r^*$.

Fig. 3. PBE example.

TT and NT that makes Other indifferent between DC and CC. It turns out that the profile $(t^*(x)\text{TT} + (1 - t^*(x))\text{NT}, q^*\text{DC} + (1 - q^*)\text{CC})$ 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 $(u^*(x)\text{NT} + (1 - u^*(x))\text{NN}, r^*\text{DD} + (1 - r^*)\text{DC})$, 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 $q\text{DC} + (1 - q)\text{CC}$ 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, $r\text{DD} + (1 - r)\text{DC}$) 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 $\alpha = a + e - 2ae \leqslant 1/(2 + v_H)$. The complete set of PBE consists of*:

1. *The GP family* (TT, CC) *for* $L(x) \leqslant L(c/v_H) - L(a)$;
2. *The SEP family* (NT, DC) *for* $L(c/v_H) + L(e) - L(a) \leqslant L(x) \leqslant L(c/v_H) + L(e) + L(a)$;
3. *The BP family* (NN, DD) *for* $L(x) \geqslant L(c/v_H) + L(a)$;
4. *The GM family* $(t^*(x)\text{TT} + (1 - t^*(x))\text{NT}, q^*\text{DC} + (1 - q^*)\text{CC})$ *for* $L(c/v_H) - L(a) \leqslant L(x) \leqslant L(c/v_H) + L(e) - L(a)$;
5. *The BM family* $(u^*(x)\text{NT} + (1 - u^*(x))\text{NN}, r^*\text{DD} + (1 - r^*)\text{DC})$ *for* $L(c/v_H) + L(a) \leqslant L(x) \leqslant L(c/v_H) + L(e) + L(a)$;
6. *The GH family* $(\text{TT}, q\text{DC} + (1 - q)\text{CC})$, *for* $q \in [0, q^*]$, *at the point where* $L(x) = L(c/v_H) - L(a)$; *and*
7. *The BH family* $(\text{NT}, r\text{DD} + (1 - r)\text{DC})$ *for* $r \in [r^*, 1]$, *at the point where* $L(x) = L(c/v_H) + L(e) + L(a)$.

The proof, in Appendix A, 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 Fig. 3, for sufficiently small proportions of the vengeful type ($L(x) \geqslant 3.30$ or $x \leqslant 0.036$) we have a Bad Pooling equilibrium:

Table 2
PBE calculations.

	Fitness function		Value in example	
	Non-vengeful type $v=0$	Vengeful type $v=v_H$	Non-vengeful type $v=0$	Vengeful type $v=v_H$
Separating	$e(2\alpha-1)$	$(1-e)(1-(2+v_H)\alpha)$	−0.036	0.418
Good Pooling	$(1-e)(1-2e)$	$(1-e)(1-(2+v_H)e)$	0.855	0.760
Bad Pooling	$-e(1-2e)$	$-e(1+v_H-(2+v_H)e)$	−0.045	−0.140
Good Mix	$(1-e)[1-2e-2q(1-e-\alpha)]$	$(1-e)[1-(2+v_H)e-qa(2+v_H)(1-2e)]$	0	0.608
Bad Mix	$e[-(1-2e)+2(1-r)(1-\alpha-e)]$	$(1-e)[1-(2+v_H)\alpha-r(2+v_H)(1-\alpha-e)]$	−0.646	−2.242

Note: Example parameter values are $a=0.1$, $e=0.05$, $c=0.5$, $v_H=2$. The hybrid equilibria will involve the fitness functions indicated for the corresponding mixed equilibria, with q and r varying within their ranges rather than fixed at particular numerical values.

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 $t^*(x)\approx 0.35\frac{x}{1-x}-0.06$) 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 0.136. 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 $x>0.75$, 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 0.75 and the GP equilibrium disappears. None of these equilibria seems stable in the long run.

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 equilibrium, the vengeful type always has lower fitness because of the extra cost $(1-e)v_He$ 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. Appendix A 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. In particular, each surviving type has expected fitness at least as great as any potential entrant at the equilibrium population state.

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 Fig. 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 $a = A(v)$, where A is a smooth, positive and decreasing function, with $A(0) = 1/2$ and $A(v) \to 0$ as $v \to \infty$. Thus the types cannot be distinguished when vengefulness is negligible, 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 $k > 0$,

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

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,

$$v_H = \arg\max_{v \in [c, v^{\max}]} \left\{E_q W^S\left(v|A(v), e\right)\right\}, \tag{2}$$

   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,

$$E_q W^S\left(0|A(0), e\right) = E_q W^S\left(v_H|A(v_H), e\right). \tag{3}$$

   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.,

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

   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; Appendix A 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[14] EPBE, one that supports mutual gains in the noisy trust game? The BM and BH families are unlikely habitats, again ruled out in Appendix A. The remaining family, Good Hybrid (GH), seems more promising

[14] Efficiency is constrained by the information structure of the game, and is not first-best efficiency.

because it allows both vengeful and unvengeful Selfs to achieve positive fitness, sometimes higher for the vengeful and sometimes higher for unvengeful. The GH strategy profiles are $(\mathrm{TT}, q\mathrm{DC} + (1-q)\mathrm{CC})$.

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 Appendix A. 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 Appendix A. 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. Appendix A 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 emotional state dependent utility components (ESDUCs) 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 specific 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 played in large unstructured groups. 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. Appendix A 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.[15] 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 Fig. 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 it

[15] For example, a system of ordinary differential equations with gradient dynamics for v and replicator dynamics for x, q and the other mixing probabilities seems to converge in time average to the "good" EPBE from nearby initial conditions, according to Matlab simulations available from the authors.

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 Eqs. (7) and (10) in Appendix A. However, there is every reason to take seriously the evolution of perception technology. Obviously Others who evolve a better perception technology (e.g., a lower value of k) would receive a fitness boost. On the other hand, so would a mutant Self with true vengeance parameter $v = 0$ who could somehow mimic the vengeful type, in effect increasing k. Friedman and Singh (2004b) refer to this possibility as the Viceroy problem, a reference to toxic 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.[16] 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. (2007) 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 several important 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 verbal definition and worked it out explicitly for a particular (and not especially simple) game of incomplete information. Appendix A 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. EPBE 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.

Appendix A. Mathematical details

A.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

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

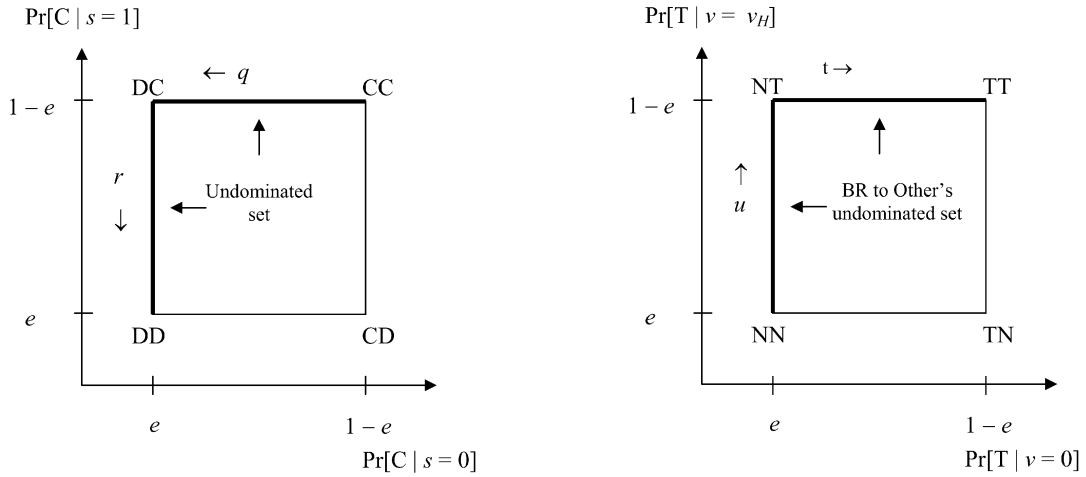
Also, let $r^* = [1+v_H - e(2+v_H)]/[(1-a)(1-2e)(2+v_H)] = q^*(\frac{1+v_H}{2+v_H} - e)/(\frac{1}{2} - e) \in (q^*, 1)$ and define $u^*(x) \in (0,1)$ as the solution to

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

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.

[16] 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.



A. Other's undominated strategies are on the NW frontier B. Self's BR to Other's undominated strategies are also on the NW frontier

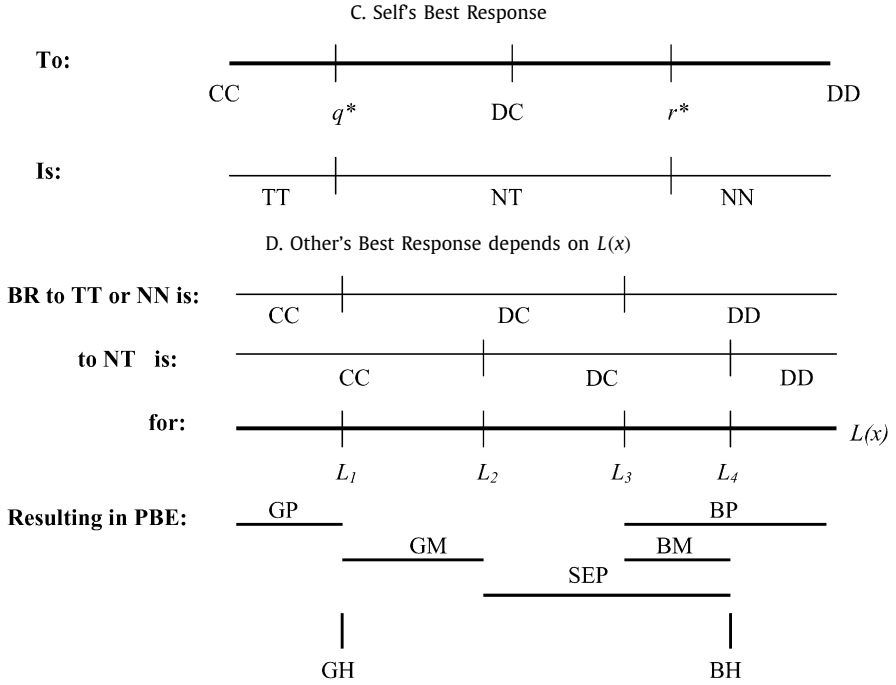


Fig. 4. Best responses and PBE.

Step 2. The only undominated Self strategies in Fig. 4A 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.

Step 3. TN is not a best response to any undominated strategy of Self. Direct computation shows that the best responses are as in Fig. 4C: 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 $\alpha = a + e - 2ae \leqslant 1/(2 + v_H)$. Suppose that Other will play DC. Then 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 $\alpha \leqslant 1/(2 + v)$, 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 $\alpha \leqslant 1/(2 + v) = 1/2$, 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 $0 = E_q W^S(\mathrm{T}|v = 0) \equiv 1(1 - \gamma) - 1\gamma$, 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^*.

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 $L(x) \leqslant L(c/v_H) + L(e) + L(a)$ 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 $L(x) \leqslant L(c/v_H) - L(a)$. That is, CC here is a best response to TT.

These computations are summarized in the first two lines of Fig. 4D. 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.

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.

Step 6. Now consider strict mixes of TT and NT. By Fig. 4D, 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 $L_1 \leqslant L(x) \leqslant L_2$ we can construct a unique PBE, which takes the Good Mix form $(t^*(x)\text{TT} + (1-t^*(x))\text{NT}, q^*\text{DC} + (1-q^*)\text{CC})$. 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 Eq. (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 $L_1 \leqslant L(x) \leqslant L_2$ 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 $L_2 \leqslant L(x) < L_4$, 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 $L_3 \leqslant L(x) \leqslant L_4$ 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.

A.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 $E_q W^S(v) = (1-e)[q(1-\alpha-(1+v)\alpha) + (1-q)((1-e)-(1+v)e)] = (1-e)[1-(2+v)e-qa(2+v)(1-2e)]$. The first order condition (FOC) $0 = dE_q W^S/dv$ for the maximization problem (2) simplifies slightly to $0 = -e - qA'(2+v)(1-2e) - qa(1-2e)$ or, separating variables,

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

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

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

and the SOC is

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

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

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

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

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

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. *Eq.* (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. 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 $4kv\exp(kv^2) = 4kv(1 + kv^2 + \cdots)$. 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

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

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

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

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 $g \geqslant 1/v$, 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. □

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. Write (9) as $S(k) \geqslant 1$. 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 $k \geqslant \varepsilon > 0$ 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. $S(k) \geqslant 1 - \epsilon$. 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 $\bar{k} \approx 0.612$. □

Eqs. (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 $(\mathrm{TT}, q\mathrm{DC} + (1 - q)\mathrm{DD})$ is a PBE. By Proposition 1, we need only check that $q \leqslant q^* \equiv 1/(2 - 2a)$. Use (14) and rearrange to obtain

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

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 $E_r W^S(v) = (1-e)[r(e-(1+v)(1-e)) + (1-r)((1-\alpha)-(1+v)\alpha)] = (1-e)[1-(1-e)(2+v) + (1-r)a(1-2e)(2+v)]$. Hence

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

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 $E_r W^S(v)$ 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 unvengeful 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 $e = \hat{e}$.

A.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) $(2+v)A'' + 2A' \geqslant 0$. Next, note that Eqs. (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)–(iii) hold,[17] 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 $k \leqslant 0.2$. 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 $O(1/v)$.

A.4. A more general definition of EPBE

Eqs. (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.

[17] 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.

Take as given a finite set of player populations $i = 1, \ldots, I$. 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) $p = (p_1, \ldots, p_I)$ over $\Theta = \Theta_1 \times \cdots \times \Theta_I$. 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 $\sigma = (\sigma_1, \ldots, \sigma_I)$. 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 vengeance model presented in the present paper 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 profile σ.

For distribution $p \in P$ let **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 conditional 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 $(\sigma, \mu) \in$ **PBE**(p); and
2. For each population $i = 1, \ldots, I$ and each $\theta_i \in S_i(p)$, the payoff $u_i(\theta_i|p, \sigma) \geqslant u_i(\tilde{\theta}_i|p, \sigma)$ for all $\tilde{\theta}_i \in \Theta_i$.

Condition 2 is the equal, maximal payoff property: equilibrium payoffs of each surviving type achieve equal and maximal payoff in each population.[18] We close with a series of remarks on the nature of EPBE.

- The original evolutionary equilibrium concept (Maynard 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

Abreu, Dilip, Sethi, Rajiv, 2003. Evolutionary stability in a reputational model of bargaining. Games Econ. Behav. 44 (2), 195–216 (August).
Becker, Gary S., 1976. The Economic Approach to Human Behavior. University of Chicago Press, Chicago.
Bergstrom, Theodore C., 2002. Evolution of social behavior: Individual and group selection. J. Econ. Perspect. 16 (2), 67–88.
Bohnet, Iris, Frey, Bruno S., Huck, Steffen, 2001. More order with less law: On contract enforcement, trust, and crowding. Amer. Polit. Sci. Rev. 95 (1), 131–144 (March).
Bolton, Gary E., Ockenfels, Axel, 2000. ERC: A theory of equity, reciprocity and competition. Amer. Econ. Rev. 90 (1), 166–193 (March).
Boyd, Robert, Gintis, Herbert, Bowles, Samuel, Richerson, Peter J., 2003. The evolution of altruistic punishment. Proceed. Nat. Acad. Sci. 100 (6), 3531–3535 (March 18).
Charness, Gary, Rabin, Matthew, 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117 (3), 817–869 (August).
Cox, James C., Friedman, Daniel, Gjerstad, Steven, 2007. A tractable model of reciprocity and fairness. Games Econ. Behav. 59 (1), 17–45 (April).

[18] Note that condition 3 on v in the EPBE definition in Section 5 is subsumed in condition 2 in the general definition above.

Dekel, Eddie, Ely, Jeffrey C., Yilankaya, Okan, 1998. The evolution of preferences. Working Paper. Northwestern University, http://www.kellogg.nwu.edu/research/math/JeffEly/working/observe.pdf.
Dufwenberg, Martin, Kirchsteiger, Georg, 2004. A theory of sequential reciprocity. Games Econ. Behav. 47 (2), 268–298 (May).
Dufwenberg, Martin, Smith, Alec, Van Essen, Matt, 2008. Hold-up: With a vengeance. Working Paper 08-10. Department of Economics, University of Arizona.
Ely, Jeffrey C., Yilankaya, Okan, 2001. Nash equilibrium and the evolution of preferences. J. Econ. Theory 97, 255–272.
Eshel, Ilan, 1983. Evolutionary and continuous stability. J. Theoretical Biology 103, 99–111.
Falk, Armin, Fischbacher, Urs, 2006. A theory of reciprocity. Games Econ. Behav. 54 (2), 293–315.
Fehr, Ernst, Gachter, Simon, 2002. Altruistic punishment in humans. Nature 415, 137–140 (January 10).
Fehr, Ernst, Schmidt, Klaus M., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114 (3), 817–868 (August).
Frank, Robert, 1988. Passions within Reason: The Strategic Role of the Emotions. W.W. Norton, New York.
Friedman, Daniel, Ostrov, Daniel, 2008. Conspicuous consumption dynamics. Games Econ. Behav. 64 (1), 121–145.
Friedman, Daniel, Singh, Nirvikar, 2004a. Negative reciprocity: The coevolution of memes and genes. Evolution Human Behav. 25 (3), 155–173.
Friedman, Daniel, Singh, Nirvikar, 2004b. Vengefulness evolves in small groups. In: Huck, Steffen (Ed.), Advances in Understanding Strategic Behavior. Palgrave, pp. 28–54.
Friedman, Daniel, Singh, Nirvikar, 2007. Equilibrium vengeance. MPRA Paper No. 4321. http://mpra.ub.uni-muenchen.de/4321/1/MPRA_paper_4321.pdf.
Fudenberg, Drew, Maskin, Eric, 1986. The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54 (3), 533–554.
Fudenberg, Drew, Tirole, Jean, 1991. Game Theory. MIT Press, Cambridge, MA.
Geanakopolis, John, Pearce, David, Stacchetti, Ennio, 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.
Gintis, Herbert, 2002. Altruism and emotions. Behavioral Brain Sci. 25, 258–259.
Güth, Werner, 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. Int. J. Game Theory 24, 323–344.
Güth, Werner, Kliemt, Hartmut, 1994. Competition or cooperation: On the evolutionary economics of trust, exploitation and moral attitudes. Metroecon. 45 (2), 155–187.
Güth, Werner, Yaari, Menachem, 1992. An evolutionary approach to explaining reciprocal behavior. In: Witt, U. (Ed.), Explaining Process and Change-Approaches to Evolutionary Economics. The University of Michigan Press, Ann Arbor.
Güth, Werner, Kliemt, Hartmut, Peleg, Bezalel, 2000. Co-evolution of preferences and information in simple games of trust. Ger. Econ. Rev. 1 (1), 83–110.
Hirshleifer, Jack, 1987. On the emotions as guarantors of threats and promises. In: Dupre, J. (Ed.), The Latest on the Best: Essays in Evolution and Optimality. MIT Press, pp. 307–326.
Heifetz, Aviad, Shannon, Chris, Spiegel, Yossi, 2007. What to maximize if you must. J. Econ. Theory 133 (1), 31–57.
Henrich, Joseph, 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. J. Econ. Behav. Organ. 53 (1), 3–35.
Herold, Florian, 2004. Carrot or stick: The evolution of reciprocal preferences in a haystack model. University of Munich Department of Economics Discussion Paper (November).
Huck, Steffen, Oechssler, Jorg, 1999. The indirect evolutionary approach to explaining fair allocations. Games Econ. Behav. 28, 13–24.
Jacobsen, Hans Jørgen, Jensen, Mogens, Sloth, Birgitte, 2001. Evolutionary learning in signalling. Games Econ. Behav. 34 (1), 34–63.
Journal of Economic Behavior and Organization, 2004. Special Issue on Evolution and Altruism. J. Econ. Behav. Organ. 53 (1).
Kaufman, Stuart, 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, NY.
Kockesen, Levent, Ok, Efe A., Sethi, Rajiv, 2000. The strategic advantage of negatively interdependent preferences. J. Econ. Theory 92 (2), 274–299 (June).
Levine, David K., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.
Maynard, Smith John, Price, George R., 1973. The logic of animal conflict. Nature 246, 15–18.
Nöldeke, Georg, Samuelson, Larry, 1997. A dynamic model of equilibrium selection in signaling markets. J. Econ. Theory 73, 118–156.
Ok, Efe A., Vega-Redondo, Fernando, 2001. On the evolution of individualistic preferences: An incomplete information scenario. J. Econ. Theory 97, 231–254.
Possajennikov, Alex, 2002. Cooperative prisoners and aggressive chickens: Evolution of strategies and preferences in 2×2 games. Discussion Paper 02-04. National Research Center 504 "Rationality Concepts, Decision Behavior, and Economic Modeling", University of Mannheim (January).
Rabin, Matthew, 1993. Incorporating fairness into game theory economics. Amer. Econ. Rev. 83, 1281–1302.
Robson, Arthur J., 2002. Evolution and human nature. J. Econ. Perspect. 16 (2), 89–106.
Rilling, James K., Gutman, David A., Zeh, Thorsten R., Pagnoni, Guiseppe, Berns, Gregory S., Kitts, Clinton D., 2002. A neural basis for cooperation. Neuron 35, 395–405.
Rubin, Paul H., Paul, C.W., 1979. An evolutionary model of taste for risk. Econ. Inquiry 17, 585–596.
Samuelson, Larry, 2001. Introduction to the evolution of preferences. J. Econ. Theory 97, 225–230.
Samuelson, Larry, Swinkels, Jeroen, 2006. Information and the evolution of the utility function. Theoretical Econ. 1, 119–142.
Sethi, Rajiv, Somanathan, E., 2003. Understanding reciprocity. J. Econ. Behav. Organ. 50, 1–27.
Sobel, Joel, 2005. Interdependent preferences reciprocity. J. Econ. Lit. 43 (2), 393–436.
van Winden, Frans, 2001. Emotional hazard exemplified by taxation-induced anger. Kyklos 54, 491–506.
Wright, Sewall, 1949. Adaption and selection. In: Jepsen, L., Simpson, G.G., Mayr, E. (Eds.), Genetics, Paleontology, and Evolution. Princeton University Press, Princeton, NJ.