Synthese (2007) 157:167–195
Knowledge, Rationality & Action 155–183
DOI 10.1007/s11229-006-9147-4

ORIGINAL PAPER

Covenants and reputations

Peter Vanderschraaf

Received: 14 March 2005 / Accepted: 18 December 2006 / Published online: 26 May 2007

Abstract In their classic analyses, Hobbes and Hume argue that offensively violating a covenant is irrational because the offense ruins one's reputation. This paper explores conditions under which reputation alone can enforce covenants. The members of a community are modeled as interacting in a Covenant Game repeated over time. Folk theorems are presented that give conditions under which the Humean strategy of performing in covenants only with those who have never offensively violated or performed with an offensive violator characterizes an equilibrium of the repeated Covenant Game. These folk theorems establish that for certain ideal settings Hobbes' and Hume's arguments against offensively violating covenants are compelling. However, these ideal settings presuppose that the community has certain mechanisms that generate common knowledge of the identities of those with whom one should perform. I analyze the results of computer simulations of the interactions in a community whose members must rely upon private communication alone. The computer simulation data show that in this community, reputation effects cannot effectively deter members from offensively violating covenants. I conclude that Hobbes' and Hume's warnings against offensive violation are compelling only on condition that the community is sufficiently structured to generate common knowledge among its members. I also conclude that even in such structured communities, the Humean strategy is not the uniquely "correct" policy.

Keywords Covenant game · Folk theorem · Hobbes · Hume · Offensive violation · Simulation · Trigger strategy

P. Vanderschraaf (✉)
Department of Philosophy, Boston University,
745, Commonwealth Ave, Boston,
MA 02215, USA
e-mail: pvanderschraaf@gmail.com

1 Introduction: 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 claims that if he enters into a covenant with another who performs her end of the covenant first or is compelled by some third party to perform, then he should not perform his end of the covenant. Hobbes does not articulate the Foole's rationale for this claim, perhaps because Hobbes thinks this rationale is obvious. In such circumstances, the Foole receives the benefits promised him by the terms of the covenant no matter what he does. Consequently, the Foole thinks 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.

Of course, if the Foole has analyzed the situation correctly, then would the other party to the agreement not replicate the Foole's reasoning before she is to act, and then act so as to forestall the Foole's exploitative conduct? 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 tomorrow. 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

[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).

external authority. Nevertheless, both Hobbes and Hume maintain that self-interested agents can have good reason to comply with the agreements they make after all, even if they are not coerced into compliance. For typically, a single interaction like the commitment problem Hume's farmers face is embedded in a complex sequence of social interactions that 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 they can trust her to honor future agreements, and so they will make and keep agreements with her in the future. Consequently, 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 believe that under certain circumstances, failing to honor a covenant one cannot expect the other parties to honor, or *defensively violating* the covenant, is acceptable. But to offensively violate a covenant is to commit an injustice. 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)

One who follows the Foole's advice and offensively violates a covenant might enjoy an immediate gain if the others perform as expected. For one then gains the benefits of the others' compliance and avoids the costs of one's own compliance. However, according to Hobbes and Hume, rational and well-informed agents will never enter into covenants with one who has ever once exploited others by offensively violating a covenant. And Hobbes and Hume think that losing all such opportunities for future benefits of others' performance in covenants far outweighs any immediate gain from refusing to reciprocate when others have performed their parts of a covenant.

This kind of reply to the Foole raises several fundamental puzzles. Some have rightly pointed out that if Hobbes' and Hume's refutations of those who reason like

the Foole and the farmers are entirely successful, then people do not need a government to enforce compliance with covenants with credible threats of force, contrary to what Hobbes maintains (Hampton, 1986; Kavka, 1986). Hobbes and Hume apparently assume that an individual like the Foole will encounter others who perform in covenants exactly with those who are in "good standing," and that offensively violating a covenant is sufficient reason to be excluded from the set of those in good standing. Performing one's ends of covenants with those in good standing and with no one else is evidently the "correct" policy according to Hume and Hobbes. But they do not clearly explain what makes this policy "correct." Moreover, it is by no means obvious that the members of a community are *able* to follow such a conditionally cooperative policy. Hume and Hobbes have little to say regarding the conditions necessary for a policy of conditional performance to regulate a community of individuals.

This paper explores conditions under which reputation alone can enforce performance in covenants. Section 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. Section 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. In particular, they show that a *Humean strategy* of performing in covenants only with those who have never offensively violated or entered into a covenant with an offensive violator characterizes an equilibrium of the repeated Covenant Game. However, these folk theorems presuppose that the community has at its disposal certain mechanisms that generate common knowledge.[2] In communities that lack such structures, reputation alone may not give a would-be Foole good reason to honor covenants. In such communities there are no equilibria of conditional cooperation, precisely because these communities cannot generate common knowledge. Consequently one cannot establish analytically for such communities that any pattern of conditional performance is stable, or even more beneficial than following the Foole's advice. I propose to analyze such communities in a different manner, that is, via computer simulations. Section 4 presents a simple computational model of a community members must rely upon private communication alone. Computer simulations of the interactions in this community show that such a community cannot effectively deter Fooles from offensively violating covenants. The concluding Sect. 5 considers some of the lessons to be drawn from the analysis of the repeated Covenant Game. I conclude that Hobbes' and Hume's warnings against offensive violation are compelling only on condition that the community is sufficiently structured so as to be able to generate common knowledge among its members. I also conclude that even in such structured communities, the Humean strategy is not the uniquely "correct" policy.

[2] Lewis (1969) presented the first analysis of common knowledge. A proposition A is *Lewis-common knowledge* among a group of agents if each agent knows that all know A and knows that all can infer the consequences of this mutual knowledge (Lewis, 1969, pp. 56–57).
Dov Samet is credited with an especially satisfactory characterization of common knowledge: A *public event* for a community is some event that cannot occur without all in the community knowing this. Anything implied by a public event is common knowledge after the public event occurs.

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. $N = \{1, \ldots, n\}$ 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. At each stage, players will interact in a *Covenant game* with assigned counterparts. 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 Covenant 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 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 $X_t : N_t \longrightarrow N_t$ 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 Fig. 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 Fig. 2.

Fig. 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

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

Fig. 2 Prisoners' Dilemma subgame

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

The Fig. 2 game is a Prisoners' Dilemma. Consequently, the Fig. 1 game is sometimes called *optional Prisoners' Dilemma* or *Prisoners' Dilemma with opting out* (Batali & Kitcher, 1995; Kitcher, 1993). 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. 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

$$\begin{aligned} x_{i1} &= \mu_i\left[a_{i(t)}(t) = P\right], \\ x_{i2} &= \mu_i\left[a_{i(t)}(t) = D\right], \text{ and} \\ x_{i3} &= \mu_i\left[a_{i(t)}(t) = B\right] = 1 - x_{i1} - x_{i2}. \end{aligned}$$

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[3] 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 $f_i^t : \Omega \longrightarrow \{P, D, B\}$ and f_i^t is $\mathcal{H}_i$-measurable. $\boldsymbol{f} = (f_1, \ldots, f_n)$ 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. The profile

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

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

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

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

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

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

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

$$E_i\left(u_i \circ \boldsymbol{f}\right) \geq E_i\left(u_i \circ \left(f_i', \boldsymbol{f}_{-i}\right)\right) \text{ for all } f_i' \in S_i.^4$$

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 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. Moreover, community members should be able to identify certain kinds of offensive violation more easily in the repeated Covenant Game than in the repeated 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 some of offensive violations of a covenant. If, for instance, one player observes another double-crossing in a given stage of repeated Prisoners' Dilemma, she might have trouble determining whether the double-crosser is offensively violating a covenant, or is merely punishing an offensive violator as a conditionally cooperative strategy for playing repeated Prisoners' Dilemma might require. In the Humean strategies for playing the repeated Covenant Game discussed below, 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

This section presents some basic folk theorems for the indefinitely repeated Covenant Game. These folk theorems establish conditions in repeated covenant situations where strategies of conditional performance are Bayesian rational and superior to strategies where one offensively violates. Hence, they partially vindicate Hobbes' and Hume's reputational argument for keeping promises. The results stated 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, Kandori and Ellison assume that every player is matched at every period and that an offensive violation is certain to start a punishment cycle. Moreover, in their models a player can follow only one of the pure strategies of the Prisoners' Dilemma at any given stage. The matching model developed here allows for the possibility that players follow mixed strategies in the Covenant Game, and does not assume that every player

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

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

Aumann (1974, 1987) gave the first precise formulation of correlated equilibrium.

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. Note that in the results stated here, 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. 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.

A first folk theorem establishes the set of average payoffs in the indefinitely repeated Covenant game that can be sustained in an equilibrium. The proofs of Proposition 1 and of the other results stated in this section are given in Appendix 1.

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 $\{P, D, B\} \times \{P, D, B\}$ as follows: $\Omega = \{\omega_1, \omega_2\}$ where $x = \mu_i[\omega = \omega_1]$ and $1 - x = \mu_i[\omega = \omega_2]$ for all $i \in N$. $s\colon \Omega \longrightarrow \{P, D\} \times \{P, D\}$ is a map that yields Player $i \in N_R$ an expected payoff of $u_R \in (0, 1]$ and counterpart Player $i(j) \in N_C$ an expected payoff of $u_C \in (0, 1]$. 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

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

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

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

Proposition 1 shows us that any point in the convex hull of payoffs of the Covenant Game such that each component is nonnegative can be sustained in an equilibrium of the corresponding ordinary indefinitely repeated game. Figure 3 gives the graph of this set for the special case where $g = l = 1$ and $c = \frac{1}{2}$.

The equilibrium of Proposition 1 is similar to some of the "contagion" equilibria of a repeated Prisoners' Dilemma presented in Kandori (1992). The basic underlying idea is as follows: The community members are partitioned into two distinct classes. 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. Here punishment spreads like

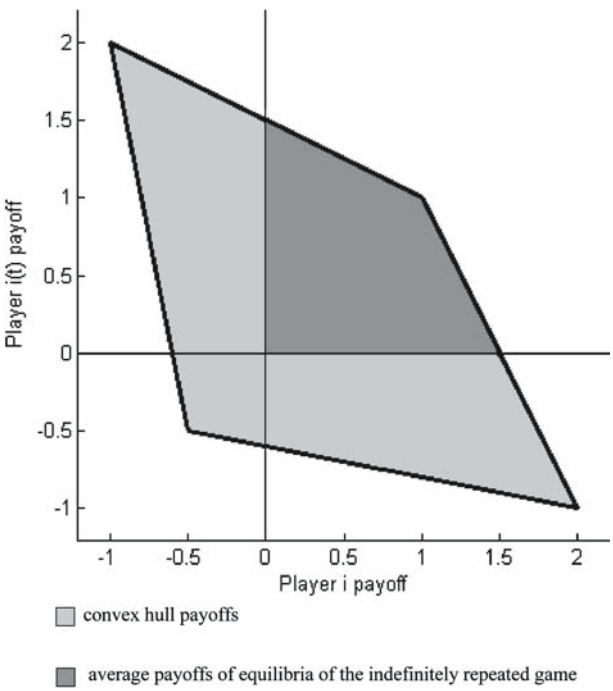

Fig. 3 Average payoff vectors of the correlated equilibria of the repeated the Covenant Game with $g = l = 1, c = \frac{1}{2}$

an epidemic throughout the community because just one offensive violator in effect "infects" everyone else.

Proposition 1 show us what kind of equilibria are possible in the indefinitely repeated Covenant Game. But the equilibria of "good reputation," if there are any, should not depend upon unusual matching protocols and should punish only those who unilaterally deviate from the equilibrium. To construct such a class of equilibria, let us define a 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 $\{P, D, B\}$.[5] $\sigma_i(t)$ 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

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

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

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

[5] 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 $\{P, D, B\}$.

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

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

For $i \in N$, define $z_i : t \rightarrow \{0, 1\}$ 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 $t > 1$,

$$\begin{aligned} C(i, t) = \forall(T \leq t-1)\forall(j \in N)[&(j = i(T) \wedge C(j, T) \rightarrow a_i(T) = \sigma_i(T)) \\ &\wedge(j = i(T) \wedge \neg C(j, T) \rightarrow a_i(T) = B)]. \end{aligned}$$

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 has always boycotted any counterpart who is not a conformist. A player becomes a nonconformist if she ever deviates from her mixed strategy when paired with a conformist or enters into a covenant with a nonconformist. Note that here Player i must enter into and perform in a covenant when she is matched with a conformist.[6] Player i is also not allowed to enter into a covenant with a nonconformist. $z_i(t)$ is Player i's "marker," and $z_i(t) = 1$ if Player i is a conformist and $z_i(t) = 0$ otherwise. Let $\omega(t) = (z_1(t), \ldots, z_n(t))$. Then Player i's matching σ_i-trigger strategy is defined by $f_{\sigma_i} = (f_{\sigma_i}(\omega(t)))$ where

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

That is, Player i 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 2 *Let* $\alpha_* = \min\{\alpha_i : i \in N\}$, *and for* $i \in N$ *let*

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

Let the probability that a given Player i is matched be constant over stages. If x_i *denotes the probability that Player i is matched at a given stage and*

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

then $\boldsymbol{f}_\sigma = (f_{\sigma 1}, \ldots, f_{\sigma n})$ *is a correlated equilibrium of the indefinitely repeated Covenant Game with matching over N.*

Proposition 2 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 in 1950 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

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

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 2 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 has the bad luck to be always matched with the "nastiest" possible counterparts. The equilibrium conditions assume only that the probability of being matched with *some* counterpart remains constant over time. One can identify weaker sufficient conditions for equilibrium than condition (1) if one places additional restrictions on the matching protocol:

Proposition 3 *If Player i's probability of being matched with a conformist is at least* $\frac{x_i(1+g)}{1-g+\Pi_i^*}$, *then* f_{σ_i} *is Player i's best response to his counterparts.*

A special case of the σ_i-trigger strategy is the *Humean strategy*, so-called because I think Hume articulates the supposed consequences of offensive violation more clearly than does Hobbes: For $i \in N$, define $z_i : t \to \{0, 1\}$ by

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

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

$$\begin{aligned} I(i, t) = \forall(T \leq t-1)\forall(j \in N)[(j = i(T) \wedge I(j, T) \to a_i(T) = P) \\ \wedge(j = i(T) \wedge \neg I(j, T) \to a_i(T) = B)]. \end{aligned}$$

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. Hobbes and Hume evidently assume in their warnings against offensive violation that innocent individuals are sure to perform in covenants with other innocents and shun the guilty, and that double-crossing even once renders one guilty. Neither say what happens if an innocent individual who enters into a covenant with a guilty individual, perhaps because this would be a violation of their assumption. Moreover, neither consider the possibility that individuals might follow mixed strategies such as the mixed strategies of σ_i-trigger. This is not surprising since the members of actual communities seldom explicitly peg their choices on random experiments in covenant situations. The Humean strategy defined here assumes that players follow only pure strategies in a given Covenant Game and also specifies what happens if an innocent player "slips" and fails to shun a guilty player. If innocent Player i double-crosses an innocent counterpart, then Player i becomes guilty, following Hobbes and Hume. If innocent Player i enters into 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 $\omega(t) = (z_1(t), \ldots, z_n(t))$. 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 immediately get the following

Corollary 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*

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

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

When the players follow the equilibrium characterized by the Humean strategy, then each fares better than he would fare by offensively violating. This equilibrium punishes exactly those who are guilty, and any guilty players do strictly worse than the innocent. So if one's counterparts follow the Humean strategy, then one had better not offensively violate.

So far, we have assumed that if a given player is guilty, the rest of the community is certain to start a punishment cycle where all boycott the guilty player. 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 q can still characterize an equilibrium of the repeated Covenant Game. This time, for $i \in N$, define $w_i : t \to \{0, 1\}$ 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

$$\begin{aligned} Q(i\,T) = \exists(j \in N)[j = i(t) \wedge \\ ((I(j,\,T) \wedge a_i(T) \neq P) \vee (\neg I(j,\,T) \wedge a_i(T) \neq B))] \end{aligned}$$

and assume that $E_i(1_{A_i(t)}|Q(j,\,T)) = q$ and $E_i(1_{A_i(t)}|\neg Q(j,\,T)) = 0$ for each $i \in N$.[7] $Q(i,\,T)$ 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 $\omega(t) = (w_1(t), \ldots, w_n(t))$. 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*

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

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

[7] 1_A is the indicator function of the proposition or event A, that is,

$$1_A = \begin{cases} 1 & \text{is } A \text{ obtains (or } A \text{ occurs)} \\ 0 & \text{otherwise} \end{cases}.$$

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 $i(t)$ that Player i exploited 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 Corollary 4. At the other extreme, 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.

Equilibria defined by Humean strategies of conditional cooperation exist for the repeated Covenant game. However, we must not slide to the conclusion that players in the repeated Covenant game will follow one of these Humean equilibria. Proposition 1 shows that even if we assume that the players eventually settle into some equilibrium, this equilibrium need not be an equilibrium of conditional cooperation. The players might settle into an equilibrium all follow the pure Humean strategy or a q-stochastic 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 2 shows that the players might even follow an equilibrium where some exploit others by sometimes double-crossing when their counterparts perform.

The conditionally cooperative strategies described in this section crucially assume that all know who bear a guilty "label" and who bear an innocent "label." The players can know this if some reliable mechanism or institution publicly announces or "broadcasts" the identities of guilty players to the entire community. If the broadcasting mechanism never fails to report the identities of the guilty, such a broadcast renders the identities of the guilty and the innocent common knowledge among the community. Then its members can follow an equilibrium where each innocent member performs in covenants exactly when she is matched with an innocent counterpart. This is the lesson to be drawn from Proposition 2 and Corollary 4. If the mechanism fails sometimes, so that sometimes offensive violators "slip through the cracks" and are not identified as guilty, 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. This equilibrium will yield innocent community members a somewhat lower expected payoff than the Humean equilibrium of a fullproof broadcasting mechanism, because sometimes an innocent community member will be exploited and yet no punishment cycle begins. Nevertheless, so long as the probability $1 - q$ of broadcast failure is sufficiently low, an innocent community member still expects to fare better at the equilibrium of the q-stochastic Humean strategy than a Foole can expect to fare. This is the lesson to be drawn from Proposition 5.

Plainly, the cooperation in the Humean-type equilibria can unravel if the players are prone to the sort of mistakes or "trembles" in executing their strategies that are used to characterize equilibrium refinements. For instance, if the players begin at the equilibrium $\boldsymbol{h}$ of the basic Humean strategy and 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. So one might consider the possibility of reputational equilibria that are based upon strategies that are more forgiving than the Humean strategy, and that allow an offensive violator to eventually regain innocence. Kandori (1992) and Vanderschraaf (2005) construct such "forgiving" equilibria. These

equilibria rely upon even more extensive knowledge requirements than the σ_i-trigger and the Humean equilibria described here. In a "forgiving" equilibrium community members must know not only who are guilty, but when they became guilty, so that they will know when to stop boycotting. Cooperation in a community of Humeans can also unravel 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. Millgrom, North, and Weingast (1990) take this 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 guilt or innocence. 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 Millgrom–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, except that it is the threat of being shunned following a guilty judgment rather than the coercive power the judge may have at his command that does the work of enforcement. 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. More specifically, the community lacks any structures that would enable anyone to make public announcements, the immediate generators of common knowledge. A multitude of people in their "naturall condition" of political freedom as Hobbes describes them would be such a community (*Leviathan* 13:9).[8] 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 make public announcements, but perhaps they can at least spread information via informal private 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."

[8] Of course, Hobbes famously goes farther and claims that the people in their naturall condition will inevitably end up in a kind of war with each other. But in the exchange with the Foole in *Leviathan* 15, Hobbes argues that one should not offensively violate covenants on reputational grounds alone. I follow Hampton (1986, pp. 64–66), Kavka (1986, §4.3) and Curley (1994, pp. xxvi–xxviii) in maintaining that Hobbes is considering the possibility that reputation alone can enforce performance in covenants even in our naturall condition.

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 who can communicate only privately analogous to the folk theorems of Sect. 3. But perhaps one can learn something of the properties of such a system by analyzing this system computationally. Gaylord and D'Andria (1998) present an early computational analysis of the spread of bad reputation. While I believe that their specific model is too crude to give much insight into how behavior in repeated Covenant Games might evolve in real human communities, they are pioneers in the use of computational models to analyze interactions in a society. Here is a description of the Gaylord–D'Andria model, which I will modify later in this section: 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 at random a direction, north, south, east or west, 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 4 depicts such a lattice where $r = 50$, so that there are 2,500 cells in total.

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

Fig. 4 50×50 Lattice of players who play the repeated the Covenant Game with random matching

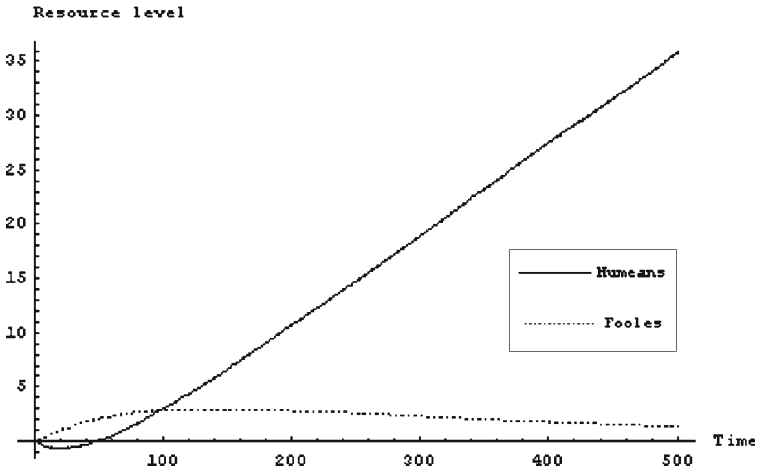

Fig. 5 Average accumulated payoffs of Humeans and naive Fooles over 500 stages of play, $g = c = 1, l = 2$

and boycott counterparts 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 5 shows the average aggregate payoffs of the Humeans and the average aggregate payoffs of the Fooles over a run of 500 stages of plays.

The results in this figure replicate Gaylord and D'Andria's findings in their computational study (1998). They appear to give 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 private 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 in terms of average aggregate payoff.

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 their model, only the identities of actual offensive violations are ever added to a player's blacklist. In effect, all agents in the system 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 become unwilling to interact with those innocent victims the Foole has added to his

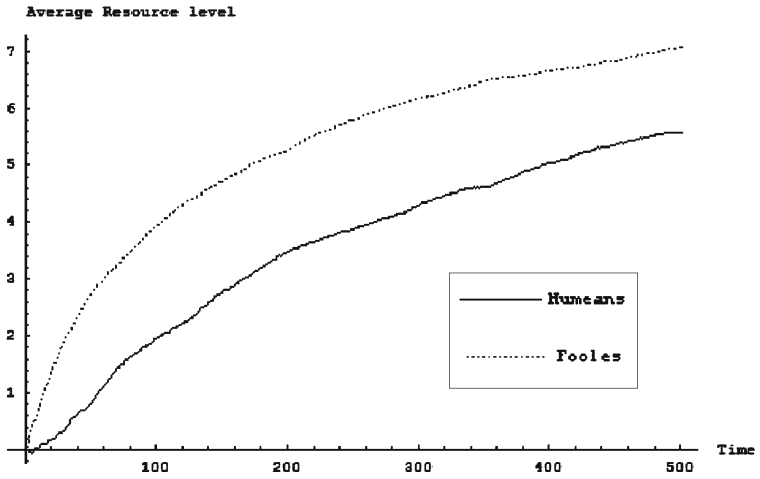

Fig. 6 Average accumulated payoffs of Humeans and more sophisticated Fooles over 500 stages of play, $g = c = 1, l = 2$

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 6 summarizes the results of a run of 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.

As in the previous simulation, here 70% of the lattice is occupied with agents. Again half the agents are Humeans and half are Fooles. The Humeans follow the same strategy as before. But now Fooles double-cross counterparts not on their blacklists at random with probability $\frac{1}{3}$. 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. Figure 6 plots the average aggregate payoffs of the Humeans and the average aggregate payoffs of the Fooles over time. As Fig. 6 shows, now Fooles fare better than Humeans on average over all stages of play. The Fooles have turned the tables on the Humeans merely by adopting a slightly more complex strategy than simply double-crossing any counterpart willing to covenant.

The results of additional computer experiments simulating the interactions of Humeans and Fooles in repeated Covenant Games are summarized in Appendix 2. These experiments show that the average prospects for Humeans, "naive" Fooles who always double-cross and always tell the truth, and more sophisticated Fooles who do not always double-cross and who lie about their victims can vary according to some of the parameters of the model such as relative portions of Fooles and Humeans in the population.[9] To be sure, the results of these specific computer experiments barely scratch the surface of what one might learn from running simulations on more

[9] Gaylord and D' Andria present the results of only one computer experiment to test their hypothesis that Humeans will on average fare better than Fooles, the experiment described in this section. Below I show by example that Fooles can do better on average even if they are "naive" and always tell the truth and always double-cross, as Gaylord and D'Andria assume.

complex models of a society that more closely mimic the conditions one would expect to find in an actual human community whose members must rely upon gossip to form their expectations regarding their counterparts. For example, in the models considered here each individual follows exactly one of two strategies all of the time, that is, each is either a Humean or one kind of Foole. More realistic models might allow for the presence of more than two strategies in the population, including the Humean strategy, a forgiving strategy, a naive Foole's strategy and a variety of more complex Foole's strategies. More realistic models might also allow for more complex migration patterns and might also allow some individuals to leave the system or "die" and some newcomers to enter the system. Future work on computer simulations of society may help philosophers and social scientists better understand the specific conditions under which private communication alone can enable Humeans to fare better than Fooles. Still, even the very simple models considered here yield an important lesson: In some of the simplest of communities whose members can communicate only privately, Fooles can fare better on average than Humeans. So Hobbes' and Hume's warnings against offensively violating one's covenants are not compelling in general.

5 Conclusion

At the start of this paper I hinted that the classic argument for keeping promises presented by Hobbes and Hume fails to address 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 provide its members the information they require in order to follow their parts of the equilibrium. Equilibria based upon reputation are viable when 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, 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. On the other hand, such a civil society might not need not enforce covenants with threats of active punishment by the government, as Hobbes supposes. Perhaps all that is required is that the civil society has an institution that disseminates the information that sustains reputational equilibria.

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, or near-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," performing in covenants exactly with those others who reason "correctly" and boycotting 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 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 a government with punishment powers in order to enforce covenants, after all.

However, this argument is too quick. In fact, it does not follow that it is never rational to offensively violate 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 Sect. 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 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: Rationality alone does not explain reciprocal cooperation.[10]

Appendix 1: Proofs of the folk theorems

Proof of Proposition 1 Given $i \in N_R$, we have

$$\begin{aligned} E_i\left(u \circ \boldsymbol{f}\right) &= \sum_{t=1}^{\infty} E_i\left(u_i\left(S\left(\omega\right)\right)\right) p_i^t \\ &= \sum_{t=1}^{\infty} u_R \cdot p_i^t. \end{aligned} \tag{1}$$

[10] Several other authors draw similar lessons in complimentary studies of the social contract (Binmore, 1994, 1998; Skyrms, 1996, 1998; Sugden, 1986). These authors focus their attention on problems of fair division and property acquisition, and conclude that norms of fair division and property rights must be the result of cultural evolution rather than rationality alone.

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

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

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

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

$$\tfrac{u_R}{1-p_i} \geq 1+g \tag{3}$$

and (3) is satisfied when

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

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

Proof of Proposition 2 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. If $x_i = 0$ then Player i is never matched and hence cannot deviate from $\boldsymbol{f}_{\boldsymbol{\sigma}}$. For the remainder of the proof we assume that $x_i > 0$ so that Player i actually plays the Covenant Game at each stage with positive probability. 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

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

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

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

because (i) if Player i deviates from $\boldsymbol{f}_{\boldsymbol{\sigma}}$ 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

either Player i is unmatched and he receives the 0 payoff of working alone or Player i's counterpart $i(t)$ boycotts. By (2) and (3), we also have

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

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

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

or

$$\frac{x_i \Pi_i^*}{1-p_i} \geq 1+g \tag{6}$$

and (6) is satisfied when

$$p_i \geq \frac{1+g-x_i\Pi_i^*}{1+g}.$$

□

Proof of Proposition 3 Let

$$y_{i1}\left(T\right) = \sum_{j\neq i}\mu_i\left[i\left(t\right)=j \wedge C\left(j,\, T\right)\right], \text{ and}$$
$$y_{i2}\left(T\right) = \sum_{j\neq i}\mu_i\left[i\left(t\right)=j \wedge \neg C\left(j,\, T\right)\right].$$

By hypothesis, $y_{i1}(T)+y_{i2}(T)=x_i$. Following the notation in the proof of Proposition 2, let f_i' be any deviation from f_{σ_i} and let T_0 be the first stage where $f_i' \neq f_{\sigma_i}$. Let $\boldsymbol{f}_{-i}^{\star}$ denote the profile of strategies Player i's counterparts follow. Then

$$\begin{aligned} E_i\left(u\left(f_{\sigma i}, \boldsymbol{f}_{-i}^{\star}\right)\right) &= \sum_{t=1}^{\infty}\left(\sum_{j\neq i}\Pi\left(i,\, j\right)\mu_i\left[i\left(t\right)=j \wedge C\left(j,\, T\right)\right]\right)\cdot p_i^t \\ &\geq \sum_{t=1}^{T_0-1}\left(\sum_{j\neq i}\Pi\left(i,\, j\right)\mu_i\left[i\left(t\right)=j \wedge C\left(j,\, T\right)\right]\right)\cdot p_i^t + \sum_{t=T_0}^{\infty} y_{i1}\left(T\right)\Pi_i^* p_i^t \end{aligned}$$

because Player i gets the positive expected payoff of following his part of a mixed strategy profile in a single Covenant Game with the conformists he meets and gets the 0 payoff of boycotting the noncomformists he meets or being unmatched. On the other hand,

$$\begin{aligned} E_i\left(u\left(f_i', \boldsymbol{f}_{-i}^{\star}\right)\right) &= \sum_{t=1}^{\infty}\left(\sum_{j\neq i}\Pi\left(i,\, j\right)\mu_i\left[i\left(t\right)=j \wedge C\left(j,\, T\right)\right]\right)\cdot p_i^t \\ &\geq \sum_{t=1}^{T_0-1}\left(\sum_{j\neq i}\Pi\left(i,\, j\right)\mu_i\left[i\left(t\right)=j \wedge C\left(j,\, T\right)\right]\right)\cdot p_i^t + \sum_{t=T_0}^{\infty} y_{i2}\left(T\right)\left(1+g\right) p_i^t \end{aligned}$$

because for each stage $t > T_0$, the conformists boycott Player i so the best Player i can do is to exploit every nonconformist he meets, assuming each of these nonconformists follows P and Player i then follows D. So f_{σ_i} is Player i's best response if

$$\sum_{t=T_0}^{\infty} y_{i1}(T)\,\Pi_i^* p_i^t > \sum_{t=T_0}^{\infty} y_{i2}(T)\,(1+g)\,p_i^t$$

or

$$y_{i1}(T)\,\Pi_i^* > y_{i2}(T)\,(1+g) = (x_i - y_{i1}(T))\,(1+g)$$

and solving for $y_{i1}(T)$ we get

$$y_{i1}(T) > \frac{x_i\,(1+g)}{1-g+\Pi_i^*}. \qquad \square$$

Proof of Corollary 4 The Humean strategy is the σ_i-trigger strategy with $\alpha_i = 1$ for every $i \in N$. Hence if each player follows h, $\Pi_i^* = 1$ and the equilibrium condition (1) follows from the proof of Proposition 2. $\square$

Proof of Proposition 5 As before, let $x_i = \sum_{j \neq i} u_i[i(t) = j]$, so that

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

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

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

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

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

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

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

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

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

and (2) is satisfied when

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

Appendix 2: Additional computer experiments

As noted in Sect. 4, the computer simulation summarized by Fig. 5 replicates the simulation reviewed in Gaylord and D'Andria (1998). Gaylord and D'Andria find that, given the parameters they input into their model, Humeans consistently do better on average than to naive Fooles. The Humeans also accumulate a steadily increasing aggregate average payoff, while the Fooles' aggregate average payoff gradually declines. The computer simulation summarized by Fig. 6 uses exactly the same parameters Gaylord and D'Andria use in their simulation, and runs the system over 500 stages of play, same as in Gaylord's and D'Andria's simulation. Again, the only difference is that the Fooles now double-cross at random $\frac{1}{3}$ of the time and include the identities of their victims on their blacklists, so that their victims end up on other Humeans' blacklists even though they are innocent. In this simulation, it is the Fooles who consistently do better on average than the Humeans. Both the Humeans' and the Fooles' aggregate average payoffs slowly increase over the 500 stages, but at every stage the aggregate average payoffs of the Fooles is strictly greater that of the Humeans.

This appendix summarizes some additional computer experiments. The simulator programs were coded and run in Mathematica 5.2 and extend the simulator program that Gaylord and D'Andria made publicly available prior to publishing their study *Simulating Society* (1998). In each of these additional experiments, individuals in the population again migrate in a 50×50 torus according to the rules described in Sect. 4, but this time do so over 1,000 cycles of play in order to better detect long term trends in the population. The Covenant Game payoffs are set by $g = c = 1$ and $l = 2$.

Figure A1 summarizes a run of 1,000 stages where the Fooles always double-cross and never lie, as was the case in Gaylord and D'Andria's original model. However, this time the ratio of Fooles to Humeans is no longer equal. Specifically, here 95% of the population are Fooles. Note that even over 1,000 stages, the average aggregate payoff of the Humeans is still lower than of the Fooles. So the Humean strategy is not necessarily better than the naive Foole's strategy even in Gaylord's and D'Andrea's original model. This lends support to a rebuttal the Foole might have raised against Hobbes' original argument against offensive violation: If most in the population are Fooles, and not Humeans as Hobbes and Hume apparently suppose, then the Foole's strategy can be the better strategy after all.

Figures A2–A10 summarize experiments that simulate populations where some in the population are Humeans and others are Humeans who offensively violate only some of the time and who spread false information about those they exploit to those they do not exploit by adding the identities of their victims to the blacklists they share with others. These experiments compare the relative effectiveness of the Humean and

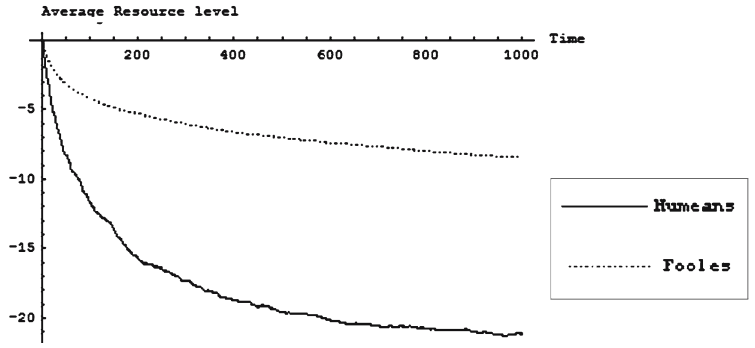

Fig. A1 Average accumulated payoffs of Humeans and naive Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 95, Defection Rate 1

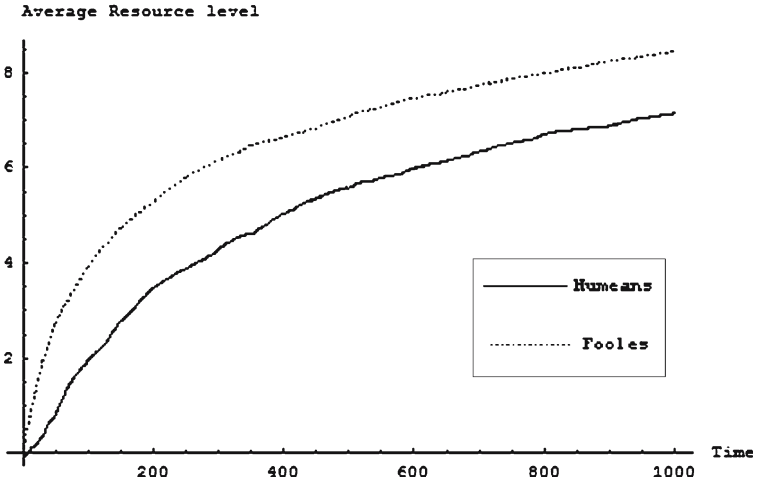

Fig. A2 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 50, Defection Rate $\frac{1}{3}$

the Foole's strategies in terms of aggregate average payoff. Figures A2–A4 summarize experiments where Humeans and Fooles who offensively violate $\frac{1}{3}$ of the time are equally represented and the population density varies. In each case, the Fooles always do better over all 1000 stages than do the Humeans, although it appears that the Fooles' relative advantage in terms of average aggregate payoff increases as the population density lowers. In the experiments Figs. A5–A7 summarize, Humeans and Fooles are again equally represented but now the rates of offensive violation of the more complex Foole's strategy vary. As the rates of offensive violation increase, the relative advantage of the Foole's strategy decreases. Figure A7 shows that if the Fooles double-cross $\frac{3}{4}$ of the time, then by the 600th stage they no longer do better on average than do the Humeans, although their average aggregate payoff does not lag far from that of the Humeans even at the 1,000th stage. Evidently if Fooles double-cross too frequently, then their strategy too closely approximates that of the naive Foole

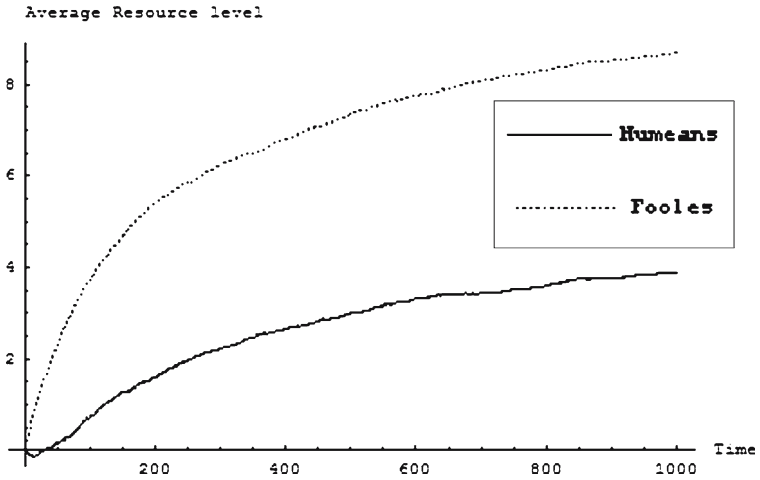

Fig. A3 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.5, Percentage of Fooles 50, Defection Rate $\frac{1}{3}$

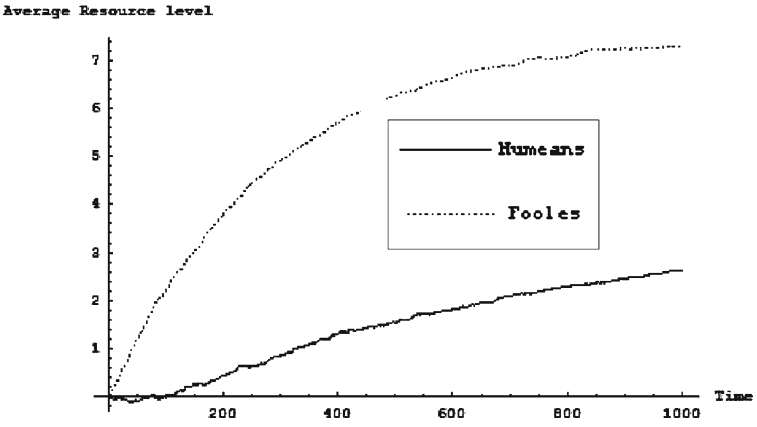

Fig. A4 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.3, Percentage of Fooles 50, Defection Rate $\frac{1}{3}$

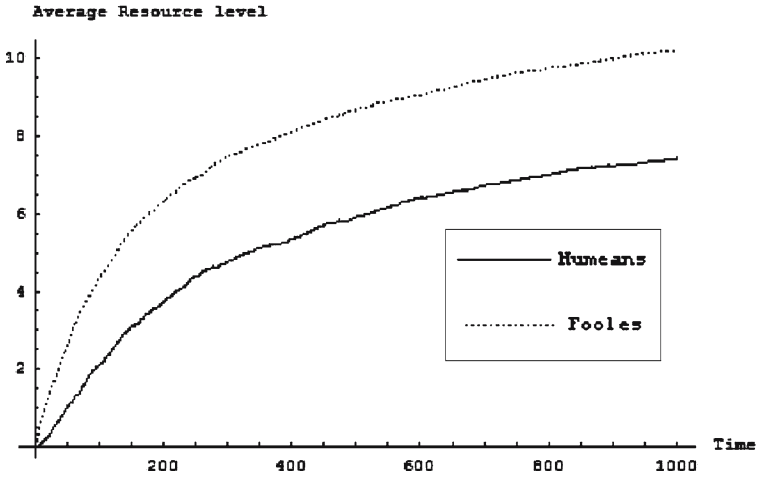

Fig. A5 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.5, Percentage of Fooles 50, Defection Rate $\frac{1}{2}$

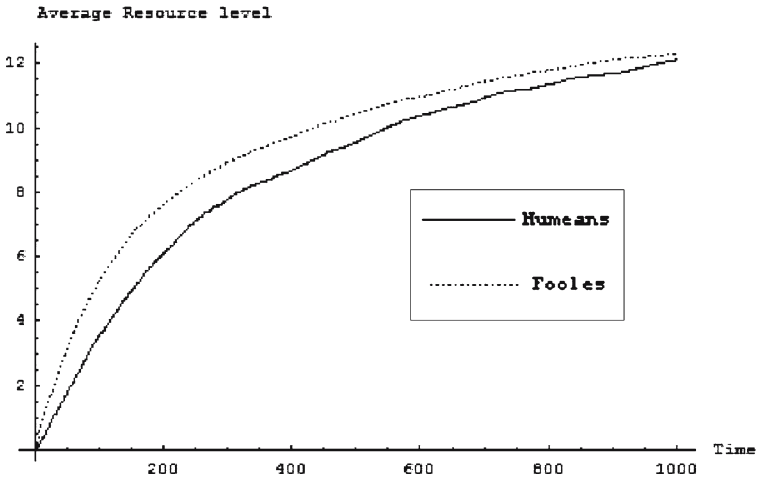

Fig. A6 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 50, Defection Rate $\frac{2}{3}$

who always double-crosses and they lose the advantage of the more complex Foole's strategy. In the experiments Figs. A8 and A9 summarize, the Fooles offensively violate $\frac{1}{3}$ of the time once more, but now the percentages of Fooles in the population vary. Figure A8 summarizes an experiment where $\frac{3}{4}$ of the population are Humeans. In the beginning, the Fooles do better on average but gradually the Humeans' average aggregate payoff overtakes that of the Fooles. This finding is a confirming instance Hobbes' and Hume's claim that following the Foole's advice is against one's self-interest, on condition that most of the population follow the Humean strategy. Figure A9 summarizes an experiment where now $\frac{3}{4}$ of the population are Fooles. Now the Humeans not only always do worse on average than do the Fooles, but the average aggregate payoff of the Humeans is always negative and steadily decreases while that

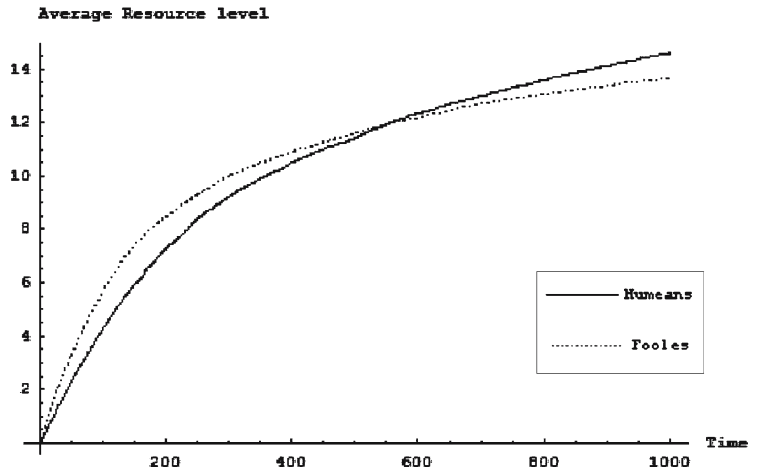

Fig. A7 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 50, Defection Rate $\frac{3}{4}$

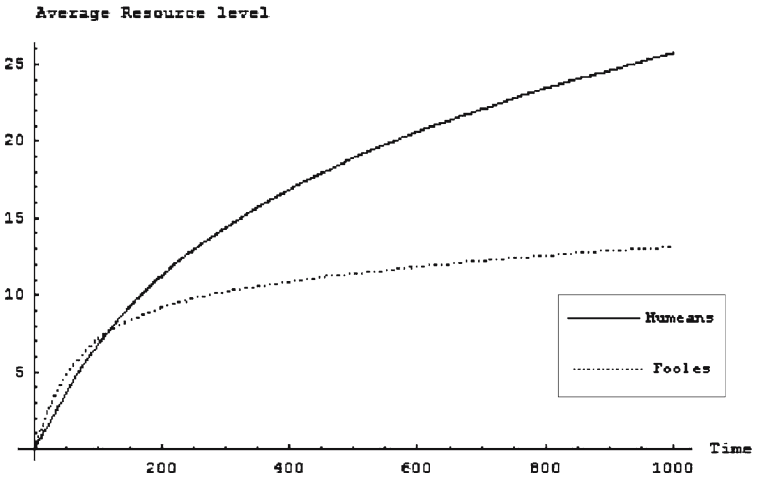

Fig. A8 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 25, Defection Rate $\frac{1}{3}$

of the Fooles is always positive and steadily increases. This suggests that the Foole's reasoning is right when most of the population follow a strategy similar to the strategy he endorses, rather than the strategy Hobbes and Hume endorse. Finally, Fig. A10 summarizes an experiment where the Fooles double-cross $\frac{3}{4}$ of the time and form $\frac{9}{10}$ of the population. Unlike Fig. A7 the case where the Humeans and the Fooles are equally represented, here the Fooles always do better on average than the Humeans, and their relative advantage grows over time. This final experiment suggests that there is no uniquely "best" complex Foole's strategy, since double-crossing $\frac{3}{4}$ of the time "beats" the Humean strategy in terms of average aggregate payoff when $\frac{9}{10}$ of the population follow this strategy but "loses" when only $\frac{1}{2}$ the population follow this strategy. Not surprisingly, the best strategy to adopt, in terms of maximizing average

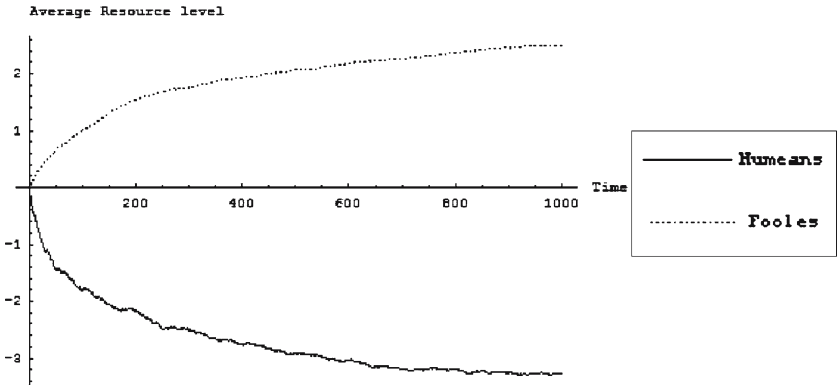

Fig. A9 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 75, Defection Rate $\frac{1}{3}$

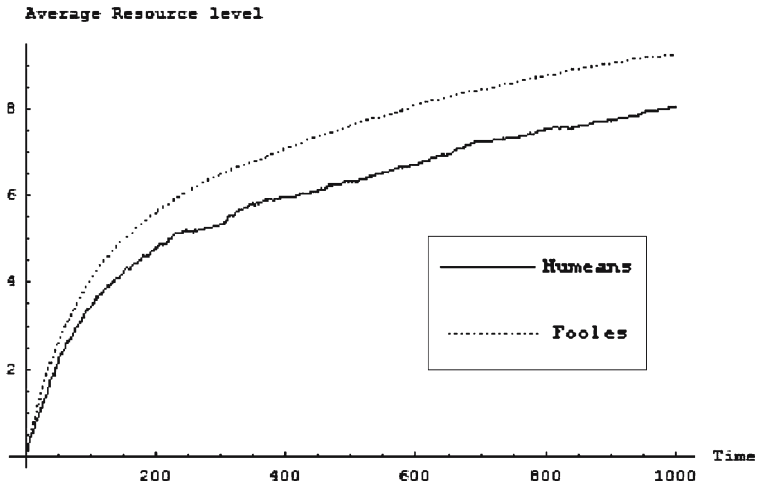

Fig. A10 Average accumulated payoffs of Humeans and complex Fooles over 1,000 stages of play. Population Density 0.7, Percentage of Fooles 90, Defection Rate $\frac{3}{4}$

aggregate payoff, will vary according to the initial conditions of the population. This underscores the claim in Sect. 5 that the best strategy to adopt in repeated covenant situations depends upon the contingent circumstances of the population.

References

Aumann, R. (1974). Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics, 1*, 67–96.
Aumann, R. (1976). Agreeing to disagree. *Annals of Statistics, 4*, 1236–1239.
Aumann, R. (1987). Correlated equilibrium as an expression of Bayesian rationality. *Econometrica, 55*, 1–18.
Axelrod, R. (1981). The emergence of cooperation among egoists. *American Political Science Review, 75*, 306–318.
Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books, Inc.

Batali, J., & Kitcher, P. (1995). Evolution of altruism in optional and compulsory games. *Journal of Theoretical Biology, 175*, 161–171.
Binmore, K. (1994). *Game theory and the social contract volume I: Playing fair*. Cambridge, Massachusetts: MIT Press.
Binmore, K. (1998). *Game theory and the social contract volume II: Just playing*. Cambridge, Massachusetts: MIT Press.
Curley, E. (1994). Introduction to Hobbes' *Leviathan*. In E. Curley (Ed.), *Leviathan* by Thomas Hobbes (pp. viii–xlvii). Indianapolis: Hackett Publishing Company.
Dekel, E., & Gul, F. (1997). Rationality and knowledge in game theory. In D. M. Kreps & K. F. Wallis (Eds.), *Advances in economics and econometrics: Theory and Applications* (pp. 87–172). Cambridge: Cambridge University Press.
Ellison, G. (1994). Cooperation in the prisoner's dilemma with anonymous matching. *Review of Economic Studies, 61*, 567–588.
Flood, M. M. (1958). Some experimental games. *Management Science, 5*, 5–26.
Gaylord, R. J., & D'Andria, L. J. (1998). *Simulating society: A mathematica toolkit for modeling socioeconomic behavior*. New York: Springer Verlag.
Hampton, J. (1986). *Hobbes and the social contract tradition*. Cambridge: Cambridge University Press.
Hill, D. E. (2004). Errors of judgment and reporting in a law merchant system. *Theory and Decision, 56*, 239–268.
Hobbes, T. (1651) (1991). *Leviathan*, ed. Richard Tuck. Cambridge: Cambridge University Press.
Hume, D. (1740) (2000). *A treatise of human nature*, ed. David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.
Kandori, M. (1992). Social norms and community enforcement. *Review of Economic Studies, 59*, 63–80.
Kavka, G. (1986). *Hobbesian moral and political theory*. Princeton: Princeton University Press.
Kitcher, P. (1993). The evolution of human altruism. *The Journal of Philosophy, 90*, 497–516.
Lewis, D. (1969). *Convention: A philosophical study*. Cambridge, Massachusetts: Harvard University Press.
Linster, B. (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., & 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* (pp. 243–266). Ann Arbor, Michigan: University of Michigan Press.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge: Cambridge University Press.
Skyrms, B. (1998). The shadow of the future. In J. Coleman & C. Morris (Eds.), *Rational commitment and social justice: Essays for Gregory Kavka* (pp. 12–22). Cambridge: Cambridge University Press.
Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Oxford: Basil Blackwell, Inc.
Vanderschraaf, P. (2005). Reputational enforcement of covenants. Technical Report No. CMU-PHIL-167.