A Theory of Moral Contagion

By Robert Innes
University of Arizona

VERY PELIMINARY

1. Introduction

Economists have long recognized the costs of selfish behavior in joint ventures. Such collaborations often take the form of Prisoner's Dilemma (PD) games in which the strategy "not to cooperate" (to defect) is in each player's self-interest, voiding collective gains from cooperation. Not surprisingly – given the importance of collaboration to economic growth and human welfare – there is a vast literature, both experimental and theoretical, studying mechanisms that can overcome the implied failure.

Chief among these mechanisms are preferences that motivate reciprocity, when generosity is rewarded and defection or stinginess is punished (Rabin, 1993). Substantial empirical evidence documents reciprocal behavior in various experimental settings (e.g., Cox, 2004; Charness and Rabin, 2002; Fehr, Gachter and Kirchsteiger, 1997; Fehr and Gachter, 2000). Theoretical work identifies the potential evolutionary basis for "punishing" preferences (vengeance) to arise and survive (Henrich and Boyd, 2001; Friedman and Singh, 2008).

In this paper, I study prospects for another behavior – also observed in experiments and also directly linked to cooperation – to survive evolutionary dynamics. Recent work documents that moral preferences are contagious; that is, when individuals perceive that most others in a relevant peer group are honest or "fair," they are more likely themselves to be honest or fair (see Bichierri and Xiao, 2007; Innes and Mitra, 2008). Moral preferences thus *change*, in any given situation, with the perceived propensity for morality by peers in that situation.

The implied change in preferences – or conditioning of preferences on social conduct – distinguishes *contagion* from *conformity*, a phenomenon that has been studied quite extensively in economics, psychology and biology. As a result, existing theories of conformity cannot explain the experimental results. Key among these theories are the role of conformity to achieve social status (Akerlof, 1980; Bernheim, 1994); the role of social sanctions to support compliance with social norms (Sugden, 1998; Fehr and Fischbacker, 2004); and the inference of valuable information from others' behavior, motivating conformity with that behavior (Banerjee, 1992, Bikchandaria, Hirschleifer, and Welch, 1992). In the experiments, however, the subjects' identities and actions are completely private, vitiating any motive for obtaining social esteem and voiding the threat of sanction. Moreover, other players' behavior – the putative stimulus for contagion in the experiments – is completely irrelevant to potential payoffs; hence, there is no potential for information cascades to motivate subjects' choices.

In principle, individuals might have some social / other-regarding preferences that produce contagion as a symptom. Indeed, a number of scholars have offered ingenious theories of preferences to explain experimental paradoxes in fairness/dictator, ultimatum, market, trust, and other games. However, none of these theories convincingly explains the experimental results on contagion (Innes and Mitra, 2008).[1]

I take a more direct approach to explain contagion in this paper, viewing contagious preferences as a trait and studying whether this trait can survive evolutionary forces. The assumed trait takes the following form: In a PD game (with outside option), an individual can suffer an additional moral penalty / disutility when behaving selfishly (vs. morally / honestly / cooperatively); if suffering this penalty, the individual is called "honest" (H) and his/her dominant strategy is to play "cooperate" in the PD game; if the individual does not suffer the penalty, then he/she is "selfish" (D) and his/her dominant strategy is to play "defect."[2] Modeling moral/cooperative impulses in this way is not new (see, for example, Frank, 1987; Henrich and Boyd, 2001). However, unlike prior literature, I am interested in a morality trait that is contagious, assigning H-type preferences to an individual when the relevant proportion of H-types (call it h) is high and assigning D-type preferences when h is low.

I argue that such contagious preferences can be advantageous in the sense that they enhance an individual's average payoff (fitness) and thus survive evolution. There are two elements to the argument. First is that there can be network externalities that make conformity in moral preferences advantageous; that is, when h is high, the expected payoff to being an H-type is higher than to being a D-type, and vice versa. The idea that network externalities can motivate uniformity is certainly not new (see, for example, Katz and Shapiro, 1986; Banerjee and Besley, 1990). However, that network externalities arise in the present context – a PD game with moral and selfish players – is not at all obvious apriori. Indeed, Frank (1987) – to my knowledge the only other author to examine this particular issue – concludes that the relationship between expected payoffs is likely to be exactly the opposite; in his analysis, *non*-conformity is advantageous.

[1] Generally speaking, prevailing theories allow for preferences (subject utility) to depend upon payoffs to others who *are affected by* one's actions. For example, in a dictator game, a dictator's utility might depend upon the payoff to the receiver, as with inequity aversion (Fehr and Schmidt, 1999) or ERC preferences that depend upon relative payoffs (Bolton and Ockenfels, 2000). For these cases, other players' behavior (outside of the sender-receiver pair) is irrelevant to the decision-maker's utility; hence, such preferences cannot explain contagion. However, if these other-regarding preferences relate to the entire subject pool – not just a sender's receiver (and vice versa) – then inequity aversion (or ERC preferences) can depend upon behavior in the overall subject pool; Innes and Mitra (2008) show, in this case, that with weakly convex inequity aversion – plausibly implying that (per unit of inequality) larger inequalities are no better than smaller ones – behavior would be the opposite of contagious for the deception game that they study: a higher fraction of subjects exhibiting truthful behavior should spur less truthfulness, not more. Other explanations for experimental paradoxes include preferences that exhibit reciprocity or depend upon social welfare (see, for example, Charness and Rabin, 2002). In the contagion experiments, however, an individual's impact on social welfare is invariant to the conduct of others. In general, there is also no role for reciprocity in the contagion experiments, as there is no receiver choice in the dictator game and no receiver knowledge of sender decisions in the deception game. However, in principle, there is scope for dictator / sender beliefs about receiver expectations of sender behavior to be affected by information about other players' conduct; if so, guilt aversion (Charness and Dufwenberg, 2006) could explain the observed contagion. Innes and Mitra (2008) attempt to control for these sender beliefs – and present evidence of success in doing so – vitiating a guilt aversion motive for their results.

[2] Kaplow and Shavell (2007) characterize an optimal moral system in which guilt and virtue serve to regulate externality-causing behavior. Here, "guilt" similarly motivates virtuous behavior, but is costless (because it is never actually borne) and, hence, is trivially optimal. I am interested in the evolutionary stability of contagious "guilt" (what I call "honesty" in this paper), rather than optimality. Kaplow and Shavell (2007) instead model plausible costs of guilt and describe optimal levels of this attribute.

By revising Frank's (1987) analysis in two (I believe) plausible directions, this conclusion is reversed. The two revisions are: (1) the assignment of partners in the PD game is random / non-assortative, rather than a costless process of assortative partner selection; and (2) the payoff to an outside option (of not playing the PD game) is less than to the joint venture for two "selfish" types. The first assumption is a standard one in evolutionary games (see, for example, Bergstrom, 2002) and is motivated by a plausibly random process for the arrival of joint venture opportunities. The second presumes that there are relatively large gains to joint ventures, even absent "honesty" / cooperation by both parties. Under these conditions, *conformity can be advantageous*. Because H-types are exploited when partnering with D-types, they quit the game – and go for the outside option – when they believe their partner is selfish. This exit hurts the D-types relatively more often when there are few of them (a high h) because they are then more likely to be rejected by H-type partners; conversely, it hurts the H-types relatively more when h is low because then the D-types are less likely to be rejected (as they are more likely to face D-type partners) and H-types are more likely to quit (as they are also more likely to face D-type partners). Hence, when there are no (or quite costly) opportunities to resample for different partners, network effects favor conformity. Moreover, less costly resampling opportunities do not upset this conclusion. With more H-types (higher h), it is the D-types that must resample more often – and bear associated resampling costs – in order to avoid the costly outside option that is forced upon them when they face H-type partners; conversely, when h is low, it is the H-types that must resample more often in order to find an H-type partner.

This logic – developed in Section 2 below – implies that conformity in moral preferences can be advantageous. But it does not necessarily imply that contagious preferences are advantageous. Hence, the second element to the argument introduces some inter-group heterogeneity and profitable inter-group opportunities for joint ventures. Incentives for conformity can give rise to different equilibria in two distinct groups, absent trade. One group can evolve to be selfish and the other to be honest. When trade is introduced, there then is an advantage to contagious preferences. For example, suppose that an H-type person in the honest group faces a positive probability of having a partner from his/her own "honest" group and a positive probability of facing a partner from the other (selfish) group. Both opportunities are profitable with contagious preferences, but only one opportunity (that with the own group partner) is profitable with non-contagious but conformist preferences. This logic – developed in Section 3 – implies that contagion (vs. conformist) preferences will arise and survive evolutionary forces in at least some subset of the overall population.

Section 4 discusses potential pitfalls and criticisms of the prior analysis. Most important in this regard is the simplifying assumption that the domain of preferences is limited to strictly "moral" (implying a dominant strategy of "cooperation" in the PD game) vs. strictly "selfish." An alternative is for moral penalties to one's own deception only to kick in when the opposing partner plays morally. In addition, given experimental evidence of reciprocity, it is natural to ask whether "punishing" preferences can arise in tandem with moral contagion. Finally, the paper would not be complete without some response to the standard criticism of evolutionary motives for virtue, namely, that it is assumed to be observable and impossible to mimic (see Sobel, 2005).

The idea that there might be contagion in preferences has been put forward in prior work, mostly in different contexts than I study here. For example, Linkbeck, Nyberg and Weibull (1999) and Besley and Coate (1992) assume that stigma from welfare is a decreasing function of the population proportion on welfare; hence, stigma is contagious, and the authors study implications of this assumption for equilibrium employment and the political equilibrium in government welfare transfers. More closely related is a fascinating paper by Henrich and Boyd (2001), which studies the role of conformity – a tendency to "copy the majority" – in supporting an evolutionarily stable equilibrium in which preferences for punishing defectors can survive. In contrast to the present paper, Henrich and Boyd (2001) *assume* that the evolutionary mechanism takes a conformist form – what they call conformist transmission, whereby the population fraction of moral (and/or punishing) preferences rises (falls) if the majority has these preferences. I instead follow convention in assuming that evolution responds to relative fitness and study whether contagion in preferences will endogenously arise and survive.

2. A Model of Conformity in Honesty

Frank (1987) models morality as a character attribute with which an individual is either endowed or not. Over time, the propensity for this attribute to exhibit itself in the population evolves based on a Darwinian principle: If moral players earn higher expected payoffs than selfish players, then the proportion of the population that is moral rises, and vice versa. The benefit of morality is that it pre-commits an individual to a cooperative strategy in a joint endeavor that takes the form of a Prisoner's Dilemma (more in a moment); hence, two moral players overcome the Prisoner's dilemma. Conversely, a selfish player plays in his / her self-interest.

Modeling morality in this way, one can ask whether in-bred "contagion in morality" is an advantageous attribute that would survive Darwinian evolution. Specifically, suppose that a population has some members that are moral, some that are selfish, and others who exhibit contagious honesty; that is, they are moral when the fraction of the overall population that is moral / "honest" (call it h) is sufficiently high, and selfish when this fraction is sufficiently low. Is the "contagious morality" attribute advantageous, so that it would evolve in a Darwinian world?

In Frank's (1987) analysis, the answer is "no." Let x denote an individual's payoff, and $E(x \mid q,h)$ denote an individual's expected payoff when he is moral / honest ($q=H$) or selfish / dishonest ($q=D$). Frank shows (under conditions specified in his paper) that there is a unique interior $h^*\varepsilon(0,1)$ such that $E(x \mid D,h) > E(x \mid H,h)$ for $h>h^*$, $E(x \mid D,h) < E(x \mid H,h)$ for $h<h^*$, and $E(x \mid D,h) = E(x \mid H,h)$ for $h=h^*$,. These inequalities imply a unique evolutionary equilibrium h^* when the morality attribute is binary (either moral or selfish). They also imply that an "anti-contagious morality" attribute would be advantageous; that is, it is advantageous under these circumstances, to be moral when most people are selfish ($h<h^*$), and to be selfish when most people are moral ($h>h^*$). Hence, as it stands, the predictions of Frank's analysis are completely at odds with the "contagious morality" that is observed in recent experiments.

However, this conclusion is not a general one, nor even, I believe, the most plausible outcome of a model like Frank's (1987). Indeed contagious morality is likely to be an advantageous attribute in simple (and plausible) model variants.

In the language of evolutionary dynamics, let $m\varepsilon(0,1)$ be the honesty mutation, where m=1 denotes honesty (H) and m=0 denotes selfishness (D). Then the fitness of an individual in a group with the proportion h of honest H individuals is the payoff,

$V(m,h) = E(x \mid D,h)\ (1-m) + E(x \mid H,h)\ m.$

In what follows, I describe circumstances in which the menu of fitness maximizing honesty mutations,

$m^*(h) = \text{argmax}_{\ m\varepsilon(0,1)}\ V(m,h),$

has the "contagion" property,

$$m^*(h) = \begin{array}{l} 1 \text{ when } h \geq h^*, \\ 0 \text{ when } h < h^*. \end{array}$$

Examining a menu of fitness maximizing choices is equivalent to examining mutations in function space, rather than is the space of a single parameter (m=0 or m=1). For the moment, I sidestep the question of whether a function mutation is advantageous (vs. a parameter mutation), but will return to this issue in the next section. However, note that the maximization of fitness is the standard criterion for evolutionary survival of an attribute.

I should be clear that, throughout the discussion, I assume that contagion is an inbred rule that assigns preferences (H or D) based on a stimulus (the observed h). Once an individual gets to the partnership game, his/her attribute (H or D) is fixed.[3]

Consider Frank's PD game, as shown in Figure 1, where $x_4 > x_3 > x_2 > x_1$. Each player can refuse to play the game, in which case he/she earns the payoff x_0. The outside option x_0 is superior if one expects to be on the receiving end of selfishness, $x_0 > x_1$, but is otherwise inferior to the joint venture, $x_0 < x_2$. [4] By assumption, an honest H individual plays the honest (H) strategy in the PD game; equivalently, I implicitly assume that, for an H individual, selfishness / dishonesty bears a monetary-equivalent cost $c \geq \max(x_4 - x_3, x_2 - x_1)$, so that H (cooperate) is a dominant strategy for H types;[5] for D types, who bear no cost of dishonesty beyond that embedded in the x payoffs, D (defect) is a dominant strategy.

Each individual plays the PD game with another individual randomly chosen from the population at large, reflecting the notion that one rarely has the luxury of choosing those with whom advantageous joint venture opportunities are available.[6] At this juncture, for simplicity, I ignore costly opportunities for resampling – rejecting the partner with whom one is currently matched and redrawing from the population for another partner; I will return to such opportunities in a moment. Hence, an individual is randomly matched with another player, and then chooses whether or not play the PD game; if not, the individual earns the outside option payoff x_0. Before partnering takes place, an individual's type (H or D) is determined, potentially based on an inbred rule

[3] The idea of stimulus-contingent preferences is common in modern economics, with reciprocity and vengeance perhaps the most studied. My interest is in a different form of stimulus-contingent preferences than studied elsewhere.

[4] The assumption that $x_0 > x_2$ departs from Frank (1987) and reflects my premise that there are large gains to joint ventures.

[5] In principle, the cost of dishonesty c could depend upon whether the other player is honest or dishonest. We return to this issue later.

[6] This assumption – called non-assortative matching – is common in the literature (see, for example, Bergstrom, 2002).

that assigns the type as a function of the perceived population propensity for honesty, t(h) ε{H,D}.

A. Perfect Signals of Type

We first suppose that an individual obtains a perfect signal of the partner's type, H or D. Then an honest (H) type, when matched with another H type, obtains the payoff x_3; however, faced with a D partner, the payoff to playing the game (x_1) is less than the outside option payoff (x_0) and, hence, the H type obtains x_0. Apriori, before knowing his/her partner, an H-type obtains the expected payoff,

$E(x \mid H,h) = h\, x_3 + (1-h)\, x_0.$

Similarly, if a D-type player is matched with an H type, the H type refuses to play and the D player obtains x_0; if matched with another D type, the two enjoy the collective (D,D) payoff x_2 ($>x_0$). Hence, before knowing his/her partner, a D type obtains the expected payoff,

$E(x \mid D,h) = h\, x_0 + (1-h)\, x_2.$

Figure 2 graphs these expected payoffs, illustrating that

$E(x \mid H,h) > (<) E(x \mid D,h)$ for $h>(<)h^*$,

with $h^* = (x_2- x_0)/\{(x_3- x_0)+(x_2- x_0)\}$ ε (0,1). Therefore, it is most advantageous for an individual to be honest when a high proportion of the population is honest ($h>h^*$) and to be dishonest when a high proportion of the population is dishonest ($h<h^*$).

Intuitively, honest individuals enjoy higher payoffs when there are more honest people in the population because they then partner with honest compatriots more often – enjoying the high joint payoff x_3; and because they have dishonest partners less often, they also go it alone – with the corresponding low payoff x_0 – less often. Conversely, dishonest individuals obtain lower payoffs when there are more honest people because honest types refuse to engage in joint ventures with them. In the extreme, when almost everyone is honest (h=1), honest types obtain the high cooperative payoff x_3 with certainty, whereas a dishonest type will be forced to go it alone with certainty, obtaining the low payoff x_0. Conversely, when almost everyone is dishonest (h=0), honest types go it alone with certainty (obtaining x_0), whereas the dishonest enjoy the higher joint venture payoff x_2 with certainty. It thus pays to be honest when others are mostly honest and, conversely, to be dishonest when others are mostly dishonest.

B. Imperfect Signals of Type

An individual's type (qε{H,D}) may not be observable, but produce an observable signal S (a tendency to blush or stand up for others, for example). Following Frank (1987), I assume that H-types draw a signal S from the density $f_H(S)$ on the support $[L_H,U_H]$, while D types draw from the density $f_D(S)$ on the support $[L_D,U_D]$, where $L_H>L_D$, $U_H \geq U_D$, and $f_H(S)/f_D(S)$ increases with S on Sε$[L_H,U_D)$. Hence, S is perfectly informative when Sε$[L_D,L_H)$ (signaling q=D) or Sε$[U_D,U_H)$ (signaling q=H), and imperfectly informative otherwise.

For dishonest D types, it is always advantageous to play the PD game (vs. the outside option) whenever the matched partner is willing to do so. However, for H types, the willingness to play depends upon the perceived probability that the partner is also honest, based on the informative S signal. Specifically, let

(1) $P(H \mid S,h) = hf_H(S)/ \{ hf_H(S)+ (1-h)f_D(S)\}$

be the probability that a partner with signal S is an H-type. Then, for the H person, the net benefit to joining the cooperative (PD) venture is

(2) $G(S,h) = P(H \mid S,h)\, x_3 + (1 - P(H \mid S,h))\, x_1 - x_0$.

It is easily seen that G(S,h) rises with S; hence, there is a critical S* such that the H-type plays the PD game when the partner's S is above S* (where G>0) and goes it alone when S is below S* (where G<0). When $h \varepsilon (0,1)$, $S^*(h) \varepsilon (L_H,U_D)$ is interior and solves:

(3) $G(S^*,h) = 0 \leftrightarrow P(H \mid S^*,h) = [x_0 - x_1]/[x_3 - x_1]\ \varepsilon\ (0,1)$.

With this groundwork, let me be more specific about the game being played and the strategy equilibrium, for a given set of H and D players in the population. First, players are randomly matched into pairs. Second, in each pair, the two players' S signal values are revealed to both players. Third, the two players simultaneously and irreversibly decide whether to "accept" the PD game or to "reject." If and only if both players choose "accept," the PD game is played and payoffs are realized accordingly. Otherwise, both players must go it alone and each obtain the payoff x_0.

In this game, a player's strategy is represented by a mapping from (S_o,S_p,q) – the player's own signal (S_o), the matched partner's signal (S_p), and the player's type, $q\varepsilon\{H,D\}$ – to an accept / reject decision, $d\varepsilon\{A,R\}$. The strategy mappings $\{d(S_o,S_p,H), d(S_o,S_p,D)\}$ represent a perfect Bayesian equilibrium if the strategies are optimal for any individual H or D player given that all other players implement these strategies.

The following are easily seen to be equilibrium strategies in this game:

(4a) $d(S_o,S_p,H) =$ A iff $S_o \geq S^*$ and $S_p \geq S^*$,
R otherwise

(4b) $d(S_o,S_p,D) =$ A for all (S_o,S_p).

Accept is clearly a dominant strategy for D-type players. To understand $d(S_o,S_p,H)$, consider the choice problem of an H-type player. For $S_o \geq S^*$ and $S_p \geq S^*$, the matched partner accepts (under the posited strategies) whether he or she is an H or D player; hence, by the definition of S* in equations (1)-(3) above, the H player obtains a higher expected payoff by playing the PD game (with accept) than by not doing so (with reject). However, if either $S_o < S^*$ or $S_p < S^*$, the matched player only accepts if he/she is a D-type player; hence, if the H player accepts, he/she either obtains x_0 (if the other player is H and thus rejects) or obtains the exploitation payoff x_1 ($<x_0$); in expectation, an accept decision thus yields a lower payoff than the x_0 obtained under a reject decision.

Before knowing with whom one is matched (and hence the partner's S_p), the expected payoff to honest (H) and dishonest (D) individuals can now be expressed, taking into account the equilibrium strategies of equation (4):

(5a) $E(x \mid H,h) = F_H(S^*)\, x_0 + (1-F_H(S^*))\{h[\, F_H(S^*)\, x_0 + (1-F_H(S^*))\, x_3]$
$+ (1-h)[F_D(S^*)\, x_0 + (1-F_D(S^*))\, x_1]\}$

(5a) $E(x \mid D,h) = (1-h)\, x_2 + h\{\, F_H(S^*)\, x_0 + (1-F_H(S^*))\, [F_D(S^*)\, x_0 + (1-F_D(S^*))\, x_4]\}$

H-types go it alone (obtaining x_0) whenever their own or their partner's S is less than S*(h). Their own S is less than S* with probability $F_H(S^*)$ (where F_q is the cumulative density / distribution function for f_q). Their partner's S is less than S* with probability $F_H(S^*)$ if the partner is also an H-type, and with probability $F_D(S^*)$ ($>F_H(S^*)$) if the partner is a D-type. When they join the game with an H-type partner, the honest players earn the cooperative payoff x_3, but with D-type partners, they suffer the exploitation payoff, x_1.

Similarly, D types always play the PD game with D-type partners, enjoying the (D,D) payoff x_2 when they do so. With an H-type partner, the D player must go it alone if either player's S is less than S* (because the partner then refuses to play). However, if both players' S values are above S*, the D player exploits the H partner, obtaining x_4.

Now consider the case of h=0, when there are virtually no honest members of the population. Then D-types always play with other D-types, obtaining x_2 with certainty. For an H-type, he will only play the PD game if the signal guarantees an H partner ($S^*=U_D$, implying $F_D(S^*)=1$). Hence, we have (as in the perfect signal case):

(6) $$E(x \mid D,0) = x_2 > x_0 = E(x \mid H,0).$$

Conversely, consider the case of h=1, when there are virtually no dishonest members of the population. Now an honest individual will play the PD game whenever the partner's S signal does not reveal a D partner with certainty; that is, $S^*=L_H$ because for any $S \geq L_H$, the inferred probability that the partner is an H-type is one. Hence, we have:

(7) $$E(x \mid H,1) = x_3\ ,\ \ E(x \mid D,0) = F_D(S^*)\, x_0 + (1-F_D(S^*))\, x_4.$$

From (7) we have the following condition for the H-type expected payoff to be higher than the D-type payoff at h=1:

(8) $$(x_3 - x_0)F_D(L_H) > (x_4 - x_3)(1-F_D(L_H)).$$

In words, equation (8) requires that (i) the signal is sufficiently informative (so that $F_D(L_H)$ is sufficiently high), and (ii) the maximum gains to cooperation ($x_3 - x_0$) are sufficiently large relative to the benefits of exploitation ($x_4 - x_3$). For example, if the cooperation gains, ($x_3 - x_0$), are twice the exploitation gains, ($x_4 - x_3$), then (8) holds provided the probability of discovery for a dishonest (D) person ($F_D(L_H)$) is more than one third.

Condition (8) implies that there is a critical propensity for honesty (h) above which it pays to be honest and below which it pays to be dishonest. For an example that satisfies equation (8), patterned after the example that frames Frank's (1987) analysis, Figure 3 graphs the expected payoff functions of equation (5).[7] Here, the functions cross once at h* = .46.

C. Resampling: The Case of Perfect Signals

So far we have assumed that a PD game is a one-shot opportunity with a single randomly selected partner. Suppose instead that, at a cost, a player can draw another partner randomly from the population. For simplicity, we will return to the case of perfect signals. Hence, an H-type person can redraw a partner if his/her initial partner is a D-type; conversely, a D-type can redraw if his initial partner is an H-type (who refuses to play the PD).

Let c>0 denote the cost of each resample. At the outset, I will assume that c is sufficiently small that it is advantageous for both H and D types to resample when matched with an opposite-type partner. (I will characterize when this is true in a moment.) The game proceeds sequentially, with all remaining players randomly matched with each other in each successive round. That is, in round 1 players are randomly matched for the overall population; in round 2, remaining players – those from round 1

[7] The example underpinning Figure 3 is as follows: f_H and f_D are normalized standard normals with means $\mu_H=4$ and $\mu_D=2$ (respectively), truncated at plus and minus two standard deviations:

$$f_q(S) = \{I(2\pi)^{.5}\}^{-1}\ \{\exp\{-(S-\mu_q)^2/2\} - \exp(-2)\},$$

where $I=\Phi(2)-\Phi(-2)-4\{\exp(-2)/\{(2\pi)^{.5}\}\}$, Φ= standard normal distribution function. Further, for simplicity, I assume: $(x_4 - x_3) = (x_3 - x_2) = (x_2 - x_0) = (x_0 - x_1) = x_1 = 1$.

who have chosen to resample – are randomly matched from the population of remaining players; and so on.

Let h_j denote the fraction of the remaining players in round j that are H types, where $h_1 = h$ trivially. By my assumption that all ill-matched players resample, we have:

(9) $h_j = [h_{j-1} (1- h_{j-1})] / \{2 [h_{j-1} (1- h_{j-1})]\} = ½$ for j=2,3,…

namely, the fraction of players that are H-types and resample in the prior round, divided by the total fraction of players that resample in the prior round. Equation (9) is intuitive; for every ill-matched player of type H, there is a corresponding D partner, and vice versa; hence, each type represents fifty percent of resampling populations, regardless of the initial h.

Because H types resample whenever matched with a D-type, which occurs with probability (1/2) in all but the first round, we have the expected payoff:

(10a) $$E(x \mid H,h) = h\,x_3 + (1\text{-}h)\{(1/2)x_3 – c + (1/2)[(1/2)x_3 –c + (1/2)(\ldots)]\}$$

$$= h\,x_3 + (1\text{-}h)\sum_{j=1}^{8} \{(1/2)^j x_3 – (1/2)^{j-1}c\} = x_3 - (1\text{-}h)2c.$$

Similarly,

(10a) $$E(x \mid D,h) = (1\text{-}h)\,x_2 + h\sum_{j=1}^{8} \{(1/2)^j x_2 – (1/2)^{j-1}c\} = x_2 - h2c.$$

At this juncture, I return to the question of when resampling is optimal. Consider, for illustrative purposes a D-type individual's choice of whether to resample once and only once, when faced with an H partner who also resamples. Without resampling, the D-type obtains the outside option payoff x_0; with resampling, he/she will face a D-type and H-type partner with equal probability. With the D-type partner, h/she obtains the PD game payoff x_2; and with the H-type partner, he/she again must go it alone. Hence, resampling is desirable if

(11) $x_0 \le (1/2)(x_2+x_0) – c \leftrightarrow c \le (1/2)(x_2\text{-}x_0).$

Clearly, (11) is a sufficient condition for D-types to resample. Moreover, by induction, if it is optimal for a player to resample in one round (when ill-matched), it is also optimal in the next round. Therefore, a necessary condition for resampling to be optimal, given resampling by H-types, is:

(11') $$x_0 \le \{(1/2)x_2 – c + (1/2)[(1/2)x_2 –c + (1/2)(\ldots)]\}$$

$$= \sum_{j=1}^{8} \{(1/2)^j x_2 – (1/2)^{j-1}c\} = x_2 - 2c \quad \leftrightarrow \quad c \le (1/2)(x_2\text{-}x_0).$$

Similarly, a necessary and sufficient condition for H-type resampling is:

(12) $c \le (1/2)(x_3\text{-}x_0).$

Because equation (11) implies equation (12), (11) is necessary and sufficient for both player types to resample.

Assuming (11) holds, we can compare the expected payoffs in (10a) and (10b). At h=1, we have

(13a) $E(x \mid H,1) - E(x \mid D,1) = x_3 - (x_2\text{-}2c) > 0.$

Similarly, at h=0,

(13b) $E(x \mid D,0) - E(x \mid H,0) = 2c – (x_3 - x_2) > 0$ if $c > (1/2)(x_3\text{-}x_2).$

Provided c satisfied the last inequality, it pays to be an H-type when the population proportion of H-types is sufficiently high (h>h*) and, conversely, to be a D-type when the proportion of D-types is sufficiently high (h<h*), with

(14) $h^* = \{2c - (x_3 - x_2)\} / 4c\ \varepsilon\ (0,1/2)$.

Intuitively, it pays to be an H-type when h is high because then H types are more frequently matched with each other and therefore need to bear resampling costs infrequently. Likewise, it pays to be a D-type when they are prevalent because then D-types need to bear resampling costs infrequently.

Following similar logic, we can characterize resampling equilibria for the full range of possible c values:

Proposition 1. Assume that $(x_3 - x_2) < (x_2 - x_0)$. In respective equilibria, allowing for resampling:

(1) if $c \leq (1/2)(x_3-x_2) < (1/2)(x_2-x_0)$, then both H and D types resample and $E(x\,|\,H,h) > E(x\,|\,D,h)$ for $h\ \varepsilon\ (0,1]$;

(2) if $(1/2)(x_3-x_2) < c \leq (1/2)(x_2-x_0)$, then both H and D types resample and $E(x\,|\,H,h) > (<)\ E(x\,|\,D,h)$ when $h > (<) h^*\ \varepsilon\ (0,1/2)$;

(3) if $(1/2)(x_2-x_0) < c \leq (x_3-x_2)$, then only H types resample; they resample only once; and, for all $h\ \varepsilon\ (0,1]$, $E(x\,|\,H,h) = x_3 - (1-h)c > hx_2 + (1-h)x_0 = E(x\,|\,D,h)$.

(4) if $\max\ \{(1/2)(x_2-x_0), (x_3-x_2)\} < c \leq (x_3-x_0)$, then only H types resample; they resample only once; and

$$E(x\,|\,H,h) = x_3 - (1-h)c > (<) hx_2 + (1-h)x_0 = E(x\,|\,D,h)$$

when $h > (<) h^* = \{c - (x_3-x_2)\}/ \{c + (x_2-x_0)\}\ \varepsilon\ (0,1/2)$,

(5) if $c > (x_3-x_0)$, then no one resamples and $E(x\,|\,H,h) > (<) E(x\,|\,D,h)$ when $h > (<) h^* = (x_2-x_0) / \{(x_3-x_0)\}+(x_2-x_0)\}\ \varepsilon\ (0,1/2)$,

Proposition 1 implies that, for a variety of cases (cases (2), (4) and (5)), conformity is advantageous.

3. Evolutionary Stability and the Genesis of Contagion

The conformist relationships described above – and illustrated in Figures 2 and 3 – imply two evolutionarily stable strategies (ESS), m=1 (all honest) and m=0 (all dishonest). This is true whether the honesty mutation evolves in function space or in parameter space. In both cases, the argument is the standard one: if h>h*, then

$$V(1,h) = E(x\,|\,H,h) > E(x\,|\,D,h) = V(0,h);$$

Hence, honesty has higher fitness and the prevalence of honesty (whether the honesty parameter m=1, or the pure honesty function m(h)=1 for all h, or the "contagious" honesty function) therefore increases over time; that is, h rises. Conversely, if h<h*, h falls. Only when h=0 or h=1 is h stable.[8] Since both ESS can evolve and be sustained in parameter space, why should the mutation evolve as a "contagion" function? I sketch a possible answer here.

Consider two groups that, in an initial period, evolve separately into the two polar ESS equilibria: h=1 in group 1 and h=0 in group 2. In a second period, members of the two groups are exposed to profitable opportunities for projects with members of the other group. Specifically, as before, each group member is randomly matched with a player from the same group for a project (the PD game); however, unlike before, a subset of group members (randomly chosen) has additional opportunities with members of the

[8] There is a vast body of literature on evolutionary games that embed these dynamics. See, for example, Weibull (1995), Guth and Kliemt (1994), Sobel (2005), Bergstrom (2002).

other group.[9] I will assume that there are gains from trade; formally, let G = (x_1, x_2, x_3, x_4) denote "raw" payoffs in the PD game, and assume that within-group partnerships (group 1, group 2, and the "interactive" group) have diminishing returns in the sense that payoffs equal $\delta(n)G$, with n=number of partnerships, $\delta'<0$, and $\delta(0)$ arbitrarily large.[10] In addition, I will assume (at least for the interactive group) that $x_1=0$, so that exploitation payoffs are never advantageous.

Of course, with the initial ESS attributes in pure (parameter) form (m=1 in group 1 and m=0 in group 2) and $x_1=0$, the potentially profitable inter-group opportunities are not actually profitable because no one from group 1 (all of whom are honest) will agree to partner with anyone from group 2 (all of whom are dishonest). However, consider the impact of a contagious honesty mutation in group 1. The first such mutant can earn the intergroup payoff $\delta(0)x_2$ with a group 2 partner, compared with $\delta(n_1/2)x_3$ with a group 1 partner (where n_1=population of group 1). With i denoting the probability that a group 1 member is exposed to an inter-group opportunity, the expected payoff to the contagious mutant is higher than to his "purely honest" compatriot:

Expected payoff to contagious group 1 mutant

$$= (1-i)\ \delta(n_1/2)x_3 + i\ \delta(0)x_2$$

$$> \delta(n_1/2)x_3 = \text{payoff to "purely honest" group 1 member,}$$

where the inequality is due to $\delta(0)$ being sufficiently large. Indeed, the contagion mutation is superior to any non-equivalent mutation so long as i is not equal to zero or one. Hence, evolutionary pressure will lead to the survival and increased prevalence of the mutation. As the number of the mutations M increases, the advantage declines until it vanishes:

$$\delta(iM^*)\ x_2 = \delta((n_1-iM^*)/2)x_3\ , \tag{15}$$

where iM* is the number of mutants who are exposed to (and accept) profitable inter-group partnership opportunities.[11]

Over time, the group members (n_1 and n_2) will change with migration and/or evolution. For example, with payoff monotonic dynamics (Weibull, 1995; Bergstrom, 2002), higher payoff populations reproduce more rapidly, a standard premise in evolutionary modeling. However, abundant evidence in population economics documents the inverse relationship between economic payoffs and reproductive rates. In view of this evidence, I invoke a more direct mechanism for population change, namely, migration incentives. In the present context, this mechanism would take the following form, assuming group membership cannot be held closed in the long run: In view of higher payoffs to group 1 members, a group 2 contagious mutation becomes advantageous. Unlike an unmutated / dishonest group 2 member, the contagious mutant can benefit from migration to group 1 – becoming honest in within-group-1 partnerships and thus enjoying the higher group 1 expected payoff. Indeed, the contagious mutation is likely to arise simultaneously in both groups – in group 1 to enable inter-group

[9] For simplicity, I suppose that inter-group opportunities are revealed to individuals, and then assortive matching of within-group members occurs after inter-group opportunities have been accepted or rejected.

[10] I avoid notational clutter by positing a common δ function across groups. Nothing changes qualitatively with the addition of group-specific functions.

[11] The implicit definition of M* presumes that i is sufficiently high (but less than one) so that $M^*<n_1$ in the solution to equation (15).

partnerships and in group 2 to enable advantageous migration. Notably, this logic suggests that migrants are more likely to be contagious.

Whether due to evolutionary pressure or migration incentives, n_1 will rise and n_2 will fall until intergroup payoffs are equalized:

$$\delta((n_2\text{-}iM^*)/2)x_2 = \delta((n_1\text{-}iM^*)/2)x_3 .$$

In sum, there is a stable evolutionary intergroup equilibrium in which (i) players are dishonest / non-cooperative in group 2 and in intergroup partnerships, (ii) a subset of group 1 members, $M^*<n_1$, has the "contagious" mutation, (iii) remaining group 1 members are either contagious or "purely honest" (m=1 for all h), and (iv) expected payoffs are equalized with a combination of within-group and intergroup partnerships. This is not the only stable equilibrium; for example, another equilibrium involves honest/cooperative behavior in intergroup partnerships and contagious mutation in the dishonest group 2.[12] In either of these cases, the contagious mutation evolves and survives in a subset of the population.

4. Competing Preference Types

A. The Nature of Deception Aversion

So far (following Frank, 1987), I have assumed that an honest individual's aversion to dishonesty / defection in the PD game makes honesty / cooperation a dominant strategy. In particular, let a_s denote the honest person's monetary-equivalent penalty to dishonesty when the partner's strategy is $s \,\varepsilon\, \{H,D\}$. I assume that $a_H>x_4\text{-}x_3>0$ and $a_D>x_2\text{-}x_1>0$. Let us now call these preferences "absolute honesty," H_A.

There is an alternative: $a_H>x_4\text{-}x_3$ and $a_D=0$. Then aversion to defection only arises when the partner is honest / cooperative.[13] Let us call these preferences "conditional honesty," H_C. In principle, H_C preferences could arise and supplant H_A preferences. What effect does this possibility have on my arguments?

Consider the model with perfect signals of type, $q \,\varepsilon\, \{D, H_A, H_C\}$, and no resampling. When conditionally honest types meet each other, there are three equilibria to the normal form game, two pure strategy (both defect (D), or both cooperate (H)) and one mixed strategy (cooperate with probability $\rho^*=(x_2\text{-}x_0)/[(x_3\text{-}x_0)+(x_2\text{-}x_0)]$). Because the mixed strategy equilibrium is unstable, I will focus on the other two.[14]

Suppose first that the advantageous (H,H) equilibrium prevails whenever there is an (H_C,H_C) pair. Letting h denote the sum of population proportions of H_A types (h_A) and H_C types (h_C), expected payoffs to H_C types, $hx_3+(1\text{-}h)x_2$, are (i) higher than for H_A types except when h=1, and (ii) higher than for D types except when h=0. Hence, for this case, there is no role for contagion; indeed, conditionally honest preferences simply dominate. The reason is that conditional honesty permits an individual to gain the benefits of

[12] Of course, completely honest or completely dishonest equilibria are also possible, but are uninteresting for our purposes.

[13] There is no reason to consider $a_H \,\varepsilon\, (0,x_4\text{-}x_3)$ or $a_D \,\varepsilon\, (0,x_2\text{-}x_1)$ because these values imply that deception aversion does not alter equilibrium strategies but reduces equilibrium payoffs and, hence, fitness. As a result, such values will not arise in an evolutionary equilibrium.

[14] If your rival is playing the mixed strategy ρ, your payoff to the (pure strategy) H is $\rho x_3+(1\text{-}\rho)x_0$ and your payoff to (pure strategy) D is $\rho x_0+(1\text{-}\rho)x_2$. ρ^* is unstable in the sense that, for permutations of your expectation of your rival's strategy ρ away from ρ^*, your optimal response is the pure strategy H (if $\rho>\rho^*$) or D (if $\rho<\rho^*$).

honesty whenever encountering another honest player without sacrificing anything when meeting a dishonest player.

If instead the disadvantageous (D,D) equilibrium prevails when two H_C partners meet, the conclusion is similar but nuanced. Here, H_C preferences dominate D preferences, but H_A preferences dominate H_C preferences as h approaches 1. The unique equilibrium point is again h=1, only now with H_A preferences, and there is again no role for contagion.

These conclusions, however, are not robust. Let us suppose (as is common in the literature) that H_A and H_C preferences permit strategic mistakes. Specifically, if any H type intends strategy s ε {H,D}, he/she mistakenly chooses the other (unintended) strategy with probability u (<1/2).[15] D-type players, I assume, continue to pursue their defect strategies without mistake (also a common premise, as in Henrich and Boyd, 2001). Once strategies (and mistakes) have been revealed, I assume that players can again exit the game and receive their outside option payoff x_0. Equilibrium expected payoffs under different partnerships are then as described in Table 1.

Now consider when the (H_C, H_C) partners support the favorable Nash Equilibrium (so that ρ=1). Then when facing H_C partners, H_C types have an advantage over H_A types because they can sometimes adapt to mistakes by playing the (D,D) outcome (and thereby obtaining x_2), which the H_A types cannot do (due to their aversion to their own defection); for the same reason, they are at a relative advantage (vs. H_A counterparts) when facing a D-type partner. H_C preferences therefore dominate H_A preferences, as they did in the model without mistakes. However, the key difference here is that, relative to D-type players, H_C players are at a disadvantage when facing D-type partners due to the H_C players' mistakes. As a result, when h=0 (virtually all players are dishonest), D preferences yield higher fitness than H_C preferences. Conversely, when h=1 – when H_C players, subject to mistakes, support the favorable (H,H) equilibrium – H_C preferences have higher fitness than D counterparts.[16] Because the relative fitness of H_C (vs. D) players is monotone in h, we again have a motive for contagion, with one central difference: the "honest" preferences are now conditional.

When (H_C, H_C) partnerships support the unfavorable Nash Equilibrium (so that ρ=0), the characterization of fitness advantages is somewhat more complicated. Now H_A players – rather than H_C players – have the advantage when facing H_C partners. As a result, as the overall propensity for honesty goes to one (h=1), H_A preferences now yield higher fitness than either H_C or D preferences. As h goes to zero, the conditional H_C preferences are still better than the absolute H_A counterparts, but both are dominated by D-type preferences that do not suffer costs of mistakes.

I now turn to the characterization of fitness-maximizing preferences for each level of h, which then implies equilibrium points in h. For higher h values, it is easily seen that there is an interior proportion of H_A types, $h_A = h_A^+$ ($0 < h_A^+ < h$) that is stable: when $h_A > h_A^+$, it is more advantageous to be an H_C type ($E(x \mid H_C, h_A, h) > E(x \mid H_A, h_A, h)$); and when $h_A < h_A^+$, it is more advantageous to be H_A.

[15] One might expect those with conditional preferences to be more subject to mistakes. However, so as not to explicitly disadvantage H_C (vs. H_A) preferences, I assume both are subject to mistakes.

[16] The qualification to this statement is the inequality, $x_3^{**} > x_2^{*}$, hold. The following (plausible) condition implies this inequality: $(x_3 - x_2) > [u(1-2u)/(1-u)^2](x_2 - x_0)$.

Formally, for each $h\varepsilon[0,1]$, define a stable distribution of the honesty attribute $a\varepsilon\{A,C\}$, $h_A^+\varepsilon[0,h]$ such that

$$V_H(h) = \{\mathrm{argmax}_{a\varepsilon\{A,C\}} \{ E(x \mid H_a, h_A^+, h) \} \} \qquad \begin{array}{lll} = A & \leftrightarrow & h_A^+=h \\ = \{A,C\} & \leftrightarrow & h_A^+\varepsilon[0,h] \\ = C & \leftrightarrow & h_A^+=0 \end{array}$$

Next define the fitness-maximizing cooperation attribute, $q\ \varepsilon\{H,D\}$,

$$q^*(H) = \mathrm{argmax}_{q\ \varepsilon\{H,D\}} \{ E(x \mid D, h_A^+, h), V_H(h)\}.$$

Together, $\{q^*(h), V_H(h), h_A^+(h)\}$ define a *stable preference correspondence*.

Proposition 2. Suppose that payoffs are as given in Table 1, with $\rho=0$, and that u is sufficiently small that: $[u^2/(1-2u)]\ (x_2-x_0) < (x_3-x_2) < [(1-2u)/u](x_2-x_0)$. There is an $h^*\varepsilon(0,1)$ such that the following is a stable preference correspondence:

$$\begin{array}{ll} q^*(h) & = D \text{ for } h<h^* \\ & = H \text{ for } h\geq h^*, \end{array}$$

and $V_H(h) = \{A,C\}$ for $h\geq h^*$, where

$$h_A^+(h) = h\ (1+k) - k\ \varepsilon\ (0,h] \quad , \quad k = (1-u)\ (x_2-x_0)/\ (x_3^*-x_2^{**}) > 0.$$

Proposition 2 indicates, again, that it is advantageous to be dishonest when there are few honest people, and to be honest when there are many. It also indicates that there are two equilibrium points, one with h=0 and D-type preferences, and the other with h=1 and H_A type preferences. Our prior arguments on contagion thus apply.

B. Guilt vs. Vengeance

The Prisoner's Dilemma can be overcome if guilt – an aversion to a non-cooperative / non-moral strategy – motivates cooperation (vs. defection), as assumed in this paper. Alternatively, vengeful preferences – which prompt punishment of defectors – can motivate cooperation. As economists who work on this subject can attest, vengeance is complicated (see, for example, Henrich and Boyd, 2001; Friedman and Singh, 2008). In what follows, I present an illustrative model showing that vengeance, even when it arises in an equilibrium, need not upset my main conclusion on moral contagion.

Suppose that there are four types of players, distinguished by whether they are (a) "honest" (H) or "selfish" (D) and (b) vengeful (v) or not (N). We will denote these types by H_v, H_N, S_v and S_N, respectively. If an individual is vengeful, he/she intentionally punishes a partner's advantageous defection. That is, if the partner (Other) plays D when the vengeful individual (Self) plays H, then Self imposes a punishment of p>0 on Other which yields Self a net (illicit) benefit b=0 but also a fitness cost of v>0 (where v≤p).[17] In principle, vengeance could imply intentional punishment whenever Other defects (even when Self also defects); I instead assume conditional punishment which, together with a favorable equilibrium assumption that I make in a moment, gives a maximum possible advantage to vengeful preferences.

Absent any mistakes in punishment, and assuming that attributes are fully observable, it is easily seen that an honest and vengeful person will achieve the optimal (H,H) equilibrium with any individual he/she encounters, so long as punishment is effective at all – that is, so long as $p>(x_4-x_3)$. Moreover, in this case, punishment is never

[17] The assumption that the illicit benefit does not affect fitness is standard, recognizing the fitness cost of vengeance.

actually made and, hence, is costless. Of course, this situation motivates only H_v type preferences and no contagion.

More interesting and realistic are situations when vengeance leads to mistaken punishments. I assume that, when punishment is intended, it is rendered without mistake. However, vengeful individuals (H_v and D_v) also punish by mistake – when they do not intend to – with probability e (<1/2).[18] I will continue to assume that punishment is effective, $p>(x_4-x_3)/(1-e)$, but not perfectly effective.[19] That is, the selfish D-types, even when they intend to be honest due to prospective punishment, make mistakes; they then mistakenly defect with probability u (<1/2) and are honest with probability (1-u). For simplicity, I assume that these are the only strategic mistakes,[20] and that when these mistakes occur, they cannot be corrected but the other player can revise strategy to defect (D) or withdraw from the game.[21] I also assume that parameter values (e,u,p,v) are such that vengeful preferences can be advantageous to H_v players (when they meet D-type partners) and to D_v players (when they meet D_v partners):

Assumption 1. (i) $(1\text{-}u)\Delta_3 - up \geq 0$, where $\Delta_3 = x_3\text{-}ep\text{-}ev\text{-}x_0$; (ii) $\beta\,\Delta_3 + (1\text{-}\beta)\,\Delta_2 \geq 0$, where $\beta = (1\text{-}u)^2$ and $\Delta_2 = x_2\text{-}ev\text{-}x_0$; and (iii) $ev \leq ep < x_3\text{-}x_2 < x_2\text{-}x_0 \leq 2ev$.

The last assumption is made to ensure an advantage to vengeance, but bound this advantage.

For these circumstances, Table 2 indicates the Nash Equilibrium (NE) strategies and fitness payoffs for the one-shot PD game. When D_v types face each other, there are two possible Nash Equilibria; for simplicity – and to give vengeance a maximum possible advantage – I assume that the players can choose the best NE (H,H), for example by playing sequentially.[22]

Now let h_v, h_N, s_v and $s_N=1\text{-}h_v\text{-}h_N\text{-}s_v$ equal the population shares of attributes H_v, H_N, D_v and D_N, respectively. Further, let $h^*_v = h_v/h$ and $s^*_v = s_v/(1\text{-}h)$ denote the respective shares of H and D-types that are vengeful, respectively, where $h=h_v+h_N$. Armed with Table 2, we can now express expected fitness payoffs to the different preference types as follows:

$$(16a)\quad E(x\,|\,H_v, h^*_v, s^*_v, h) = (x_3\text{-}ev)\,[1 - u(1\text{-}h)] - ep\,(h_v+s_v(1\text{-}u)) + u(1\text{-}h)(x_0\text{-}v)$$

$$(16b)\quad E(x\,|\,H_N, h^*_v, s^*_v, h) = h\,x_3 - ep\,h_v + (1\text{-}h)\,x_0$$

$$(16c)\quad E(x\,|\,D_v, h^*_v, s^*_v, h) = (x_3\text{-}ev\text{-}ep)\,(1\text{-}u)\,[h_v + s_v(1\text{-}u)] + x_0\,(uh_v+h_N) - h_v up + [(1\text{-}h)\text{-}s_v(1\text{-}u)^2](x_2\text{-}ev)$$

[18] The probability e may be more than a tiny fraction. For example, suppose that payoffs are all random, each drawn from the two point distribution {0,X}. It is reasonable to suppose that when a vengeful individual draws a zero payoff, even though this is no fault of his partner, he/she sometimes gets mad and mistakenly acts vengefully. If so, the probability of mistake is proportional to the probability of a zero outcome, which may be substantial.

[19] If punishment is not effective, it is clearly disadvantageous. Moreover, as higher punishments will cost more in fitness, an optimal punishment will be just barely effective, $p \approx (x_4\text{-}x_3)/(1\text{-}e)$.

[20] Honest individuals are thus honest without mistake, and defectors defect without mistake.

[21] I thus add a "strategy revision" stage of the game, and assume mistakes are persistent in this stage. Punishments are made (or not) after the revision stage.

[22] Note also that Table 2 assumes that mistaken defections are punished, even if the wronged partner withdraws from the game. All that follows is robust to the absence of punishment in such cases.

(16d) $E(x \mid D_N, h_v^*, s_v^*, h) = (x_3\text{-}ep)\ (1\text{-}u)\ h_v + x_0\ (uh_v+h_N) - h_vup + (1\text{-}h)x_2 - s_vep$
where $h_v= h_v^*h$, $h_N=(1\text{-}\ h_v^*)h$, and $s_v= s_v^*(1\text{-}h)$. Studying these expressions reveals the following.

Observation 1. (a) If $h=h_v=0$ and $s_v=1$, then (i) $E(x \mid H_v, .) > E(x \mid H_N, .)$; (ii) $E(x \mid D_v, .) > E(x \mid D_N, .)$; and (iii) $E(x \mid D_v, .) > E(x \mid H_v, .)$. (b) If $h=1$, then (i) $E(x \mid H_N, .) > E(x \mid H_v, .)$; (ii) $E(x \mid D_N, .) \geq E(x \mid D_v, .)$; and (iii) $E(x \mid H_N, .) > E(x \mid D_N, .)$. (c) Incentives for the preference choice H_v, vs. H_N, decline with h and s_v. (d) Incentives for the preference choice D_v, vs. D_N, rise with s_v, fall with h_v (for given h), and fall with h (when $s_v=1\text{-}h$)); in addition, a positive s_v is necessary for D_v preferences to be preferred to D_N preferences.

Observation 1(a)-(b) considers two extremes (which turn out to be the relevant equilibrium points), namely, when there are no honest people (h=0) and when all people are honest (h=1). In the former case, vengeful preferences are advantageous (under our assumptions) because they enable selfish partners to be persuaded to play cooperatively. Moreover, selfish (and vengeful) preferences are more advantageous than honest (and vengeful) preferences. The reason is that honest types suffer more from the mistakes of their selfish partners than do selfish types. H types withdraw from the game when faced with a (mistaken) defection by their selfish partner, whereas D-types adapt by themselves defecting. As before, the latter strategy is more advantageous ex-post, again giving rise to incentives for like types to pair with one another.

Conversely, when all are honest (h=1), non-vengeful preferences are more advantageous because there is no need for punishment to elicit cooperation, and punishment has costs. Moreover, as before, honest (non-vengeful) types are better off than selfish (non-vengeful) counterparts because the selfish individuals are turned out from the PD game.

As I will describe in more detail in a moment, these properties imply the optimality of contagion similar to that discussed earlier, with a twist. Now more honesty in the population snuffs out vengeance, and selfishness breeds vengeance.

Observation 1(c)-(d) describe how incentives for vengeance vary with population propensities for honesty and vengeance, respectively. As h rises, incentives for vengeance decline because there are fewer potential selfish partners in need of vengeful persuasion to elicit cooperation. Perhaps more interesting is that incentives for D-types to be vengeful are "contagious" – that is, they rise the propensity for vengefulness s_v. This is because, for D-types, vengeful preferences are only effective – they only elicit cooperative outcomes that would not otherwise occur – when their partner is also a vengeful D-type.

In principle, a complete (contagious) assignment rule defines a mapping from the population propensities, (h, h_v, s_v) to the domain of preference types, $(q\varepsilon\{H,D\},V\varepsilon\{v,N\})$, so as to maximize fitness:

$$(q^*,V^*) = \text{argmax}_{\ q\varepsilon\{H,D\},V\varepsilon\{v,N\}}\ E(x \mid qv.)$$

Rather than describing this full assignment rule, I will examine, for each h, the possible equilibrium points for choice of the vengeance attribute, which will in turn permit identification of possible ESS equilibria for the full range of preference types.

Formally, for each $h\varepsilon[0,1]$ and each $q\varepsilon\{H,D\}$, define a stable set of vengeful preferences, $V_q\varepsilon\{N=0,v=1\}$, such that

(17) $V_q = \text{argmax}_{\,V\varepsilon\{N,v\}}\ E(x \,|\, qv,\ h^*_v=V_H,\ s^*_v=V_D,h)$

That is, given the equilibrium (V_H, V_D), V_H and V_D are fitness-maximizing choices of vengeful preferences for individual H and D types, respectively. Next, given the stable preferences defined in (2), $V_q(h)$, we can define the fitness-maximizing "cooperation attribute," $q\varepsilon\{H,D\}$:

(18) $q^*(h) = \text{argmax}_{\,q\varepsilon\{H,D\}}\ E(x \,|\, qv,\ h^*_v=V_H,\ s^*_v=V_D,h)$

I will define a stable preference correspondence as one that satisfies (17)-(18).

Because vengeance by other D-types is necessary for a vengeful preference to be optimal for individual D-types, it is easily seen that $V_D=0$ for all h is stable. I will focus on "vengeance maximizing" stable outcomes, those that involve $V_D=1$ when such a stable outcome exists. Following some preliminary observations, I can now state the main result of this Section.

Observation 2. (a) There is an $h_0\varepsilon(0,1)$ such that vengeful preferences are advantageous for H-types if and only if h is less than h_0:

(19) $E(x \,|\, H_N,\ h^*_v=1,\ s^*_v=1,h) = , < , > E(x \,|\, H_v,\ h^*_v=1,\ s^*_v=1,h)$

for $h=h_0$, $h_0<0$, and $h>h_0$, respectively. In addition, for all lower h, $h\varepsilon[0,h_0]$, vengeance is advantageous for D-types,

(20) $E(x \,|\, D_v,\ h^*_v=1,\ s^*_v=1,h) - E(x \,|\, D_N,\ h^*_v=1,\ s^*_v=1,h) > 0,$

and, for all higher h, $h\varepsilon[0,1)$, vengeance is also advantageous for D-types,

(21) $E(x \,|\, D_v,\ h^*_v=0,\ s^*_v=1,h) - E(x \,|\, D_N,\ h^*_v=0,\ s^*_v=1,h) > 0.$

(b) A higher proportion of vengeful H-types, h^*_v, is disadvantageous to H-types and advantageous to D-types:

(22) $E(x \,|\, H_v,\ h^*_v=0,\ s^*_v=1,h) - E(x \,|\, H_v,\ h^*_v=k\ \varepsilon\ (0,1],\ s^*_v=1,h) = ephk \geq 0$,

(23) $E(x \,|\, D_v,\ h^*_v=1,\ s^*_v=1,h) - E(x \,|\, D_v,\ h^*_v=0,\ s^*_v=1,h) = \{(1-u)\Delta_3-up\}h \geq 0;$

(c) There is an $h_1\varepsilon(0,1)$:

(24) $E(x \,|\, H_v,\ h^*_v=1,\ s^*_v=1,h) = , < , > E(x \,|\, D_v,\ h^*_v=1,\ s^*_v=1,h)$

for $h=h_1$, $h_1<0$, and $h>h_1$, respectively.

(d) There is an $h_2\varepsilon(0,1)$:

(25) $E(x \,|\, H_v,\ h^*_v=0,\ s^*_v=1,h) = , < , > E(x \,|\, D_v,\ h^*_v=0,\ s^*_v=1,h)$

for $h=h_2$, $h_2<0$, and $h>h_2$, respectively. Further, $h_2\leq h_1$.

Proposition 3. There are three possible vengeance-maximizing stable preference correspondences (VSPC): (1) If $h_1\leq h_0$, then

$$\{q^*(h),V_{q^*}(h)\} = \begin{array}{l} (D,v) \text{ for } h\leq h_1 \\ (H,v) \text{ for } h\varepsilon(h_1,h_0) \\ (H,N) \text{ for } h\geq h_0 \end{array}$$

(2) If $h_1>h_0$ and $h_2\leq h_0$, then

$$\{q^*(h),V_{q^*}(h)\} = \quad (D,v) \text{ for } h\leq h_0 \text{ and } (H,N) \text{ for } h>h_0.$$

(3) If $h_1>h_0$ and $h_2>h_0$, then

$$\{q^*(h),V_{q^*}(h)\} = \quad (D,v) \text{ for } h\leq h_2 \text{ and } (H,N) \text{ for } h>h_2.$$

Figure 4 graphs the VSPC for the first case. Note that, in all cases, there are incentives for contagion in both “honesty” and vengeance in the following sense. Honesty is more advantageous when there is a higher propensity for honesty in the population (and vice versa), and vengeance is more advantageous when there is a higher propensity for both selfishness and vengefulness in the population. The three cases give rise to two ESS equilibrium points, one in which all are selfish and vengeful (D_V) and another in which all are honest and non-vengeful (H_N). The logic above for population interactions argues for a fitness advantage to contagion in these two polar preference types.

Intuitively, vengeance is more advantageous when selfishness is more prevalent because (i) vengeance is only needed to elicit cooperation from D-type partners, not from the H-types (who cooperate anyway), and (ii) it is costly, so only advantageous when there are a lot of D-type people. Honesty is more advantageous when there are more honest people for essentially the same reasons as before: There are fewer potential D-type partners that lead to the breakdown of cooperative efforts; likewise, dishonesty/selfishness is less advantageous under these circumstances because such individuals are more likely to be matched with H-types who reject the partnership. Conversely, honesty is relatively less advantageous when D-types are prevalent, even with vengeful preferences, because mistakes by D-types who intend to cooperate, but don’t are more costly to the H-types who do not adapt by themselves defecting and thereby bringing about a (less advantageous) symmetric equilibrium to the PD game, but still an equilibrium that is more advantageous than the go-it-alone option (with $x_2>x_0$); the D-types, in contrast, adapt and thus suffer less from their own and their partner’s mistakes.

Appendix

Proof of Proposition 2. First we have:

(A1) $E(x \mid H_A,h_A,h) - E(x \mid H_C,h_A,h) = (h-h_A)\,(x_3^*-x_2^{**}) - (1-h)(1-u)\,(x_2-x_0)$

< 0, all $h_A \varepsilon [0,h]$, for $h<h^{**} = [(1-u)\,(x_2-x_0)]/[\,(x_3^*-x_2^{**}) + (1-u)(x_2-x_0)]$

$<(=,>)\ 0$ for $h \geq h^{**}$, $h_A >(=,<)\ h(1+k)-k\ \varepsilon\ [0,h]$

where $(x_3^*-x_2^{**}) > 0$ (by the lefthand inequality assumed in the Proposition). (A1) directly implies that

(A2) $h_A^+(h) = 0$ for $h \leq h^{**}$

$= h(1+k)-k\ \varepsilon\ [0,h]$ for $h>h^{**}$

By (A2),

$V_H(h) = E(x \mid H_C, h_A^+(h),h)$

and to derive q*(h), it suffices to evaluate

(A3) $\Delta(h) = E(x \mid H_C, h_A^+(h),h) - E(x \mid D, h_A^+(h),h)$

$= h_A^+ \{(x_3^*-x_2^{**}) + (x_2^*-x_0)\} + h\{(\,x_2^{**} - x_2^*) + (x_2 - x_2^*)\} - (x_2 - x_2^*)$

where $(x_3^*-x_2^{**}) > 0$, $(x_2^* - x_2^{**}) > 0$ (righthand inequality assumed in the Proposition), $(x_2^* - x_0) > 0$, $(x_2 - x_2^*) > 0$, and

$\{(\,x_2^{**} - x_2^*) + (x_2 - x_2^*)\} = u^2\,[x_2+x_3-2x_0] > 0.$

Hence,

(A4) $d\Delta/dh = (\partial h_A^+/\partial h)\ \{(x_3^*-x_2^{**}) + (x_2^*-x_0)\} + \{(x_3^*-x_2^{**}) + (x_2 - x_2^*)\} > 0,$

where $(\partial h_A^+/\partial h) = 0$ for $h<h^{**}$, $=1+k>0$ for $h \geq h^{**}$.

(A5) $\Delta(h^{**}) = h^{**}\,(\,x_2^{**} - x_2^*) - (1-h)\,(x_2 - x_2^*) < 0.$

(A6) $\Delta(1) = x_3^* - x_0 > 0.$

Together with the Intermediate Value Theorem, (A4)-(A6) imply that there is an $h^* \varepsilon (h^{**},1)$ such that $\Delta=0$ at $h=h^*$, $\Delta<0$ for all $h<h^*$, and $\Delta>0$ for $h>h^*$. The Proposition now follows. QED.

Proof of Observation 1. Using (16), we have

(A7) $E(x \mid D_v, .) - E(x \mid D_N, .) = Z_0 - h\,[Z_0 + (1-u)\,h_v^* ev],$

where $Z_0 = [(x_3-ep-x_2)\,(1-u)^2+ep]\,s_v^* - ev$; and $Z_0>0$ when $s_v^*=1$ (by Assumption 1(iii));

(A8) $E(x \mid H_N, .) - E(x \mid H_v, .) = h[\,Z_2 + ev + (1-u)(1- s_v^*)ep]$

$-\quad [\,Z_2 + (1-u)(1- s_v^*)ep],$

where $Z_2 = (1-u)\Delta_3 - uv \geq 0$ (by Assumption 1(i));

(A9) $E(x \mid D_v, h_v^*=1, s_v^*=1,h) - E(x \mid H_v, h_v^*=1, s_v^*=1,h) = Z_1 - h\,[Z_1 + p + \Delta_3],$

where $Z_1 = (2-u)\,)\Delta_2 + v - (1-u)\,)\Delta_3 > 0$ (by Assumption 1(iii), $x_3-x_2<x_2-x_0$, and $\max(e,u)<1/2$). The Observation follows directly from (A7)-(A9). QED.

Proof of Observation 2. (a) Eq. (19) follows from Observation 1(a)-(b) and continuity; from equation (A8), we have (using Assumption 1(ii)),

$h_0 = Z_2 / (Z_2+ev)$, $Z_2 > 0$.

Equation (20) follows from monotonicity of

$E(x \mid D_v, h_v^*=1, s_v^*=1,h) - E(x \mid D_N, h_v^*=1, s_v^*=1,h)$

in h (equation (A7)) and

$$E(x \mid D_v, h_v^*=1, s_v^*=1,h_0) - E(x \mid D_N, h_v^*=1, s_v^*=1,h_0) \overset{s}{=} Z_0 - (1-u)Z_2$$
$$= - (1-u)^2(x_2-x_0)+ep+v(1-u)(u+(1-u)e) \geq ve(1-(1-u)_2)+u(1-u)v > 0$$

where the first inequality is due to Assumption 1(iii). Equation (21) follows from:

$$E(x \mid D_v, h_v^*=0, s_v^*=1,h) - E(x \mid D_N, h_v^*=0, s_v^*=1,h) = Z_0 (1-h),$$

where Z_0 is as defined in the proof of Observation 1.

(b)-(c) follow from (16), Assumption 1(i), and (A9).

(d) follows from continuity, the Intermediate Value Theorem, and the following inequalities: (i) at h=1,

$$E(x \mid H_v, h_v^*=0, s_v^*=1,h) > E(x \mid H_v, h_v^*=1, s_v^*=1,h) > E(x \mid D_v, h_v^*=1, s_v^*=1,h)$$
$$> E(x \mid D_v, h_v^*=0, s_v^*=1,h),$$

where the first and last inequalities are due to Observation 2(b) above, and the second inequality is due to equation (A9); (ii) at h=0, by the same logic,

$$E(x \mid H_v, h_v^*=0, s_v^*=1,h) = E(x \mid H_v, h_v^*=1, s_v^*=1,h) < E(x \mid D_v, h_v^*=1, s_v^*=1,h)$$
$$= E(x \mid D_v, h_v^*=0, s_v^*=1,h).$$

Finally, at $h=h_1$, we have

$$E(x \mid H_v, h_v^*=0, s_v^*=1,h) > E(x \mid H_v, h_v^*=1, s_v^*=1,h) = E(x \mid D_v, h_v^*=1, s_v^*=1,h)$$
$$> E(x \mid D_v, h_v^*=0, s_v^*=1,h),$$

by Observation 2(b) above and the definition of h_1 in (c) above. $h_2 \leq h_1$ now follows from the definition of h_2 in (d). QED.

Proof of Proposition 3. Proof for Proposition 3 part (1) (parts (2) and (3) are similar and thus omitted): First note that, by Observation 2(d), we have $h_2 \leq h_1 \leq h_0$. Now consider $h \geq h_0$. By Observation 2(a)-(b), V=N=0 yields a higher fitness payoff for the H-types when $s_v^*=V_D=1$. Similarly, by Observation 2(a), V=v=1 yields a higher fitness payoff for the D-types when $h_v^*=V_H=0$. Hence, a VSPC sets $V_H(h)=0(=N)$ and $V_D(h)=1(=v)$. Further,

$$\text{(A10)} \quad E(x \mid H_N, h_v^*=0, s_v^*=1,h) \geq E(x \mid H_v, h_v^*=0, s_v^*=1,h) > E(x \mid H_v, h_v^*=1, s_v^*=1,h)$$
$$\geq E(x \mid S_v, h_v^*=1, s_v^*=1,h) > E(x \mid S_v, h_v^*=0, s_v^*=1,h).$$

The first inequality is due to Observation 2(a) and equation (A8) (showing that $E(x \mid H_N,.)$- $E(x \mid H_N, .)$ is invariant to h_v^*). The second is due to Observation 2(b). The third follows from Observation 2(c) and $h_1 \leq h_0 \leq h$. The fourth follows from Observation 2(b). By (A10), $\{q^*(h),V_{q^*}(h)\} = (H,N)$ for $h \geq h_0$.

For $h<h_0$, Observation 2(a) implies that a VSPC sets $V_H(h)=1(=v)=V_D(h)$. Further, for $h \geq h_1$, Observation 2(c) implies that, given $V_H(h)=V_D(h)=1$, q*(h)=H. And, for $h<h_1$ $(\leq h_0)$, Observation 2(c) implies that q*(h)=D. QED.

References

Akerlof, G. “A Theory of Social Custom, of Which Unemployment May Be One Consequence.” *Quarterly Journal of Economics* 94 (1980): 749-75.

Banerjee, A. “A Simple Model of Herd Behavior.” *Quarterly Journal of Economics* 107 (1992): 797-817.

Banerjee, A., and T. Besley. “Peer Group Externalities and Learning Incentives: A Theory of Nerd Behavior.” John M. Olin Discussion Paper no. 68. Princeton, N.J.: Princeton University, December 1990.

Bergstrom, T. “Evolution of Social Behavior: Individual and Group Selection.” *Journal of Economic Perspectives* 16 (2002): 67-88.

Bernheim, B.D.. “A Theory of Conformity.” *Journal of Political Economy* 102 (1994): 841-77.

Besley, T., and S. Coate. “Understanding Welfare Stigma: Taxpayer Resentment and Statistical Discrimination.” *Journal of Public Economics* 48 (1992): 165-83.

Bicchieri, C., and E. Xiao. “Do the Right Thing: But Only if Others Do So.” University of Pennsylvania Working Paper, Philosophy, Politics and Economics Program, August 2007.

Bikhchandaria, S., D. Hershleifer, and I. Welch. “A Theory of Fads, Fashion, Custom, and Cultural change as Informational Cascades.” *Journal of Political Economy* 100 (1992): 992-1026.

Bolton, G. and A. Ockenfels. “ERC: A Theory of Equity, Reciprocity, and Competition.” *American Economic Review* 90 (2000): 166-93.

Charness, G., and M. Dufwenberg. “Promises and Partnership.” *Econometrica* 74 (2006): 1579-1601.

Charness, G. and M. Rabin. “Understanding Social Preferences with Simple Tests.” *Quarterly Journal of Economics* 117 (2002): 817-69. 2002.

Cox, J. “How to Identify Trust and Reciprocity.” *Games and Economic Behavior* 46 (2004): 260-81.

Fehr, E., and U. Fischbacker. “Social Norms and Human Cooperation.” *Trends in Cognitive Sciences* 8 (2004): 185-90.

Fehr, E., and S. Gachter. “Fairness and Retaliation: The Economics of Reciprocity.” *Journal of Economic Perspectives* 14 (2000): 159-81.

Fehr, E., and S. Gachter. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review* 90 (2000): :980-94.

Fehr, E., and S. Gachter. “Altruistic Punishment in Humans.” *Nature* 415 (2002): 137-40.

Fehr, E., S. Gachter, and G. Kirchsteiger. “Reciprocity as a Contract Enforcement Device.” Econometrica 65 (1997): 833-60.Fehr, E. and K. Schmidt. “A Theory of Fairness, Competition, and Cooperation.” *Quarterly Journal of Economics* 114 (1999): 817-68.

Frank, R. “If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience?” *American Economic Review* 77 (1987): 593-604.

Friedman, D., and N. Singh. “Equilibrium Vengeance.” Working Paper, Economics Department, U.C. Santa Cruz, 2008.

Gneezy, U. “Deception: The Role of Consequences.” *American Economic Review* 95 (2005): 384-94.

Guth, W. and H. Kliemt. "Competition or Cooperation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes." *Metroeconomica* 45 (1994): 155-87.

Henrich, J., and R. Boyd. "Why People Punish Defectors." *Journal of Theoretical Biology* 208 (2001): 79-89.

Kandori, M., G. Mailath, and R. Rob. "Learning, Mutation, and Long Run Equilibria in Games." *Econometrica* 61 (1993): 29-56.

Kaplow, L., and S. Shavell. "Moral Rules, the Moral Sentiments, and Behavior: Toward a Theory of an Optimal Moral System." *Journal of Political Economy* 115 (2007): 494-514.

Katz, M., and C. Shapiro. "Technology Adoption in the Presence of Network Externalities." *Journal of Political Economy* 94 (1986): 822-41.

Lindbeck, A., S. Nyberg, and J. Weibull. "Social Norms and Economic Incentives in the Welfare State." *Quarterly Journal of Economics* 114 (1999): 1-35.

Rabin, M. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83 (1993): 1281-1302.

Sobel, J. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43 (2005): 393-436.

Sugden, R.. "Normative Expectations: The Simultaneous Evolution of Institutions and Norms." In: Ben-Ner, A. and Putterman, L. (eds.), *Economics, Value, and Organization.* Cambridge: Cambridge University Press, 1998.

Weibull, J. *Evolutionary Game Theory*. Cambridge: MIT Press, 1995.

Figure 1. Payoffs in the Prisoner's Dilemma Game

Player 2 Strategy	Player 1 Strategy H	 D
H	(x_3,x_3)	(x_4,x_1)
D	(x_1,x_4)	(x_2,x_2)

Figure 2. Expected Payoffs with Perfect Signals of Type

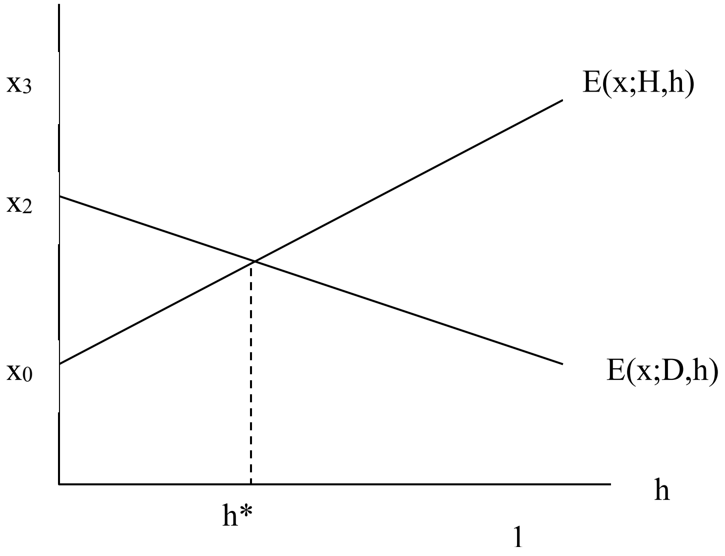


Figure 3. Expected Payoffs with Imperfect Signals

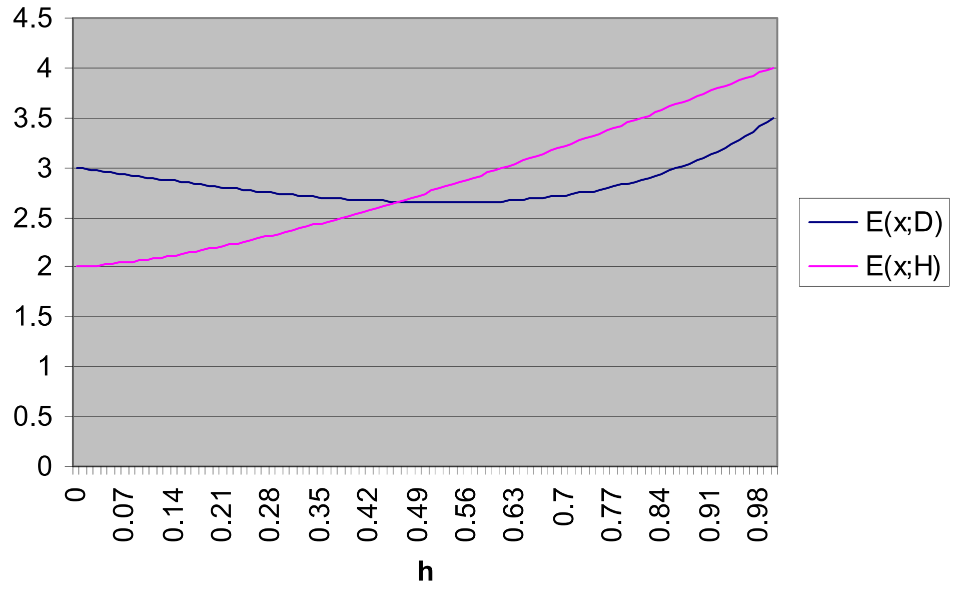
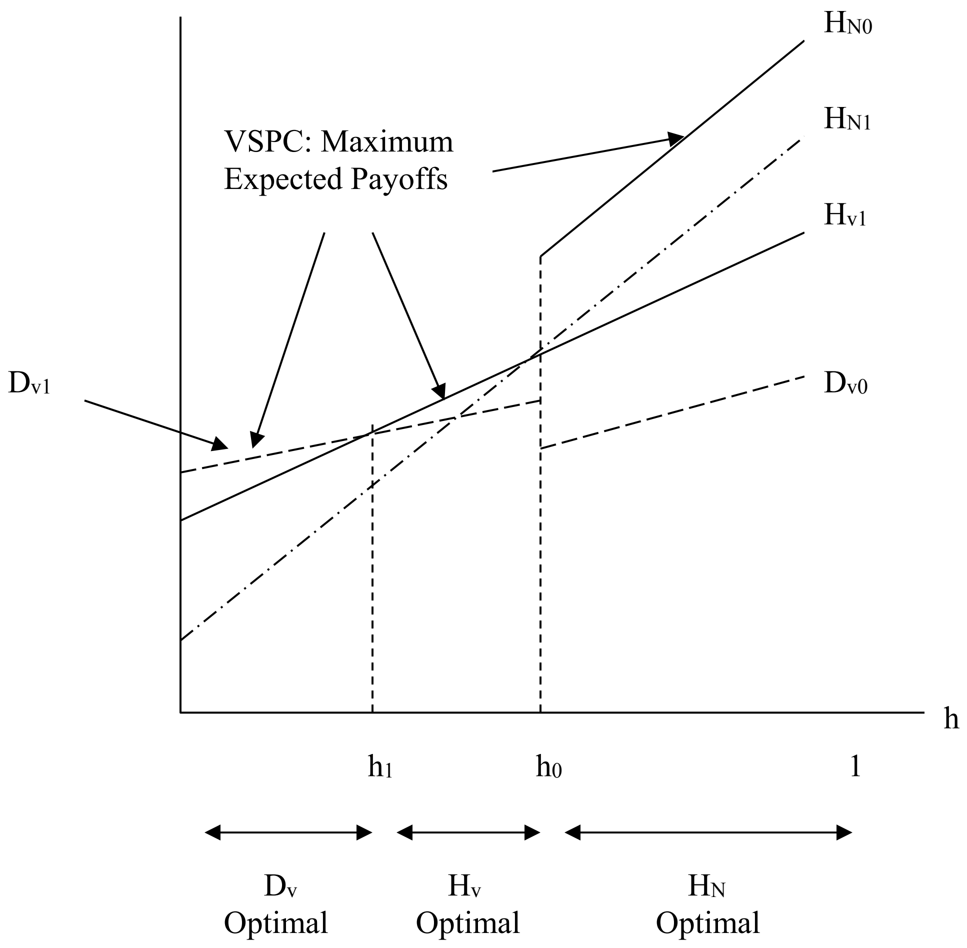


Figure 4. Equilibrium with Vengeance



Notes: H_{N0} denotes $E(x \mid H_N, h_v^*=0, s_v^*=1,h)$, expected H_N payoff when $h_v^*=0$

H_{N1} denotes $E(x \mid H_N, h_v^*=1, s_v^*=1,h)$, expected H_N payoff when $h_v^*=1$

H_{v1} denotes $E(x \mid H_v, h_v^*=1, s_v^*=1,h)$, expected H_v payoff when $h_v^*=1$

D_{v0} denotes $E(x \mid D_v, h_v^*=0, s_v^*=1,h)$, expected D_v payoff when $h_v^*=0$

D_{v1} denotes $E(x \mid D_v, h_v^*=1, s_v^*=1,h)$, expected D_v payoff when $h_v^*=1$

Table 1. Expected Payoffs to Self under Nash Equilibrium Strategies With Conditionally Honest Preferences

	Other		
Self	H_A	H_C	D
H_A	x_3^*	x_3^*	x_0
H_C	x_3^*	$\rho\, x_3^{**} + (1-\rho)\, x_2^{**}$ $\rho \varepsilon \{0,1\}$	x_2^*
D	x_0	x_2^*	x_2

Note:

$x_3^* = (1-u)^2 x_3 + (1-(1-u)^2) x_0$

$x_2^* = (1-u) x_2 + u\, x_0$

$x_3^{**} = (1-u)^2 x_3 + 2u(1-u) x_0 + u^2\, x_2$

$x_2^{**} = (1-u)^2 x_2 + 2u(1-u) x_0 + u^2\, x_2$

Table 2. Nash Equilibrium (NE) Strategies and Fitness Payoffs to Self With Vengeful Preferences

	Other							
	H_V		H_N		D_V		D_N	
Self	NE (Prob)	Payoff	NE (Prob)	Payoff	NE (Prob)	Payoff	NE (Prob)	Payoff
H_V	(H,H) (1)	x_3-ev-ep	(H,H) (1)	x_3-ev	(H,H) (1-u) (H,D) (u)	x_3-ev-ep x_0-v	(H,H) (1-u) (H,D) (u)	x_3-ev x_0-v
H_N	(H,H) (1)	x_3-ep	(H,H) (1)	x_3	(H,D) (1)	x_0	(H,D) (1)	x_0
D_V	(H,H) (1-u) (D,H) (u)	x_3-ev-ep x_0-p	(D,H) (1)	x_0	(H,H) $(1-u)^2$ (D,D) $(1-(1-u)^2)$	x_3-ev-ep x_2-v	(D,D) (1)	x_2-ev
D_N	(H,H) (1-u) (D,H) (u)	x_3-ep x_0-p	(D,H) (1)	x_0	(D,D) (1)	x_2-ep	(D,D) (1)	x_2