Reputational Enforcement of Convenants

Peter Vanderschraaf

February 28, 2005

Technical Report No. CMU-PHIL-167

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

Reputational Enforcement of Covenants

Peter Vanderschraaf

§1. The Classic Reputational Justification of Keeping Promises

A venerable tradition, starting with Plato, maintains that following norms of justice generally serves one's self-interest. In *Leviathan*, Hobbes defends this position against a particularly severe challenge. Suppose a Foole alleges that sometimes he is better off by violating a fundamental norm of justice, namely, that one ought to honor the agreements one makes with others.[1]

> the question is not of promises mutuall, where there is no security of performance on either side; as when there is no Civill Power erected over the parties promising; for such parties are no Covenants: But either where one of the parties has performed already, or where there is a Power to make him performe; there is the question whether it be against reason, that is, against the benefit of the other to performe, or not. (*Leviathan* 15:5)

The Foole contends that if he enters into an agreement with another who honors her commitment first, then he has already received whatever benefits were promised him by the terms of the agreement. In this instance, the Foole claims it would be irrational for him to honor his commitment, for if he did so he would incur a cost to himself with no expectation of any further benefit.

---

[1]Citations of passages in Hobbes' *Leviathan* and Hume's *A Treatise of Human Nature* include chapter or section and paragraph number. In Chapter 15 of *Leviathan*, Hobbes (1651) actually defines justice as the keeping of one's covenants (15:7). In Chapter 14, Hobbes gives a more general definition of justice as fulfilling all of one's obligations, a special case of which are the obligations created by covenants (14:7).

Of course, if the Foole's analysis of the situation is correct, then would the other party to the agreement not anticipate the Foole's response to agreements honored, and act accordingly? Hume raises this very point in *A Treatise of Human Nature*. Hume has us consider the following example:

> Your corn is ripe today; mine will be so tomorrow. 'Tis profitable for us both, that I shou'd labour with you to-day, and that you shou'd aid me to-morrow. I have no kindness for you, and know you have as little for me. I will not, therefore, take any pains on your account; and should I labour with you upon my own account, in expectation of a return, I know I shou'd be disappointed, and that I shou'd in vain depend upon your gratitude. Here then I leave you to labour alone: You treat me in the same manner. The seasons change; and both of us lose our harvests for want of mutual confidence and security. (*Treatise* 3.2.5:8)

The farmer whose corn ripens later reasons that if she were to help the other farmer, then when her corn ripens he would be in the position of Hobbes' Foole, having already benefited from her help. He would no longer have anything to gain from her, so he would not help her, sparing himself the hard labor of a second harvest. Since she cannot expect the other farmer to return her aid when the time comes, she will not help when his corn ripens first, and of course the other farmer does not help her when her corn ripens later.

The problem raised by Hobbes' Foole and Hume's farmers suggests that rational, self-interested agents would never honor their commitments unless forced to by some external authority. Nevertheless, both Hobbes and Hume maintain that self-interested agents can have good reason to comply with agreements after all, even if they are not coerced into doing so. For typically, a single interaction like the commitment problem Hume's farmers face is embedded in a complex sequence of social interactions which occur over time. If an agent reasons that she can expect many opportunities for mutually

beneficial cooperation with others in her community, then by honoring an agreement on one occasion, she indicates to others that she can be counted upon, so that they will be willing to make and keep agreements with her in the future. As Hume puts it,

> I learn to do a service to another, without bearing him any real kindness; because I forsee, that he will return my service, in expectation of another of the same kind, and in order to maintain the same correspondence of good offices with me or with others. And accordingly, after I have serv'd him, and he is in possession of the advantage arising from my action, he is induc'd to perform his part, as foreseeing the consequences of his refusal. (*Treatise* 3.2.5:9)

What are the consequences of refusing to reciprocate? Such refusal constitutes what Kavka terms an unexpected unilateral or *offensive violation* of a covenant (1986, p. 139). Hobbes and Hume give similar warnings against offensive violation. Hume declares that

> When a man says he promises any thing, he in effect expresses a resolution of performing it; and along with that, by making use of this form of words, subjects himself to the penalty of never being trusted again in case of failure. (*Treatise* 3.2.5:10)

In his response to the Foole, Hobbes argues that

> He therefore that breaketh his covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any society, that unite themselves for Peace and Defense, but by the errour of them that receive him; nor when he is received, be retayned in it, without seeing the danger of their errour; which errours a man cannot reasonably reckon upon as the means of his security: and therefore if he be left, or cast out of

> Society, he perisheth; and if he live in Society, it is by the errours of other men, . . . (*Leviathan* 15: 5)

According to Hobbes and Hume, one who follows the Foole's advice and offensively violates an agreement might enjoy an immediate gain by taking advantage of others' compliance. However, rational and well-informed agents will never enter into covenants with one who has ever once exploited others by offensively violating a covenant. Losing all such opportunities for future benefits of cooperation far outweighs any immediate gain from refusing to reciprocate when others have honored their parts of a covenant, think Hobbes and Hume.

This kind of reply to the Foole is exceptionally controversial. Among the many puzzles this reputational argument for keeping one's promises raises, some have rightly notes that if Hobbes' and Hume's arguments are entirely successful than people do not need a government to enforce compliance with covenants (Hampton 1986, Kavka 1986), contrary to what Hobbes and Hume both conclude. Hobbes and Hume apparently assume that the agents an individual like the Foole will encounter keep their agreements with all but offensive violators. But it is by no means obvious that agents are *able* to follow such a conditionally cooperative policy. Hume and Hobbes have little to say regarding the conditions necessary for a convention of conditional cooperation to regulate a community of individuals. This paper explores conditions under which reputation alone can enforce covenants. §2 introduces a formal model of interaction in a community. The members of a community engage in a *Covenant Game* that is repeated over time. §3 presents several folk theorems that establish conditions under which performing in covenants with those who follow Hume's and Hobbes' advice constitutes an equilibrium of the repeated Covenant Game. These folk theorems establish that in certain settings Hobbes' and Hume's arguments against offensively violating covenants are indeed decisive. However, these settings presuppose that the community has certain mechanisms that generate

common knowledge at their disposal. In communities that lack these structures, reputation alone may not give a would-be foole good reason to honor covenants. §4 presents a simple computational model where community members who must rely upon private communication alone cannot effectively deter fooles from offensively violating covenants. The concluding §5 considers some of the lessons to be drawn from the analysis of the repeated Covenant Game.

§2. The Indefinitely Repeated Covenant Game Played by Community Members

We first give a formal description of how individual members in a community interact in pairwise covenant situations. $N = \{1, ..., n\}$ is a set or community of *players* where $n \geq 2m$, $m \in \mathbb{N}$ and $m \geq 2$. Ω denotes a set of *possible worlds*. At each time period or *stage* t, one world $\omega(t) \in \Omega$ obtains at t. A description of each possible world at t includes all of the information relevant to the agents' decisions and acts at stage t, including a description of the stage game, the assignment, and the beliefs each player has regarding the counterparts, as in Aumann (1987) and Dekel and Gul (1997). Each Player $i \in N$ has a subjective probability distribution $\mu_i(\cdot)$ over the propositions in Ω, a private information partition $\mathcal{H}_i$ of Ω, and an expectation operator $E_i(\cdot)$ based upon $\mu_i(\cdot)$. At each stage t, a set $N_t \subseteq N$ such that $\text{card}(N_t) = m_t$ is divisible by 2 is selected. Each Player $i \in N_t$ is matched with a *counterpart* $i(t) \in N_t - \{i\}$ according to a bijective random vector $X_t : N_t \rightarrow N_t$ with no fixed points. The sequence (X_t) is the *matching protocol*. If $N_t = N$ at each stage, that is, every player is matched at every stage, then the players are in an *ordinary repeated matching game*. Note that in this case we must have $n = 2m$ and $m_t = n$ at every stage t. If Player i is unmatched at some stage t, then at this stage Player i receives a constant *noninteraction payoff* $\underline{u} = 0$. If Player i is matched at period t, then Player i and his counterpart Player $i(t)$ play the *Covenant Game* summarized in Figure 1.

Figure 1. Covenant Game

		Player $i(t)$		
		P	D	B
Player i	P	$(1,1)$	$(-l, 1+g)$	$(0,0)$
	D	$(1+g, -l)$	$(-c, -c)$	$(0,0)$
	B	$(0,0)$	$(0,0)$	$(0,0)$

$P =$ perform, $D =$ double-cross, $B =$ boycott

$$g > 0, l > c \geq 0, g - l < 1$$

In this game, parties can enter into a covenant by exchanging promises. Either party can *boycott* (B) by refusing to enter into a covenant. If the matched parties do enter into a covenant, then each can either *perform* (P) or *double-cross* (D). If either or both boycott, then each receives the payoff 0 of working alone, which is strictly worse than her payoff if both perform. The subgame that results if the players exchange promises is given in Figure 2.

Figure 2. Prisoner's Dilemma Subgame

		Role 2: P	Role 2: D
Role 1	P	$(1, 1)$	$(-l, 1+g)$
	D	$(1+g, -l)$	$(-c, -c)$

The Figure 2 game is a Prisoners' Dilemma. Consequently, the Figure 1 game is sometimes called *optional Prisoners' Dilemma* or *Prisoners' Dilemma with opting out* (Kitcher 1993, Batali and Kitcher 1995). We write $a_i(t) \in \{P, D, B\}$ to denote the pure strategy a Player i matched at stage t selects in the Covenant Game. We will assume that if a Player i is unmatched at any stage, then at this stage he receives the payoff $\underline{u} = 0$, same as the payoff when he is matched but his counterpart boycotts him. Figure 3 depicts the convex hull of the payoffs of the Covenant Game for the case where $g = l = 1$ and $c = \frac{1}{2}$.

[See Figure 3.]

If $c > 0$, (B, B) is the unique Nash equilibrium of the Covenant Game. To see why, suppose that it is mutual knowledge throughout the community N that each Player $i \in N$ paired in the Covenant Game is *Bayesian rational*, that is, he acts so as to maximize expected payoff, and each player knows the structure of the Covenant Game. For each Player $i \in N$, let

$$x_{i1} = \mu_i\big[a_{i(t)}(t) = P\big],$$
$$x_{i2} = \mu_i\big[a_{i(t)}(t) = D\big], \text{ and}$$
$$x_{i3} = \mu_i\big[a_{i(t)}(t) = B\big] = 1 - x_{i1} - x_{i2}\ .$$

D weakly dominates P, so any matched Player i rules out opting for P and also rules out the possibility that counterpart Player $i(t)$ chooses P, that is, $x_{i1} = 0$. In the remaining subgame, B can fail to be Player i's Bayesian rational strategy only if $-cx_{i2} > 0$, which is impossible since $c > 0$. By a similar argument, if $c = 0$ then all Nash equilibria of the Covenant Game are characterized by Player i and Player $i(t)$ both following a mixed strategy[2] over $\{D, B\}$.

Now we define formally the strategies that the players in N can follow in the indefinitely repeated Covenant Game. A generic *strategy* for Player i is a sequence of functions $f_i = (f_i^t)$ where $f_i^t : \Omega \to \{P, D, B\}$ and f_i^t is $\mathcal{H}_i$-measurable.
$\boldsymbol{f} = (f_1, ..., f_n)$ is a generic *strategy profile*. S_i denotes the set of all strategies Player i can follow, and $S = S_1 \times \cdots \times S_n$. At a given stage t, $f_i^t(\omega(t)) \in \{P, D, B\}$ defines the pure strategy $a_i(t)$ that Player i follows in Γ at stage t. We stipulate that $f_i^t(\omega(t)) = B$ if $i \notin N_t$ in order to avoid trivial complications.

$$\boldsymbol{f}^t(\omega(t)) = \big(f_1^t(\omega(t)), \ldots, f_n^t(\omega(t))\big) \in \{P, D, B\}^n$$

is the set of pure strategies $(a_1(t), \ldots, a_n(t))$ the players follow at t. Player i's expected payoff at stage t given $\omega(t) \in \Omega$ is

$$E_i\big(u_i\big(\boldsymbol{f}^t(\omega(t))\big)\big) = \sum_{j \neq i} E_i\big(u_i\big(f_i^t(\omega(t)), f_j^t(\omega(t))\big)\big)\mu[i(t) = j]\ .$$

Let $p_i \in (0, 1)$ be Player i's *discount factor*. Player i's overall expected payoff is

---

[2]That is, Player i and Player $i(t)$ select either D or B according to the oputcomes of independent random experiments.

$$E_i(u_i \circ \boldsymbol{f}) = \sum_{t=1}^{\infty} E_i\big(u_i\big(\boldsymbol{f}^t(\omega(t))\big)\big)p_i^t\,.$$

A strategy profile f is a *correlated equilibrium* of the indefinitely repeated Covenant game if, and only if, for each $i \in N$,

$$E_i(u_i \circ \boldsymbol{f}) \geq E_i(u_i \circ (f_i', \boldsymbol{f}_{-i})) \text{ for all } f_i' \in S_i\,.^3$$

In the sequel we will examine the prospects for reciprocal cooperation among a community of Bayesian rational players who engage in the Covenant Game when they meet. Why base the analysis of this paper on the repeated Covenant Game rather than the repeated Prisoners' Dilemma? The Covenant Game reflects the arguments for keeping promises in the classic works of Hume and Hobbes better than the Prisoners' Dilemma. Hobbes and Hume argue that offensive violators will be *shunned* by others, not that others will then try to exploit them. Shunning is formalized by introducing the boycott strategy of the Covenant Game, which allows players to ignore those they encounter. However, there is a more substantive reason for using the Covenant Game rather than the Prisoners' Dilemma. In the repeated Prisoners' Dilemma, equilibria of conditional cooperation require an exploited player to punish the offensive violator by double-crossing, at least for a certain number of stages after the violation. In a repeated Prisoners' Dilemma between a fixed pair of players, each has no trouble knowing when to punish because the counterpart is fixed. Matters are far more complicated for the members of a community of players who are matched with different counterparts over time. In such a community, if the stage game is the Prisoners' Dilemma, then members might have trouble identifying offensive violations of a covenant. If, for instance, one

---

[3]The subscript '$-i$' is the "jacknife" notation that indicates the results of removing the ith component of an ordered n-tuple or n-fold Cartesian product. Here,

$$(f_i', \boldsymbol{f}_{-i}) = (f_1, ..., f_{i-1}, f_i', f_{i+1}, ..., f_n)\,.$$

player observes another double-crossing in a given stage of repeated Prisoners' Dilemma, she might have trouble determining whether the double-crosser is violating a covenant offensive, or is merely punishing an offensive violator. In the Humean-type strategies for playing the repeated Covenant Game, this problem is avoided to a large extent because punishment always takes the form of a boycott. In the repeated Covenant Game, double-crossing is always an offensive violation.

§3. Folk Theorems

We will prove several basic folk theorems for the indefinitely repeated Covenant Game. These folk theorems establish that a strategy of conditional performance can be a Bayesian rational strategy in repeated covenant situations, and that conditional performance yield a greater average payoff than always double-crossing when one is matched. This gives a partial vindication of the reputational argument for keeping promises we find in Hobbes and Hume. However, we shall see that this is a partial vindication only. The results in this section are similar in spirit to the folk theorems proved for repeated Prisoners' Dilemma played in a random matching model in Kandori (1992) and Ellison (1994). However, the matching model developed here does not assume that every player is matched with a counterpart in every period. The stochastic strategies considered below also allow for the possibility that an offensive violation starts no punishment cycle. Kandori and Ellison assume that every player is matched at every period and that an offensive violation is certain to start a punishment cycle. Note that in Propositions 2-6, the only restriction placed on the matching protocol is that the probability a Player $i \in N$ is matched remains constant over the stages of play, and in Proposition 7 each player is matched at each stage but otherwise there is no restriction on the matching protocol. Finally, readers should be aware that the analysis here differs fundamentally from the evolutionary analyses of strategies such as "tit for tat" for playing repeated Prisoners' Dilemma developed by authors such as Axelrod (1981, 1984) and

Linster (1992). Such evolutionary analyses focus on how strategies can emerge in repeated Prisoners' Dilemma games played over time between fixed pairs of agents. Here, the members of a community play the Covenant Game when they are matched, and might at any stage change their partners.

The first result of this section establishes the set of average payoffs in the indefinitely repeated Covenant game that can be sustained in an equilibrium.

Proposition 1. Let $N_t = N$ for each period t, let the matching protocol be such that half of the players lie in the same set N_R at each stage and their counterparts lie in the set $N_C = N - N_R$, and let $f = (f^t)$ define a sequence of correlated strategies over $\{P, D, B\}^2$ as follows: $\Omega = \{\omega_1, \omega_2\}$ where $x = \mu_i[\omega = \omega_1]$ and $1 - x = \mu_i[\omega = \omega_2]$ for all $i \in N$. $s : \Omega \to \{P, D\}^2$ is a map that yields Player $i \in N_R$ an expected payoff of $u_R > 0$ and counterpart Player $i(j) \in N_C$ an expected payoff of $u_C > 0$. Then at each stage t,

$$f^t(\omega) = \begin{cases} s(\omega) & \text{if all players have followed } s(\omega) \text{ at every stage } T < t \\ B & \text{otherwise} \end{cases}.$$

If

$$p_i \geq \frac{1 + g - u_R}{1 + g} \text{ for each } i \in N_R$$
$$p_i \geq \frac{1 + g - u_C}{1 + g} \text{ for each } i \in N_C$$

then $\boldsymbol{f}$ is a correlated equilibrium of the indefinitely repeated Covenant Game with this matching protocol.

PROOF. Given $i \in N_R$, we have

$$E_i(u \circ \boldsymbol{f}) = \sum_{t=1}^{\infty} E_i(u_i(s(\omega)))p_i^t \tag{1}$$
$$= \sum_{t=1}^{\infty} u_R \cdot p_i^t \,.$$

Now consider any strategy f_i' where Player i deviates from $\boldsymbol{f}$. Let T_0 be the first stage such that $f_i'(\omega(T_0)) \neq f(\omega(T_0))$. Then

$$E_i(u_i(f_i', \boldsymbol{f}_{-i})) \leq \left(\sum_{t=1}^{T_0-1} u_R \cdot p_i^t\right) + (1+g) \cdot p_i^{T_0} \tag{2}$$

because (*i*) at stage T_0 Player i gains at most the discounted gain $(1+g)p_i^{T_0}$ of exploiting Player $i(T_0)$, and (*ii*) at each subsequent stage $t > T_0$, Player i receives 0 because now all players in N_C always boycott. By (1) and (2), $E_i(u_i \circ \boldsymbol{f}) \geq E_i(u_i(f_i', \boldsymbol{f}_{-i}))$ when

$$\frac{u_R p_i^{T_0}}{1-p_i} = \sum_{t=T_0}^{\infty} u_R p_i^t \geq (1+g)p_i^{T_0}$$

or

$$\frac{u_R}{1-p_i} \geq 1+g \tag{6}$$

and (6) is satisfied when

$$p_i \geq \frac{1+g-u_R}{1+g}.$$

The argument for $i \in N_C$ is similar. □

Proposition 1 shows us that any point in the convex set of points within the convex hull of the Covenant Game such that each component is nonnegative can be sustained in an equilibrium of the corresponding ordinary indefinitely repeated game. Figure 4 gives the graph of this set for the special case where $g = l = 1$ and $c = \frac{1}{2}$.

[See Figure 4.]

The equilibrium of Proposition 1 is similar to one of the "contagion" equilibria presented in Kandori (1992). The basic underlying idea is as follows: The community members are partitioned into two distinct classes, and at each stage every member of the community is matched with a member of the class other than her own. So long as everyone has always

conformed to the strategy s, each player continues to follow her required end of s and all members of the one class continue to receive the discounted payoff u_R and all members of the other class continue to receive the discounted payoff u_C. But if anyone deviates from s at any stage, then subsequently everyone always boycotts. Clearly, this equilibrium requires a rather contrived matching protocol and requires all to punish each other for the offense of just a single member of the community. Proposition 1 show us what kind of equilibria are possible in the indefinitely repeated Covenant Game. But what we want to consider are equilibria that do not depend upon unusual matching protocols and that punish only offensive violators.

The remaining results in this section characterize some of the most important reputational equilibria that are possible in the repeated Covenant Game with matching. First, let us define a basic *Humean strategy* for playing this indefinitely repeated game: For $i \in N$, define $z_i : t \rightarrow \{0, 1\}$ by

$$z_i(t) = \begin{cases} 1 & \text{if } I(i,t) \text{ obtains} \\ 0 & \text{otherwise} \end{cases}$$

where $I(i, 1)$ obtains for each $i \in N$ and for $t > 1$,

$$I(i,t) = \forall(T \leq t-1)\forall(j \in N)[(j = i(T) \wedge I(j,T) \rightarrow a_i(T) = P) \\ \wedge (j = i(T) \wedge \neg I(j,T) \rightarrow a_i(T) = B)] \ .$$

The definition of $I(i, 1)$ formalizes the idea that a player is *innocent* at stage t if, and only if, over the first $t-1$ stages she has never failed to perform with an innocent counterpart and has always boycotted counterparts who are *guilty*, that is, not innocent. So $z_i(t) = 1$ exactly when Player i is innocent at stage t. We can think of $z_i(t)$ as Player i's guilt-or-innocence "marker". Note that here Player i must enter into and perform in a covenant when she is matched with an innocent counterpart.[4] Player i is also not allowed to enter

[4]It is possible to define a variant of $I(i, t)$ where Player i remains innocent if he boycotts innocent parties, and to derive results similar to those we discuss here.

into a covenant with a guilty counterpart. If Player i does form a covenant with a guilty counterpart, then Player i becomes guilty whether he performs or double-crosses. Intuitively, the community punishes fellow members who succumb to the temptation to try to profit by dealing with the guilty, no matter how these deals may turn out. Let $\omega(t) = (z_1(t), ..., z_n(t))$. Then Player i's Humean strategy is defined by $h = (h(\omega(t)))$ where

$$h(\omega(t)) = \begin{cases} P & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 1 \\ B & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 0 \end{cases}.$$

That is, Player i performs when he is matched in the Covenant Game if he is matched with an innocent counterpart and boycotts if he is matched with a guilty counterpart. We can prove a folk theorem that gives conditions under which the Humean strategy characterizes an equilibrium of the repeated Covenant Game.

Proposition 2. Let the probability that a given Player $i \in N$ is matched be constant over stages. If x_i denotes the probability that Player i is matched at a given stage and

$$p_i \geq \frac{1+g-x_i}{1+g}, \; i \in N$$

then $\boldsymbol{h} = (h, ..., h)$ is an equilibrium of the indefinitely repeated Covenant Game with matching over N.

PROOF. Let $x_i = \sum_{j \neq i} \mu_i[i(t) = j] = \mu_i[i(t) \in N_t]$ and $1 - x_i = \mu_i[i(t) \notin N_t]$, that is, x_i is Player i's probability of being matched. Note that at stage t, Player i's undiscounted expected payoff for following his end of $\boldsymbol{h}(t) = (h(t), ..., h(t))$ is

$$\begin{aligned} E_i(u_i(\boldsymbol{h}(t))) &= 0 \cdot \mu_i[i(t) \notin N] + \sum_{j \neq i} u_i(P, P) \cdot \mu_i[i(t) = j] \\ &= 0 \cdot (1 - x_i) + u_i(P, P) \cdot x_i \\ &= x_i \end{aligned}$$

because by hypothesis all the players in N follow their respective ends of $\boldsymbol{h}$.

So we have

$$E_i(u_i \circ \boldsymbol{h}) = \sum_{t=1}^{\infty} x_i p_i^t = \frac{x_i p_i}{1 - p_i}.$$

Note that if $x_i = 0$, that is, Player i is never matched, then $E_i(u_i \circ \boldsymbol{h}) = 0$ and Player i can never gain by deviating from $\boldsymbol{h}$ because he never interacts. The interesting case is where $x_i > 0$. Now consider any strategy f_i' where Player i deviates from the sequence $(h(\omega(t)))$. Let T_0 be the first stage such that $f_i'(\omega(T_0)) \neq h(\omega(T_0))$. Then we have two cases to consider:

Case *i*. If $f_i'(\omega(T_0)) = B$, then

$$\begin{aligned} E_i(u_i(f_i', \boldsymbol{h}_{-i})) &= \sum_{t=1}^{T_0-1} x_i p_i^t + 0 \cdot p_i^{T_0} + \sum_{t=T_0+1}^{\infty} 0 p_i^t \\ &= \sum_{t=1}^{T_0-1} x_i p_i^t \\ &= x_i \cdot \frac{p_i - p_i^{T_0}}{1 - p_i} \end{aligned}$$

because Player i follows B against Player $i(T_0)$ at stage $t = T_0$, so at this stage he will net the discounted payoff $0 \cdot p_i^{T_0}$ of boycotting Player $i(T_0)$, and in each subsequent stage $t > T_0$, Player i will gain the $0 \cdot p_i^t$ of acting alone, because for each $t > T_0$ if Player i is matched then Player i's counterpart $i(t)$ follows B. Note that

$$\sum_{t=1}^{\infty} x_i p_i^t > \sum_{t=1}^{T_0-1} x_i p_i^t$$

so in this case, $E_i(u_i \circ \boldsymbol{h}) > E_i(u_i(f_i', \boldsymbol{h}_{-i}))$ for *any* positive value of p_i.

Case *ii*. If $f_i'(\omega(T_0)) = D$, then

$$\begin{aligned} E_i(u_i(f_i', \boldsymbol{h}_{-i})) &= \sum_{t=1}^{T_0-1} x_i p_i^t + (1+g) \cdot p_i^{T_0} + \sum_{t=T_0+1}^{\infty} 0 p_i^t \\ &= \sum_{t=1}^{T_0-1} x_i p_i^t + (1+g) \cdot p_i^{T_0} \end{aligned}$$

because Player i follows D against Player $i(T_0)$ at stage $t = T_0$, then at this stage he will net the discounted gain $(1+g)p_i^{T_0}$ of exploiting Player $i(T_0)$, and in each subsequent stage, Player i will gain the discounted payoff $0 \cdot p_i^t$ of acting alone, as in Case i. In this case, we have

$$E_i(u_i \circ \boldsymbol{h}) \geq E_i(u_i(f_i', \boldsymbol{h}_{-i}))$$

when

$$\sum_{t=T_0}^{\infty} x_i p_i^t \geq (1+g) \cdot p_i^{T_0}$$

that is,

$$\frac{x_i p_i^{T_0}}{1-p_i} \geq (1+g) \cdot p_i^{T_0}$$

or

$$\frac{x_i}{1-p_i} \geq 1+g \tag{1}$$

and (1) is satisfied when

$$p_i \geq \frac{1+g-x_i}{1+g} \,. \quad \square$$

A first variant on the basic Humean strategy has players punish an offensive violation sometime after the violation. In order to follow their parts of the equilibrium $\boldsymbol{h}$ of the basic Humean strategy, at every period $t > T_0$, each player must know the identity of any Player i who has deviated unilaterally from $\boldsymbol{h}$ at some period $t < T_0$. Suppose that when a Player i double-crosses an innocent counterpart or covenants with a guilty counterpart, there is a *lag* before all the other players learn that Player i is guilty. There is still an equilibrium of conditional cooperation, but it depends upon a "delayed reaction" on the part of the rest of the community when someone double-crosses. Let us define a k-*delayed Humean strategy* $h[k]$ as follows: Let $k \in \mathbb{N}$, and for $i \in N$, let $z_i : t \to \{0, 1\}$

be defined as

$$z_i(t) = \begin{cases} 1 & \text{if } I(i, t-k) \text{ obtains} \\ 0 & \text{otherwise} \end{cases}$$

where $I(i,t)$ is defined as before, and we stipulate that $I(i, t-k) = I(i,0) = 1$ if $t - k \leq 1$. Then a player's k-delayed Humean strategy is defined by $h[k] = (h[k](\omega(t)))$ where

$$h[k](\omega(t)) = \begin{cases} P & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 1 \\ B & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 0 \end{cases}.$$

Here each player performs with any counterpart who from the 1st to the $t-k$th stages never failed to perform with the innocent and always boycotted the guilty . But if a counterpart became guilty k or more stages ago, then a $h[k]$-follower boycotts. Why would a community of players follow a k-delayed Humean strategy that requires each of them to wait for k periods before boycotting, even if she knows her counterpart is guilty? They have good reason to wait, if it takes at least k periods for Player i's guilt to become common knowledge[5] among $N - \{i\}$. Suppose Player j knows that Player i is guilty and $j(t) = i$, but Player j believes that some of the others in $N - \{i, j\}$ do not yet know that Player i is guilty. Then if Player j boycotts Player i at stage t, she knows that some of the others might mistakenly infer that she is boycotting an innocent counterpart and infer from this that *she* is guilty. But if Player i's guilt becomes common knowledge among $N - \{i\}$ k stages after the offense, then each player in $N - \{i\}$ knows he can punish Player i from now on without anyone else thinking he is deviating from $h[k]$ rather than punishing. One can think of this situation as one where it takes k stages from the time an

---

[5]A proposition $A \subset \Omega$ is common knowledge among the players of $M \subset N$ if each Player $i \in M$ knows A, each Player $i \in M$ knows that each Player $j \in M$ knows A, each Player $i \in M$ knows that each Player $j \in M$ knows that each Player $k \in M$ knows A, and so on (Lewis 1969, Aumann 1976).

offense occurs for the identity of the offender to be "broadcast" to the entire community. Proposition 3 is another folk theorem that gives equilibrium conditions for a community of a $h[k]$-followers.

Proposition 3. Let the probability that a given Player $i \in N$ is matched be constant over stages. If $x_i > 0$ denotes the probability that Player i is matched at a given stage and

$$(1) \qquad p_i^{k+1} + p_i - \left(1 + x_i - \frac{x_i}{1+g}\right) \geq 0, \; i \in N$$

then $\boldsymbol{h}[k] = (h[k], ..., h[k])$ is an equilibrium of the indefinitely repeated Covenant Game with matching over N.

Note that for $k \geq 2$, the polynomial equation $p_i^{k+1} + p_i - \left(1 + x_i - \frac{x_i}{1+g}\right) = 0$ will have a root in $(0, 1)$, so in general there will be discount factors that do satisfy (1). However, as $k \to \infty$,

$$p_i^{k+1} + p_i - \left(1 + x_i - \frac{x_i}{1+g}\right) \to p_i - \left(1 + x_i - \frac{x_i}{1+g}\right) < 0$$

because $1 < 1 + x_i - \frac{x_i}{1+g}$, so (1) cannot be satisfied in the limit. This is not surprising, if we keep in mind that k is the number of stages where the players in $N - \{i\}$ are "biding their time" before they start to punish a Player i who deviates from $\boldsymbol{h}[k]$. The longer they wait before they start the punishment cycle, the higher each player's discount factor must be to make $h[k]$ a best response to $\boldsymbol{h}[k]_{-k}$. And if $k \to \infty$, the players in $N - \{i\}$ never start to punish a deviator, so conforming to $h[k]$ will never be a Player's best strategy in this limiting case.

PROOF. As in Proposition 2, let $x_i = \sum_{j \neq i} \mu_i[i(t) = j]$ denote Player i's probability of being matched, and note that at stage t, Player i's undiscounted expected payoff for following his end of $h[k] = (h[k](t), ..., h[k](t))$ is

$$E_i(u_i(\boldsymbol{h}[k](t))) = x_i \, .$$

We have

$$E_i(u_i(\boldsymbol{h}[k])) = \sum_{t=1}^{\infty} x_i p_i^t$$

Suppose Player i deviates from $h[k]$ by following some strategy $f_i \neq h[k]$. If T_0 is the first period where $f_i(T_0) \neq h[k](T_0)$, then

$$E_i(u_i(f_i, \boldsymbol{h}[k]_{-i})) \leq \sum_{t=1}^{T_0-1} x_i p_i^t + (1+g)p_i^{T_0} + \sum_{t=T_0+1}^{T_0+k} x_i(1+g)p_i^t + \sum_{t=T_0+k+1}^{\infty} 0 \cdot p_i^t \tag{2}$$

because the right member of (2) is the expected payoff Player i receives if he exploits his counterpart $i(T_0)$ at period T_0 and then successfully exploits every counterpart he meets from the T_0 to the $T_0 + k$th period before the punishment begins at the $T_0 + k + 1$st period. So $E_i(u_i(\boldsymbol{h}[k])) \geq E_i(u_i(f_i, \boldsymbol{h}[k]_{-i}))$ if

$$\sum_{t=T_0}^{\infty} x_i p_i^t \geq (1+g)p_i^{T_0} + \sum_{t=T_0+1}^{T_0+k} x_i(1+g)p_i^t \,. \tag{3}$$

Simplifying (3) we get

$$\frac{x_i p_i^{T_0}}{1-p_i} \geq (1+g)p_i^{T_0} + x_i(1+g) \cdot \frac{p_i^{T_0} - p_i^{T_0+k+1}}{1-p_i}$$

or

$$\frac{x_i}{1-p_i} \geq (1+g)\left(1 + \frac{x_i - x_i p_i^{k+1}}{1-p_i}\right) \tag{4}$$

and (4) is satisfied when $p_i + p_i^{k+1} \geq 1 + x_i - \frac{x_i}{1+g}$. □

Another variant on the basic Humean strategy is to allow for forgiveness after a period of punishment. For $i \in N$, define $z_i : t \to \{0, 1\}$ by

$$z_i(t) = \begin{cases} 1 & \text{if } I[K](i,t) \text{ obtains} \\ 0 & \text{otherwise} \end{cases}$$

where $I[K](i, 1)$ obtains for each $i \in N$ and for $t > 1$,

$$I[K](i,t) = \neg\exists(T : t-K+1 \leq T \leq t-1 \wedge j \in N)[(j = i(T)) \\ \wedge I[K](j,T) \wedge a_i(T) = D) \\ \vee (j = i(T) \wedge \neg I[K](j,T) \wedge a_i(T) \neq B)] \ .$$

In words, a Player i is now innocent if, over the K most recent periods in the past, i has not exploited an innocent counterpart or entered into a covenant with a guilty counterpart. Again let $\omega(t) = (z_1(t), ..., z_n(t))$. Then Player i's *K-step Humean strategy* is defined by $h_K = (h_K(\omega(t)))$ where

$$h_K(\omega(t)) = \begin{cases} P & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 1 \\ B & \text{if } i \in N_t \text{ and } z_{i(t)}(t) = 0 \end{cases} .$$

As with the basic Humean strategy h, here Player i performs when he is matched in the Covenant Game if the counterpart is innocent and boycotts if the counterpart is guilty. The difference is that now a party who unilaterally deviates from the K-step Humean strategy is guilty for only K stages after the violation. Over the punishment period of K stages after a unilateral violation, violator Player i's counterparts boycott when i is matched. At the end of the punishment period, Player i in effect regains his innocence and is treated accordingly. We now show that this more "forgiving" Humean strategy can characterize an equilibrium of the indefinitely repeated Covenant Game.

Proposition 4. Let the probability that a given Player $i \in N$ is matched be constant over stages. If x_i denotes the probability that Player i is matched at a given stage and

$$x_i p_i^{K+1} + 1 \leq (1+g)p_i, \ i \in N, \tag{1}$$

then $\boldsymbol{h}_K = (h_K, ..., h_K)$ is a correlated equilibrium of the indefinitely repeated Covenant Game with matching over N.

For $K \geq 2$, the polynomial equation $x_i p_i^{K+1} - (1+g)p_i + 1 = 0$ will have a root in $(0,1)$, so in general there will be discount factors that do satisfy (1).

PROOF. Note that as in Proposition 2, Player i's undiscounted expected payoff for following his end of $h_K = (h_K(t), ..., h_K(t))$ is

$$E_i(u_i(\boldsymbol{h}_K(t))) = x_i \, .$$

Consider any strategy $f_i^{\star}$ where Player i deviates unilaterally from the sequence $(\boldsymbol{h}_K(\omega(t)))_{t=1}^{\infty}$. Let T_l, $l \in \mathbb{N}$ be any stage such that $f_i^{\star}(\omega(T_l)) \neq h_K(\omega(T_l))$ when the other players follow $\boldsymbol{h}_{K_{-i}}(\omega(T_l))$. Then Player i's counterparts follow B at each stage t where $i \in N_t$ where $T_l + 1 \leq t \leq T_l + K$, and then revert back to P at stage $t = T_l + K + 1$ if $i \in N_t$, so that Player i now faces the sequence

$$(\boldsymbol{h}_{K_{-i}}(\omega(t)))_{t=T_l+K+1}^{\infty} = (\boldsymbol{h}_{K_{-i}}(\omega(t)))_{t=1}^{\infty}$$

that is, at stage $t = T_l + K + 1$ Player i faces the same situation he faced at the beginning stage $t = 1$, except that Player i's overall expected payoff is multiplied by the discount factor $p_i^{T_l+K+1}$. Now note that for any T_l, over the K stages from $t = T_l + 1$ to $t = T_l + K$ Player i's expected payoff if he follows $f_i^{\star}$ is no greater than

$$(1 + g) \cdot p_i^{T_l}$$

because (*i*) if Player i deviates from h_K at $t = T_l$, then at stage T_l he will net at most the discounted gain of exploiting Player 2, and (*ii*) from $t = T_l + 1$ to $t = T_l + K$ the most Player i can gain in each subsequent is 0, because over these K stages Player 2 follows H and so H is Player i's unique best response to Player 2. If, on the other hand, Player i follows h_K from $t = T_l + 1$ to $t = T_l + K$, then over these stages Player i's expected payoff equals

$$\sum_{t=T_l}^{T_l+K} x_i \cdot p_i^t = x_i \cdot \frac{p_i^{T_l} - p_i^{T_l+K+1}}{1 - p_i} \, .$$

So in order for $f_i^{\star}$ to be Player i's best response to h_K, we must have for each T_l

$$(1 + g) \cdot p_i^{T_l} > x_i \cdot \frac{p_i^{T_l} - p_i^{T_l+K+1}}{1 - p_i}$$

or

$$1 + g > x_i \cdot \frac{1 - p_i^{K+1}}{1 - p_i}$$

Hence $E_1(u \circ \boldsymbol{h}_K) \geq E_1\big(u_1\big(f_i^{\bigstar}, \boldsymbol{h}_{K_{-i}}\big)\big)$ when $1 + g \leq x_i \cdot \frac{1-p_i^{K+1}}{1-p_i}$ or

$$x_i p_i^{K+1} - (1+g)p_i + 1 \leq 0\,. \ \square$$

So far, we have assumed that if a given player is guilty, the rest of the community is certain to punish him in time. But suppose that an offensive violation is discovered by the rest of the community only with probability $q < 1$. Then a *stochastic* Humean strategy $h[q]$ where the players in N punish a guilty player with probability can still characterize an equilibrium of the repeated Covenant Game. This time, for $i \in N$, define $w_i : t \rightarrow \{0, 1\}$ by

$$w_i(t) = \begin{cases} 0 & \text{if } 1_{A_i(T)} = 1 \text{ for some } T \leq t \\ 1 & \text{otherwise} \end{cases}$$

where $A_i(T)$ implies that

$$Q(i,T) = \exists(j \in N)[j = i(T) \wedge \\ ((I(j,T) \wedge a_i(T) \neq P) \vee (\neg I(j,T) \wedge a_i(T) \neq B))]$$

and assume that $E_i\big(1_{A_i(t)}|Q(j,T)\big) = q$ and $E_i\big(1_{A_i(t)}|\neg Q(j,T)\big) = 0$ for each $i \in N$. $Q(i,T)$ obtains either if at period T Player i offensively violates a covenant or enters into a covenant with a guilty counterpart. One can think of the event $A_i(t)$ as Player i's offense against the community being "found out" by everyone. If $Q(i,T)$ does obtain, the offense is "found out" with probability q. Let $\omega(t) = (w_1(t), ..., w_n(t))$. Then Player i's *Humean stochastic q-strategy* is defined by $h[q] = (h[q](\omega(t)))$ where

$$h[q](\omega(t)) = \begin{cases} P & \text{if } i \in N_t \text{ and } w_{i(t)}(t) = 1 \\ B & \text{if } i \in N_t \text{ and } w_{i(t)}(t) = 0 \end{cases}.$$

Proposition 5. Let the probability that a given Player $i \in N$ is matched be constant over stages. If x_i denotes the probability that Player i is matched at a given stage and

$$p_i \geq \frac{1+g-x_i}{1+g-x_i+qx_i},\ i \in N \tag{1}$$

then $\boldsymbol{h}[q] = (h[q], ..., h[q])$ is an equilibrium of the indefinitely repeated Covenant Game with matching over N.

We can think of the players hearing a "broadcast" report of a Player i's guilt at the time of offense with probability q, which makes Player i's guilt common knowledge. If no broadcast occurs, then all continue to cooperate with Player i, including the innocent Player $i(t)$ who did not benefit from Player i's required performance at period t. The players in $N - \{i, i(t)\}$ continue to cooperate with Player i because they don't know that Player i is guilty, and Player $i(t)$ cooperates after the offense because if he were to punish Player i unilaterally, then the others might hear a "broadcast" report that Player $i(t)$ is guilty. Note that if $q = 1$, then (1) reduces to the equilibrium condition of Proposition 2. On the other hand, if $q = 0$, then (1) can never be satisfied, which makes intuitive sense since in this case no offense is ever "broadcast" so the members of the community never have common knowledge of who are guilty.

PROOF. As before, let $x_i = \sum_{j \neq i} \mu_i[i(t) = j]$, so that

$$E_i(u_i(\boldsymbol{h}[q](t))) = x_i\,.$$

Again, the interesting case is where $x_i > 0$. Let f_i' be such that Player i deviates from the sequence $(h[q](\omega(t)))$, and let T_0 be the first stage such that $f_i'(\omega(T_0)) \neq h[q](\omega(T_0))$. Then we have two cases to consider:

Case i. If $f_i'(\omega(T_0)) = B$, then

$$E_i(u_i(f_i', \boldsymbol{h}[q]_{-i})) = \sum_{t=1}^{T_0-1} x_i p_i^t + 0 \cdot p_i^{T_0} + (1-q) \cdot \sum_{t=T_0+1}^{\infty} x_i p_i^t$$

because Player i follows B against Player $i(T_0)$ at stage $t = T_0$ and gets the discounted payoff $0 \cdot p_i^{T_0}$ of boycotting Player $i(T_0)$, and with probability $1 - q$, in each subsequent stage $t > T_0$, Player i will gain $1 \cdot p_i^t$ when he is matched. Note that

$$\sum_{t=1}^{\infty} x_i p_i^t > \sum_{t=1}^{T_0-1} x_i p_i^t + (1-q) \cdot \sum_{t=T_0+1}^{\infty} x_i p_i^t$$

so in this case, $E_i(u_i \circ \boldsymbol{h}[q]) > E_i(u_i(f_i', \boldsymbol{h}[q]_{-i}))$ for any positive value of p_i.

Case *ii*. If $f_i'(\omega(T_0)) = D$, then

$$E_i(u_i(f_i', \boldsymbol{h}[q]_{-i})) = \sum_{t=1}^{T_0-1} x_i p_i^t + (1+g) \cdot p_i^{T_0} + (1-q) \cdot \sum_{t=T_0+1}^{\infty} x_i p_i^t$$

because Player i follows D against Player $i(T_0)$ at stage $t = T_0$, then at this stage he will net the discounted gain $(1+g)p_i^{T_0}$ of exploiting Player $i(T_0)$, and for $t > T_0$, Player i gets the payoffs of cooperation with probability $1 - q$. So

$$E_i(u_i \circ \boldsymbol{h}[q]) \geq E_i(u_i(f_i', \boldsymbol{h}[q]_{-i}))$$

when

$$\sum_{t=T_0}^{\infty} x_i p_i^t \geq (1+g) \cdot p_i^{T_0} + (1-q) \cdot \sum_{t=T_0+1}^{\infty} x_i p_i^t$$

that is,

$$x_i p_i^{T_0} + \sum_{t=T_0+1}^{\infty} x_i p_i^t \geq (1+g) \cdot p_i^{T_0} + (1-q) \cdot \sum_{t=T_0+1}^{\infty} x_i p_i^t$$

or

$$x_i + \sum_{t=1}^{\infty} x_i p_i^t \geq (1+g) + (1-q) \cdot \sum_{t=1}^{\infty} x_i p_i^t \,. \tag{1}$$

(1) is equivalent to

$$q x_i \cdot \frac{p_i}{1-p_i} = q x_i \cdot \sum_{t=1}^{\infty} p_i^t \geq 1 + g - x_i \tag{2}$$

and (2) is satisfied when

$$p_i \geq \frac{1+g-x_i}{1+g-x_i+qx_i} \,.\ \square$$

We can also define Player i's *k-delayed Humean stochastic q-strategy* is defined by $h[k,q] = (h[k,q](\omega(t)))$ where

$$h[q](\omega(t)) = \begin{cases} P & \text{if } i \in N_t \text{ and } w_{i(t)}(t-k) = 1 \\ B & \text{if } i \in N_t \text{ and } w_{i(t)}(t-k) = 0 \end{cases} .$$

Proposition 6. Let the probability that a given Player $i \in N$ is matched be constant over stages. If $x_i > 0$ denotes the probability that Player i is matched at a given stage and

$$\left[(1+g)x_i - (1-q)^{k+1}x_i\right]p_i^{k+1} + (1+g)(1-x_i)p_i - (1+g-x_i) \geq 0,\ i \in N$$

then $\boldsymbol{h}[k,q] = (h[k,q], ..., h[k,q])$ is an equilibrium of the indefinitely repeated Covenant game with matching over N.

PROOF. Once more, let $x_i = \sum_{j \neq i} \mu_i[i(t) = j]$ denote Player i's probability of being matched, and note that at stage t, Player i's undiscounted expected payoff for following his end of $\boldsymbol{h}[k,q](t) = (h[k,q](t), ..., h[k,q](t))$ is

$$E_i(u_i(\boldsymbol{h}[k,q](t))) = x_i \,.$$

Hence Player i's overall expected payoff if he follows his end of $\boldsymbol{h}[k,q]$ is

$$E_i(u_i(\boldsymbol{h}[k,q])) = \sum_{t=1}^{\infty} x_i p_i^t$$

Suppose Player i deviates from $\boldsymbol{h}[k,q]$ by following some strategy $f_i' \neq h[k,q]$. If T_0 is the first period where $f_i'(T_0) \neq h[k,q](T_0)$, then

$$(1) \qquad E_i(u_i(f_i', \boldsymbol{h}[k,q]_{-i})) \leq \sum_{t=1}^{T_0-1} x_i p_i^t + (1+g)p_i^{T_0} + \sum_{t=T_0+1}^{T_0+k} (1+g)x_i p_i^t + (1-q)^{k+1} \cdot \sum_{t=T_0+k+1}^{\infty} x_i p_i^t$$

because the right member of (1) is the expected payoff Player i receives if: (a) Player i exploits $i(T_0)$ at period T_0, (b) then Player i successfully exploits every counterpart he meets from the T_0 to the $T_0 + k$th period, and (c) *none* of Player i's exploitations from T_0 to $T_0 + k$ are "broadcast", so that in the end Player i is not punished starting at the $T_0 + k + 1$st period or at any later no later period, and this latter event occurs with probability $(1-q)^{k+1}$. So $E_i(u_i(\boldsymbol{h}[k,q])) \geq E_i(u_i(f_i', \boldsymbol{h}[k,q]_{-i}))$ if

$$\sum_{t=T_0}^{\infty} x_i p_i^t \geq (1+g)p_i^{T_0} + \sum_{t=T_0+1}^{T_0+k} (1+g)x_i p_i^t + (1-q)^{k+1} \cdot \sum_{t=T_0+k+1}^{\infty} x_i p_i^t$$

or

$$p_i^{T_0} \cdot \sum_{t=0}^{\infty} x_i p_i^t \geq (1+g)p_i^{T_0} + p_i^{T_0} \cdot \sum_{t=1}^{k} (1+g)x_i p_i^t + p_i^{T_0}(1-q)^{k+1} \cdot \sum_{t=k+1}^{\infty} x_i p_i^t . \tag{2}$$

(2) simplifies to

$$x_i \cdot \sum_{t=0}^{\infty} p_i^t \geq (1+g) + \sum_{t=1}^{k} (1+g)x_i p_i^t + p_i^{k+1}(1-q)^{k+1} x_i \cdot \sum_{t=0}^{\infty} p_i^t$$

that is,

$$x_i \cdot \frac{1}{1-p_i} \geq (1+g) + (1+g)x_i \frac{p_i - p_i^{k+1}}{1-p_i} + p_i^{k+1}(1-q)^{k+1} x_i \cdot \frac{1}{1-p_i}$$

or

(3) $$x_i \geq (1+g)(1-p_i) + (1+g)x_i(p_i - p_i^{k+1}) + p_i^{k+1}(1-q)^{k+1}x_i$$

and (3) is satisfied when

$$p_i^{k+1}\left[(1+g)x_i - (1-q)^{k+1}x_i\right] + p_i(1+g)(1-x_i) \geq 1+g-x_i\ .\ \square$$

One additional folk theorem is of special interest. In this case, we will suppose for simplicity's sake that we have an ordinary repeated matching game, so that everyone plays at every stage. Let us define the *σ_i-trigger strategy* for playing the repeated Covenant Game with matching: At each stage of play, Player $i \in N$ follows a mixed strategy $\sigma_i(\omega)$ for $\omega \in \Omega$ over $\{P, D, B\}$.[6] $\sigma_i(t)$ denotes the pure strategy in $\{P, D, B\}$ specified by σ_i at stage t. We assume that each σ_i assigns positive probabilities to P and D only. The mixed strategies $\sigma_1, \ldots, \sigma_n$ are probabilistically independent. Let

$$\alpha_i = \mu_i(\sigma_i(t) = P),\ 1 - \alpha_i = \mu_i(\sigma_i(t) = D)$$

as defined by σ_i, and let $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n)$. Let

$$\Pi(i,j) = \alpha_i\alpha_j + (1-\alpha_i)\alpha_j(1+g) - \alpha_i(1-\alpha_j)l - (1-\alpha_i)(1-\alpha_j)c$$

so that at each stage, if each player $j \in N$ follows the mixed strategy σ_j, Player i's undiscounted expected payoff in Covenant Game is

$$E_i(u_i(\sigma_i, \boldsymbol{\sigma}_{-i})) = \sum_{j \neq i} \Pi(i,j)\mu[i(t) = j]\ .$$

For $i \in N$, define $z_i : t \to \{0, 1\}$ by

$$z_i(t) = \begin{cases} 1 & \text{if } C(i,t) \text{ obtains} \\ 0 & \text{otherwise} \end{cases}$$

where $C(i, 1)$ obtains for each $i \in N$ and for $t > 1$,

---

[6]That is, Player i pegs his choice of pure strategy on the results of a random experiment that defines a random variable σ_i with outcomes in $\{P, D, B\}$.

$$C(i,t) = \forall(T \leq t-1)(C(i(T),T) \rightarrow a_i(T) = \sigma_i(T)) \ .$$

In words, a player is a *conformist* at stage t if she has always followed her mixed strategy over the first $t-1$ stages with other conformists, and is a *nonconformist* if she has ever deviated from her mixed strategy when paired with a conformist over the first $t-1$ stages. So the definition of $C(i,t)$ is similar to the definition of innocence that is used to define the Humean strategies. As before $z_i(t)$ is Player i's "marker", and $z_i(t) = 1$ if Player i is a conformist and $z_i(t) = 0$ otherwise. Let $\omega(t) = (z_1(t), ..., z_n(t))$. Then Player i's matching σ_i-trigger strategy is defined by $f_{\sigma_i} = (f_{\sigma_i}(\omega(t)))$ where

$$f_{\sigma_i}(\omega(t)) = \begin{cases} \sigma_i & \text{if } z_{i(t)} = 1 \\ B & \text{if } z_{i(t)}(t) = 0 \end{cases} .$$

That is, Player i follows his mixed strategy σ_i in the stage game if his counterpart $i(t)$ is a conformist, and otherwise Player i punishes his counterpart by boycotting.

The following result gives conditions under which a profile of σ_i-trigger strategies forms an equilibrium of the repeated Covenant Game with matching.

Proposition 7. Let $\alpha_* = \min\{\alpha_i : i \in N\}$, and for $i \in N$ let

$$\Pi_i^* = \alpha_i\alpha_* + (1-\alpha_i)\alpha_*(1+g) - \alpha_i(1-\alpha_*)l - (1-\alpha_i)(1-\alpha_*)c \ .$$

If $N_t = N$ for each period t and

$$p_i \geq \frac{1+g-\Pi_i^*}{1+g}, \ i \in N \tag{1}$$

then $\boldsymbol{f_\sigma} = (f_{\sigma_1}, ..., f_{\sigma_n})$ is a correlated equilibrium of the indefinitely repeated Covenant Game with matching over N.

PROOF. Let $M = \{j \in N : \alpha_j = \alpha_*\}$. Note that for each pair $i, j \in N$,

$$\Pi(i,j) \geq \Pi_i^* \ . \tag{2}$$

Given $i \in N$, we have

$$(3) \qquad E_i(u \circ \boldsymbol{f}_{\sigma}) = \sum_{t=1}^{\infty} E_i(u_i(\boldsymbol{\sigma}))p_i^t$$
$$= \sum_{t=1}^{\infty} \left(\sum_{j \neq i} \Pi(i,j)\mu_i[i(t) = j] \right) \cdot p_i^t \, .$$

Now consider any strategy f_i' where Player i deviates from the sequence $(f_{\sigma_i}(\omega(t)))$. Let T_0 be the first stage such that $f_i'(\omega(T_0)) \neq f_{\sigma_i}(\omega(T_0))$. Then

$$(4) \qquad E_i(u_i(f_i', \boldsymbol{f}_{\sigma_{-i}})) \leq \left(\sum_{t=1}^{T_0-1} \left(\sum_{j \neq i} \Pi(i,j)\mu_i[i(t) = j] \right) \cdot p_i^t \right) + (1+g) \cdot p_i^{T_0}$$

because (*i*) if Player i deviates from $\boldsymbol{f}_{\sigma}$ for the first time at $t = T_0$, then at stage T_0 he will net at most the discounted gain $(1+g)p_i^{T_0}$ of exploiting Player $i(T_0)$, and (*ii*) at each subsequent stage $t > T_0$ Player i receives 0 because at each of these stages Player i's counterpart $i(t)$ boycotts. By (2) and (3), we also have

$$(5) \qquad E_i(u_i \circ \boldsymbol{f}_{\sigma}) \geq \sum_{t=1}^{T_0-1} \left(\sum_{j \neq i} \Pi(i,j)\mu_i[i(t) = j] \right) \cdot p_i^t + \sum_{t=T_0}^{\infty} \Pi_i^* p_i^t$$

because the right member of (5) is Player i's expected payoff if, starting at stage T_0, he is always paired with counterparts in M. By (4) and (5), $E_i(u_i \circ \boldsymbol{f}_{\sigma}) \geq E_i(u_i(f_i', \boldsymbol{f}_{\sigma_{-i}}))$ when

$$\frac{\Pi_i^* p_i^{T_0}}{1 - p_i} = \sum_{t=T_0}^{\infty} \Pi_i^* p_i^t \geq (1+g)p_i^{T_0}$$

or

$$(6) \qquad \frac{\Pi_i^*}{1 - p_i} \geq 1 + g$$

and (6) is satisfied when

$$p_i \geq \frac{1 + g - \Pi_i^*}{1 + g} \, . \; \square$$

Proposition 7 is a generalization of the early folk theorem that says that "grim trigger" is an equilibrium of the indefinitely repeated Prisoners' Dilemma played between a fixed pair of players, a result first discovered by John Nash (Flood 1958). For all $i \in N$, $\Pi_i^* > 0$ when $\alpha_* > \frac{\alpha_i l + c - c\alpha_i}{1+g-g\alpha_i+\alpha_i l+c-c\alpha_i}$, so there is always an equilibrium characterized by the σ_i-trigger strategies whenever the latter inequality is satisfied for each $i \in N$. One can think of the players in M as the "nastiest" players in the system, since they double-cross most often against other conformists. The key idea underlying the proof of Proposition 7 is that following one's own end of $\boldsymbol{f}_\sigma$ can be a best response to the others' strategies even if from a certain stage onward one is always paired with the "nastiest" possible counterparts. The equilibrium conditions assume only that the indefinitely repeated game is ordinary. One can identify weaker necessary conditions for equilibrium than condition (1) if one places additional restrictions on the matching protocol.

As illustrations like Figure 4 make clear, it is by no means a foregone conclusion that a community of agents will follow a Humean-type equilibrium of conditional cooperation in the indefinitely repeated Covenant Game. Even if we assume that such a community ultimately settles into some equilibrium of the indefinitely repeated game, this is no guarantee that the equilibrium is one of conditional cooperation. The players of this community might settle into one of the equilibria where all follow a Humean strategy, or the equilibrium where all boycott always, or some intermediate equilibrium where all perform some of the time and boycott some of the time. Proposition 7 shows that the players might even follow an equilibrium where some exploit others by double-crossing sometimes. Note that in order for players to follow any of the Humean strategies described in this section correctly at each stage, certain facts regarding their situation must be known to all players. In particular, if players start a punishment cycle, all must know who are "labeled" guilty and who are "labeled" innocent. This can occur if there is some mechanism or institution that publicly announces or "broadcasts" the identities of

guilty players to the entire community. Such a broadcast makes the identities of the guilty and the innocent common knowledge among the community, assuming the broadcasting mechanism never fails to report the identities of the guilty. If the mechanism can fail, so that some of the actually guilty "slip through the cracks", then it is still possible for the community to follow an equilibrium of conditional cooperation based upon one of the stochastic Humean strategies discussed above.

Plainly, the cooperation in the Humean-type equilibria can unravel if the players are prone to the sort of mistakes "trembles" in executing their strategies that are used to characterize equilibrium refinements. For instance, if the equilibrium $\boldsymbol{h}$ of the basic Humean strategy is amended so that at each stage t, a given Player i deviates from $\boldsymbol{h}$ with any positive probability ϵ_i, then in time with probability one everyone in the community will be boycotting everyone else. Another way for cooperation in a community of Humeans to be undermined is if the reporting institution broadcasts erroneous reports of certain players' guilt, either by mistake or because this institution is corrupt. One might try to construct reputational equilibria that are more robust against these kinds of errors by adding additional structure to the base game. Milgrom, North and Weingast (1990) take this sort of approach in their analysis of merchant trade in 14th century Europe. In their model, individual traders at a fair may present complaints of being cheated to a judge, who for a fee renders a judgment of innocence or guilt. If judged guilty, a merchant must either make restitution as determined by the judge or in effect be excluded from future trading at this fair. Hill (2004) shows that the cooperation in the Milgrom-North-Weingast model is robust with respect to mistakes on the part of the judge so long as the error rate is sufficiently low. Similarly, one can construct Humean reputation equilibria that are more robust to trembles and false reports by introducing more structure to the model, in effect extending the interaction of the Covenant Game into a more complex game where players have to defer to the judgments of a central agent and support this agent at some personal cost. This is to take steps much like those that

institute the Leviathan that Hobbes claims is generally necessary to enforce compliance with covenants. However, in the next section I wish to explore a different possibility. The reputational equilibria developed in this section tacitly presuppose that some formal mechanism exists that can generate common knowledge among community members. Below, I will consider the possibility that performance in covenants can be enforced by informal communication only.

§4. Decentralized Reputation Effects

Suppose that no mechanism or institution exists in the community that can generate common knowledge among its members. Can such a community sustain a norm of conditional cooperation that excludes offensive violators from covenants over time? The members in such a community cannot generate common knowledge, but perhaps they can at least spread information via informal communication. Perhaps performance in covenants can be enforced by *gossip*. More precisely, perhaps the members of a community can exchange information when they interact, with the result that the identities of offensive violators in the community are spread through "the community grapevine". Now Humeans still perform in covenants with those counterparts they believe to be innocent and boycott those they believe to be guilty, but they have to rely upon their individual experiences and the information they receive from other counterparts to form their beliefs. Such a system will generally not be in equilibrium if any offensive violators are present. For if a member of the community offensively violates a covenant, the exploited party will know that the violator is guilty but others in the system may not know this, because by assumption nothing in the community serves as an authority that can make innocence or guilt common knowledge.

Consequently, it is not possible to establish any analytical results for such a community of individuals analogous to the folk theorems of §3. But perhaps we can learn something of the properties of such a system if we analyze this system

computationally. Gaylord and D'Andria (1998) present an early computational analysis of the spread of bad reputation. While we shall see that their model is altogether too crude to give much insight into how behavior in covenant games might evolve in real human communities, they are pioneers in the use of computational models to analyze the spread of reputation. To model this situation, let the members of a community occupy positions in an $r \times r$ lattice whose edges "wrap around", so that their territory is topologically equivalent to a torus. At each stage, a member chooses a direction, north, south, east or west, at random in the cell he occupies. If the cell this member faces is empty, the member migrates into this new cell. If the cell this member faces is occupied by a second member whose direction faces the cell of the first, they are matched and they play the Covenant Game. Otherwise this member does not interact and gets the payoff of working alone. Figure 5 depicts such a lattice where $r = 50$.

[See Figure 5.]

In this lattice, 70% of the cells are occupied. Half of the community members are Humeans, who when matched perform with counterparts they believe to be innocent and boycott counterpart they believe to be guilty. The other half are Fooles who are willing to offensively violate a covenant. The parameters of this particular lattice are identical to those Gaylord and D'Andria use in their analysis. Gaylord and D'Andria have the agents in this system play the Covenant Game when they are matched with parameters $g = c = 1$and $l = 2$. They assume that a Foole always chooses D if matched unless the counterpart has double-crossed this Foole before, in which case the Foole chooses B. A Humean always chooses B if the counterpart is on this Humean's "blacklist", and otherwise the Humean chooses P. If the Humean does enter into a Covenant, this Humean and the counterpart exchange blacklists. The counterpart is added to the Humean's blacklist as well only if this counterpart chooses D, which is an offensive

violation. Figure 6 shows the average accumulated payoffs of the Humeans and the Fooles over 500 stages of plays.

[See Figure 6.]

The results in this figure replicate Gaylord and D'Andria's findings in their computational study (1998). They appear to give a powerful evidence that a Humean-type policy of performing in covenants with those one has not learned to be guilty through personal experience or informal information exchange is far superior to a policy of following the Foole's advice and offensively violating covenants when one can. In this model, a bad reputation spreads rapidly throughout the community, leading the Fooles to fare very poorly compared with their Humean counterparts.

Nevertheless, the relative prospects of Fooles and Humeans change dramatically if the Fooles are even slightly more sophisticated than those of the Gaylord-D'Andria model. In that model, only actual offensive violations are ever added to a player's blacklist. In effect, all agents are assumed to tell only the truth always. Suppose that a Foole is willing to lie as well as double-cross in a covenant. Specifically, suppose a Foole adds the identity of a counterpart he double-crosses to his own blacklist. This Foole's motivation for lying about those he exploits is simple: If the Foole gives false information to others who might be Humeans, they may be unwilling to interact with those innocent victims the Foole has added to his own blacklist and therefore will not learn that the Foole is on *their* blacklists. Such Humeans might be especially willing to believe a Foole who does not double-cross all of the time. If the Foole double-crosses only occasionally, then the counterparts he does not exploit have no reason as of yet to believe that this Foole is not another Humean like themselves. Figure 7 summarizes the results over 500 stages of repeated Covenant Game with Random matching where now the Fooles adopt a more complex strategy than simply double-crossing all of the time.

[See Figure 7.]

In this computer simulation, the Humeans follow the same strategy as before, but now Fooles double-cross counterparts not on their blacklists at random with probability $\frac{1}{2}$. If a Foole double-crosses a Humean, the Foole adds the identity of this Humean to his blacklist and spreads this identity to the blacklists of all who interact with him. As Figure 7 shows, now Fooles fare better than Humeans on average over all stages of play. The Fooles have turned the tables on the Humeans simply by adopting a slightly more complex strategy than simply double-crossing any counterpart willing to covenant.

§5. Conclusion

At the start of this paper I hinted that the classic argument for keeping promises presented by Hobbes and Hume avoids a fundamental question: Can the members of a community follow a policy of performing in covenants made with innocent members and boycotting guilty members? Not surprisingly, the answer to this question is contingent upon the circumstances of this community. The folk theorems for the indefinitely repeated Covenant Game show that a variety of equilibria where players perform with the innocent and boycott the guilty are possible. In a community that follows such a Humean-type equilibrium, would-be fooles do have a decisive reason to perform in covenants. The would-be fooles should perform in order to maintain their reputation, which in the folk theorems is summarized by the innocence or guilt "marker". But in order to sustain any such Humean-type equilibrium of conditional performance, the community requires an institution that can generate the common knowledge its members require in order to follow their parts of the equilibrium. Humean-type equilibria presuppose that the identities of the guilty and the innocent are common knowledge, or at the very least that the identities of the guilty and the innocent are made common knowledge with high probability. Such common knowledge can exist in communities

that have a reliable judge together with a reliable communication network. Reputation alone can enforce good conduct among the members of a clan who meet regularly and receive information from certain designated members who are "elders", or a church with truthful ministers, or in a larger civil society with a reliable broadcasting network. But in a community with no such structures, reputation alone is far less likely to enforce good conduct. The computational models considered above show that fooles can fare better than Humeans in a community that must rely upon private information or "gossip" only to spread information. So the key to reputational enforcement of covenants is common knowledge, which presupposes mechanisms that can make certain information public. This conclusion dovetails with Hobbes' analysis of life in the State of Nature. Hobbes expressly denies that people in a State of Nature can have any of the means such as navigation or letters that in civil society facilitate the transmission of knowledge (*Leviathan* 13:9). So Hobbes has the means available to argue that his rebuttal to the foole is consistent with his claims that civil society is a necessary condition for the rationality of forming and keeping covenants.

One might conclude that the real lesson of Hobbes' and Hume's reputational defense of performing in covenants is that offensive violation is not rational when one resides in an ideal community where all or at least most are rational and have the common knowledge necessary to sustain a Humean-type equilibrium. If only we all lived in a community where all, or nearly all, reason "correctly" and perform in covenants exactly those others who reason "correctly" and boycott those who follow the foole's advice, then preserving one's reputation preservation really would give each member of the community sufficient reason to always perform in their covenants. Since we in fact live in communities where not all reason "correctly" and one cannot easily distinguish the fooles from the Humeans, we need government to enforce covenants, after all. However, this argument is too quick. In fact, it does not follow that it is never rational to offensively violating a covenant even if everyone in the community is rational and all

have the common knowledge needed to distinguish the guilty from the innocent. As the folk theorems of §3 show, there are equilibria of the indefinitely repeated Covenant Game where some of the players double-cross others some of the time. In these equilibria, those players who are occasionally exploited do not try to punish the offensive violators by boycotting because the costs of starting a punishment cycle are even greater than tolerating the occasional double-cross. These equilibria can even allow some of the community members to achieve greater payoffs by occasionally double-crossing than all would achieve by following a Humean-type equilibrium. So it does not follow that the members of an ideal community will settle into a pattern of always performing in the covenants they make as a consequence of their rationality and common knowledge alone. The analysis of the repeated Covenant Game yields a rather different lesson, similar to the lesson several other authors present in complimentary studies of the social contract (Sugden 1986, Binmore 1994, 1998, Skyrms 1996, 1998): Rationality alone does not explain reciprocal cooperation.

REFERENCES

Aumann, Robert. 1976. 'Agreeing to Disagree.' *Annals of Statistics* 4: 1236-9.

Aumann, Robert. 1987. 'Correlated Equilibrium as an Expression of Bayesian Rationality', *Econometrica* 55, 1-18.

Axelrod, Robert. 1981. 'The emergence of cooperation among egoists'. *American Political Science Review*, 75, 306-318.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books, Inc.

Batali, John and Kitcher, Philip. 1995. 'Evolution of Altruism in Optional and Compulsory Games.' *Journal of Theoretical Biology* 175: 161-171.

Binmore, Ken. 1994. *Game Theory and the Social Contract Volume I: Playing Fair*. Cambridge, Massachusetts: MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract Volume II: Just Playing.* Cambridge, Massachusetts: MIT Press.

Dekel, Eddie and Gul, Faruk. 1996. "Rationality and Knowledge in Game Theory", working paper, Northwestern and Princeton Universities.

Ellison, Glenn. 1994. 'Cooperation in the Prisoner's Dilemma with anonymous matching.' *Review of Economic Studies*, 61: 567-588.

Flood, Merrill M. 1958. 'Some Experimental Games.' *Management Science* 5, pp. 5-26.

Gaylord, Richard J. and D'Andria, Louis J. 1998. *Simulating Society: A **Mathematica** Toolkit for Modeling Socioeconimic Behavior.* New York: Springer Verlag.

Hampton, Jean. 1986. *Hobbes and the Social Contract Tradition.* Cambridge: Cambridge University Press.

Hill, Douglass E. 2004. 'Errors of Judgment and Reporting in a Law Merchant System.' *Theory and Decision* 56: 239-268.

Hobbes, Thomas. (1651) 1991. *Leviathan*, ed. Richard Tuck. Cambridge: Cambridge University Press.

Hume, David. (1740) 2000. *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Kandori, Michihiro. 1992. 'Social norms and community enforcement.' *Review of Economic Studies*, 59: 63-80.

Kavka, Gregory. 1986. *Hobbesian Moral and Political Theory.* Princeton: Princeton University Press.

Kitcher, Philip. 1993. 'The Evolution of Human Altruism.' *The Journal of Philosophy* 90: 497-516.

Lewis, David. 1969. *Convention: A Philosophical Study.* Cambridge, Massachusetts: Harvard University Press.

Linster, Bruce. 1992. 'Evolutionary stability in the infinitely repeated Prisoners' Dilemma played by two-state Moore machines.' *Southern Economic Journal*, 58: 880-903.

Millgrom, P. R., North, D. C., and Weingast, B. R. (1990) 1997. 'The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs.' *Economics and Politics* 2: 1-23. Reprinted in Klein, Daniel B., *Reputation: Studies in the Voluntary Elicitation of Good Conduct.* Ann Arbor, Michigan: University of Michigan Press, pp. 243-266.

Skyrms, Brian. 1996. *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

Skyrms, Brian. 1998. 'The shadow of the future', in *Rational Commitment and Social Justice: Essays for Gregory Kavka*, ed. Jules Coleman and Christopher Morris. Cambridge: Cambridge University Press: 12-22.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation and Welfare.* Oxford: Basil Blackwell, Inc.

Figure 3. Convex Hull of the Covenant Game with $g = l = 1, c = \frac{1}{2}$

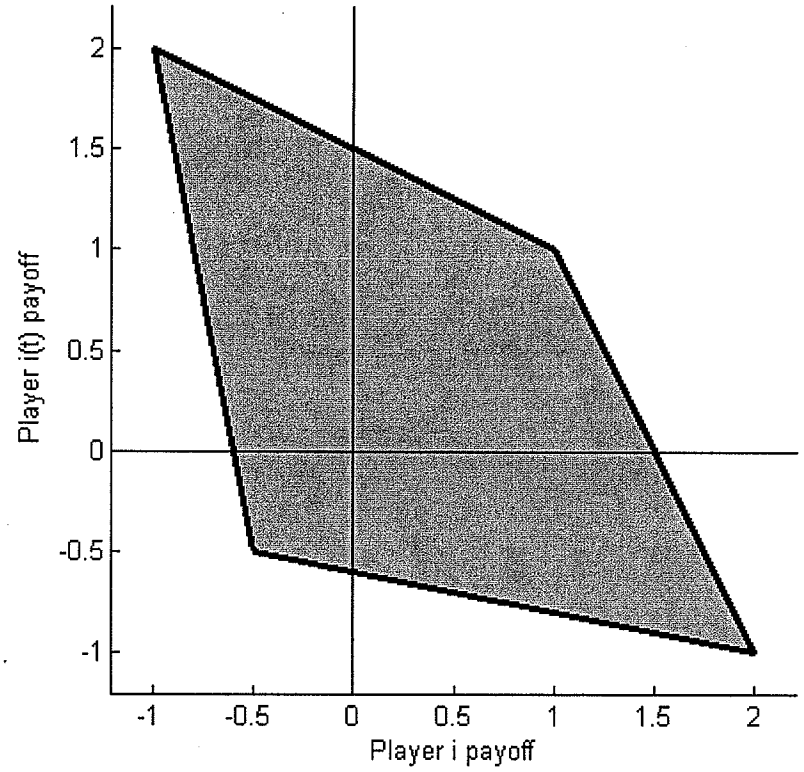
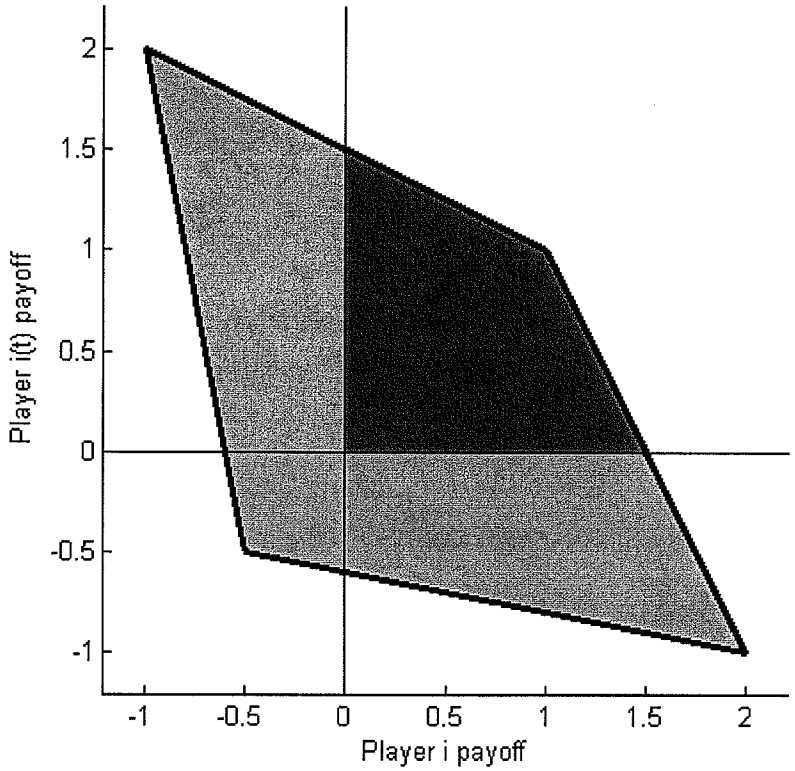


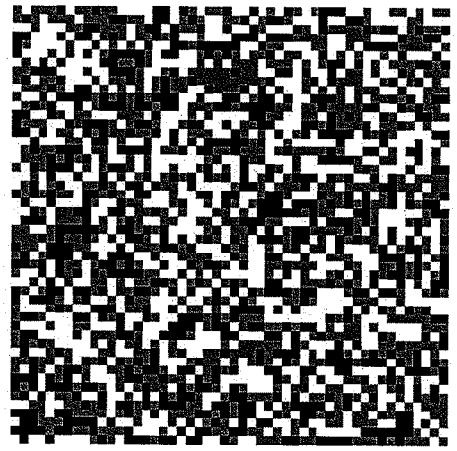
Figure 4. Average Payoff Vectors of the Correlated Equilibria of the Repeated the Covenant Game with $g = l = 1, c = \frac{1}{2}$



convex hull payoffs

average payoffs of equilibria of the indefinitely repeated game

Figure 5. 50 × 50 Lattice of Players who Play the Repeated the Covenant Game with Random Matching



Humean, Foole, unoccupied cell

Figure 6. Average Accumulated Payoffs of Humeans and Naive Fooles over 500 Stages of Play

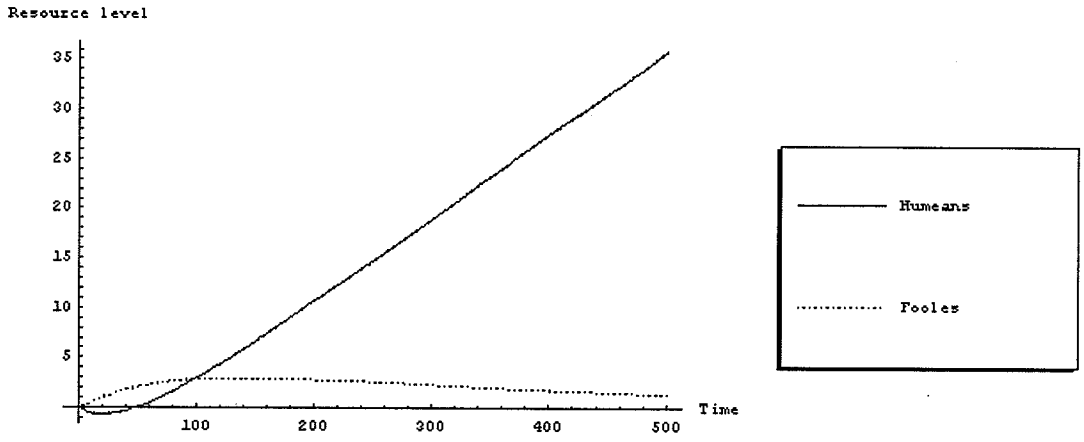


Figure 7. Average Accumulated Payoffs of Humeans and More Sophisticated Fooles over 500 Stages of Play