Modeling Cooperation with Self-regarding Agents

Herbert Gintis*

June 19, 2007

Abstract

The n-player public goods game, the basic model of decentralized social cooperation in non-market settings, has a unique Nash equilibrium in which all players defect. The Folk Theorem asserts that near-Pareto-optimal payoffs can be supported if the game is indefinitely repeated and the discount factor is sufficiently near unity. This paper advances the view that the Folk Theorem does not explain why or how individuals cooperate in these settings. First, with imperfect public signaling, the Folk Theorem's implication that near-optimal payoffs do not depend on the number of players, given a fixed finite upper limit on the informational processing capacity of players, is based on an impermissible order of taking limits. Second, the Folk theorem demonstrates the existence of an equilibrium, but does not show that this equilibrium can be implemented; i.e., that there is some plausible social process by means of which it can be instantiated. We offer a social implementation mechanism using *focal rules*. This mechanism uses public signals that agents treat as common knowledge, so instances of cooperation based on the repeated public goods game must be publicly observable. The fact that no examples of cooperation based on the n-player public goods game ($n > 2$) with public focal rules has been observed suggests that modeling cooperation with self-regarding agents remains unsolved problem.

1 Introduction

Since Adam Smith (1937[1776]), economists have favored the hypothesis that economic cooperation occurs when self-regarding individuals face incentives that harness selfish motives to the satisfaction of social needs. This idea attained a sophisticated expression through the *Fundamental Theorem of Welfare Economics*, in the hands of Kenneth Arrow, Gérard Debreu and others (Arrow and Debreu 1954).

*Santa Fe Institute and Central European University. I would like to thank Samuel Bowles for careful commentaries on several drafts of this paper, Robert Boyd for helpful suggestions, and the John D. and Catherine T. MacArthur Foundation for financial support.

The general equilibrium model, from which the Fundamental Theorem flows, assumes judicial enforcement, and hence treats the binding nature of contracts as the product of *organizational* rather than *market* forces. What induces self-regarding individuals within the judiciary to enforce private contracts? One could posit that such individuals also have agreed to costlessly enforceable third-party contracts. Aside from the empirical question as to whether such contracts exist and govern judicial behavior, this appears to entail an infinite regress. More important, costless third party enforcement is empirically counterfactual at the level of marketable commodities and factors of production (Gintis 1976, Klein and Leffler 1981, Stiglitz 1987). This situation has led economists to shift attention from *market* models with *contractual* enforcement to *game-theoretic* models with self-enforcement in the form of Nash equilibrium.

Repeated game theory analyzes a stage game $\mathcal{G}$ played an indefinite number of times with a discount factor $\delta \in (0, 1)$. This paper will deal only with the case where $\mathcal{G}$ is an n-player public goods game, although our results can be extended to a far broader set of stage games. In the public goods game, each player has pure strategy set $\{C, D\}$, and by playing C (cooperate), a benefit b is generated that is shared equally by all players, at a cost c to the player himself, where $0 < b/n < c < b$; playing D (defect) generates neither costs nor benefits. We assume there is a noisy public or private signal indicating the cooperation and defection of each player in the previous period, although our results can be extended to the case of private signals.[1]

Let $V \subset \mathbf{R}^n$ be the convex hull of all possible payoffs in $\mathcal{G}$, and let $V^* \subseteq V$ consist of those payoff vectors $v = (v_1, \ldots, v_n)$ in the interior of V^* such that $v_i \geq 0$ for all players i. Suppose each player emits a signal, received by all other players, that indicates defection if the player defected on the previous round, and indicates defection with probability $\epsilon > 0$ even if the player cooperated on the previous round, so with probability $1-\epsilon$ the signal correctly indicates the player cooperated. We assume ϵ satisfies $b^* = b(1-\epsilon) - c > 0$. Then, as we show below, $\mathcal{G}$ conforms to the requirements of Fudenberg, Levine and Maskin (1994), who prove a Folk Theorem asserting that for any point $v^* \in V^*$, there is a $\underline{\delta} < 1$ such that for all $\delta \in (\underline{\delta}, 1)$, v^* represents a payoff vector corresponding to a sequential equilibrium the repeated game based on $\mathcal{G}$ with discount factor δ. Since the Pareto-optimal vector $(b^*, \ldots, b^*)$ is a vertex of V^*, near-Pareto optimality can be so achieved.

Section 2 shows that the dynamic programming approach of Fudenberg et al. (1994), which proves the Folk Theorem for imperfect public signals, applies to

[1] A signal is *public* if the same signal is received by all players, and is otherwise *private*. A signal is *imperfect* if there is a strictly positive probability of indicating that a player defected when he cooperated, or *vice-versa*.

the public goods game. We derive expressions linking the signal error rate ϵ, the number of players n, and the discount factor δ, and show that for any given δ, there is a maximum $n\epsilon$ (order of magnitude unity) that supports cooperation. Moreover, information theoretic arguments suggest that, *ceteris paribus*, the error rate ϵ will be an increasing function of the number of players n. Thus, for any given discount factor, there is a finite maximum number of players that can support an equilibrium with a positive level of cooperation, and the maximum degree of efficiency is a decreasing function of n.

Section 3 suggests that repeated game models and their associated Folk Theorems are merely a first step in understanding social cooperation. Proving the existence of a sequential Nash equilibrium must be followed by developing a mechanism whereby individuals would come to adopt the strategies entailed by the equilibrium or should this occur, why they would persist in doing so. Aumann (1987) showed that the correct solution concept corresponding to the common knowledge of rationality, when the common prior assumption is fulfilled, is the *objectively correlated equilibrium*. Aumann does not speculate as to the nature of the correlating device, but an understanding of human social organization suggests that *social institutions* generally play exactly that role (Durkheim 1967[1902], Weber 1978[1914]). Indeed, the assumption of common priors, widely debated in the game theory literature, can be shown to be justifiable in terms of such a social mechanism. I develop the notion of socially promulgated *focal rules* that implement a sequential equilibrium by serving as a correlating mechanism. Section 7 draws on the this analysis to critique the Folk Theorem for private signals on implementability grounds.

I conclude by noting that my analysis shows that *methodological individualism*, often considered to be a guiding principle of game theory, is incorrect and blocks the application of game theory to many complex social processes. The full potential of game theory to inform social theory will be realized one when interactive epistemology recognizes that humans are social creatures by nature and the character of their evolutionary history, not by reason alone.

2 The Folk Theorem for Imperfect Public Signals

Fudenberg et al. (1994) consider a stage game consisting of players $i = 1, \ldots, n$, each with a finite set of pure strategies $a_{i1}, \ldots, a_{im_i} \in A_i$. A vector $a \in A \equiv \prod_{j=1}^{n} A_i$ is called a pure strategy *profile*. For every profile $a \in A$ there is a probability distribution $y|a$ over the m possible public signals $y \in Y$. Player i's payoff, $r_i(a_i, y)$, depends only on his own strategy choice and the resulting public signal. If $\pi(y|a)$ is the probability of $y \in Y$ given strategy profile $a \in A$, i's expected payoff

from a is given by

$$g_i(a) = \sum_{y \in Y} \pi(y|a) r_i(a_i, y).$$

Mixed strategies and mixed strategy profiles, as well as their payoffs, are defined in the usual way, and denoted by greek letters, so $\alpha \in \Delta A$ is a mixed strategy profile, and $\pi(y|\alpha)$ is the probability distribution generated by mixed strategy α.

In the case of the n-player public goods game, each player can cooperate (C) by producing b/n for each of the n players at a personal cost c, or can defect (D), producing zero benefit at zero cost. Each pure strategy set consists of the two elements {C,D}. We will assume that players choose only pure strategies. It is then convenient to represent the choice of C by 1 and D by 0. Let A be the set of strings of n zeros and ones, representing the possible pure strategy profiles of the n players, the k^{th} entry representing the choice of the k^{th} player. Let $\tau(a)$ be the number of ones in $a \in A$, and write a_i for the ith entry in $a \in A$. For any $a \in A$, the random variable $y \in A$ represents the imperfect public information concerning $a \in A$. We assume defections are signaled correctly, but intended cooperation fails and appears as defection with probability $\epsilon > 0$. Let $\pi(y|a)$ be the probability that signal $y \in A$ is received by players when the actual strategy profile is $a \in A$. Clearly, if $y_i > a_i$ for some i, then $\pi(y|a) = 0$. Otherwise

$$\pi(y|a) = \epsilon^{\tau(a)-\tau(y)}(1-\epsilon)^{\tau(y)} \qquad \text{for } y \leq a.$$

The payoff to player i who chooses a_i and receives signal y is given by $r_i(a_i, y|a) = b\tau(y)(1-\epsilon) - a_i c$. The expected payoff to player i is just

$$g_i(a) = \sum_{y \in A} \pi(y|a) r_i(a_i, y) = b\tau(a)(1-\epsilon) - a_i c.$$

Moving to the repeated game, we assume in each period $t = 0, 1, \ldots$, the stage game is played with public outcome $y^t \in Y$. The sequence $\{y^0, \ldots, y^t\}$ is thus the *public history* of the game through time t, and we assume that the strategy profile $\{\sigma^t\}$ played at time t depends only on this public history (Fudenberg, Levine, and Maskin show that allowing agents to condition their play on their previous private profiles does not add any additional equilibrium payoffs). We call a profile $\{\sigma^t\}$ of public strategies a *perfect public equilibrium* if, for any period t and any public history up to period t, the strategy profile specified for the rest of the game is a Nash equilibrium from that point on. Thus, a public perfect equilibrium is subgame perfect Nash equilibrium implemented by public strategy profiles. The payoff to player i is then the discounted sum of the payoffs from each of the stage games.

The *minimax* payoff for player i is largest payoff i can attain if all the other players collude to choose strategy profiles that minimize i's maximum payoff. In

the public goods game, the minimax payoff is zero for each player, since the worst the other players can do is universally defect, in which case i's best action is to defect himself, giving payoff zero. Let V^* be the convex hull of stage game payoffs that dominate the minimax payoff for each player. A player who intends to cooperate and pays the cost c (which is not seen by the other players) can fail to produce the benefit b (which is seen by the other players) with probability $\epsilon > 0$. In the two-player case, V^* is the quadrilateral ABCD in Figure 1, where $b^* = b(1-\epsilon) - c$ is the expected payoff to a player if everyone cooperates.

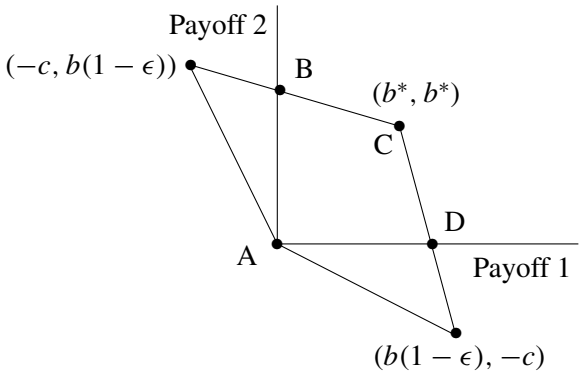

Figure 1: Two-player Public Goods Game

The Folk Theorem (Theorem 6.4, p. 1025 in Fudenberg, Levine, and Maskin, 1994) is then as follows.[2] We say $W \subset V^*$ is *smooth* if W is closed and convex, has a nonempty interior, and such that each boundary point $v \in W$ has a unique tangent hyperplane P_v that varies continuously with v (e.g., a closed ball with center interior to V^*). Then if $W \subset V^*$ is smooth, there is a $\underline{\delta} < 1$ such that for all δ satisfying $\underline{\delta} \leq \delta < 1$, each point in W corresponds to a strict perfect public equilibrium with discount factor δ, in which a pure strategy profile is played in each period. In particular, we can choose W to have a boundary as close as we might desire to $\mathbf{v}^* \equiv (b^*, \ldots b^*)$, in which case the full cooperation payoff can be approximated as closely as desired.

The only condition of the theorem that must be verified in the case of the public goods game is that the full cooperation payoff $\mathbf{v}^* = \{b^*, \ldots, b^*\}$ is on the boundary of an open set of payoffs in $\mathbf{R}^n$, assuming players can used mixed strategies. Suppose player i cooperates with probability x_i, so the payoff to player i is $v_i = \pi_i - cx_i$,

[2] I am suppressing two conditions on the signal y that are either satisfied trivially or irrelevant in the case of the public goods game.

where

$$\pi_i = b \sum_{j=1}^{n} x_j - x_i.$$

If J is the Jacobian of the transformation $x \to v$, it is straightforward to show that

$$\det[J] = (-1)^{n+1}(b-c)\left(\frac{b}{n-1} + c\right)^{n-1},$$

which is non-zero, proving the transformation is not singular.

The method of recursive dynamic programming used to prove this theorem in fact offers an constructive algorithm. Given a set $W \subset V^*$, a discount factor δ, and a strategy profile α, we say α is *enforceable* with respect to W and δ if there is a payoff vector $v \in \mathbf{R}^n$ and a *continuation function* $w : Y \to W$ such that for all i,

$$v_i = (1-\delta)g_i(a_i, \alpha_{-i}) + \delta \sum_{y \in Y} \pi(y|a_i, \alpha_{-i}) w_i(y)$$
$$\text{for all } a_i \text{with } \alpha_i(a_i) > 0, \tag{1}$$
$$v_i \geq (1-\delta)g_i(a_i, \alpha_{-i}) + \delta \sum_{y \in Y} \pi(y|a_i, \alpha_{-i}) w_i(y)$$
$$\text{for all } a_i \text{with } \alpha_i(a_i) = 0. \tag{2}$$

We interpret the continuation function as follows. If signal $y \in Y$ is observed (the same signal will be observed by all, by assumption), each player switches to a strategy profile in the repeated game that gives player i the long-run average payoff $w_i(y)$. We thus say that $\{w(y)_{y \in Y}\}$ *enforces* α with respect to v and δ, and that the payoff v is *decomposable* with respect to α, W, and δ. To render this interpretation valid, it must be shown that $W \subseteq E(\delta)$, where $E(\delta)$ is the set of average payoff vectors that correspond to equilibria when the discount factor is δ.

Equations (1) and (2) can be used to construct an equilibrium. First, we can assume that the equations in (1) and (2) are satisfied as equalities. There are then two equations for $|Y| = 2^n$ unknowns $\{w_i(y)\}$ for each player i. To reduce the underdetermination of the equations, we shall seek only pure strategies that are symmetrical in the players, so no player can condition his behavior on having a particular index i. In this case, $w_i(y)$ depends only on whether or not i signaled C (which occurs with probability $1 - \epsilon$ if the player chose C), and the number of other players who signaled C. This reduces the number of continuation strategies for a player from 2^n to $2(n-1)$. In the interests maximizing efficiency, we assume that in the first period all players cooperate, and as long as y indicates universal cooperation, all players continue to play C.

To minimize the amount of punishment meted out in the case of observed defections while satisfying (1) and (2), we first assume that if more than one agent signals defect, all continue to cooperate, but that a single defection is punished an amount that just satisfies the incentive compatibility equations (1) and (2). There is of course no assurance that this will be possible, but if so, there will be a unique punishment p^* such that the observed defector receives payoff $b(1-\epsilon) - p^*$. Each cooperator must then produce at level $1 - p^*/b(1-\epsilon)$, and receives payoff

$$\frac{b(1-\epsilon)}{n-1} + (n-2)\frac{b(1-\epsilon) - p^*}{n-1} - c\frac{b - p^*/(1-\epsilon)}{b}.$$

Comparing this with the payoff to the defector, which is $b(1-\epsilon) - p^* - c$, we find that the difference Δp^* between a cooperator and the defector when there is a single defection is

$$\Delta p^* = \left(\frac{1}{n-1} + \frac{c}{b(1-\epsilon)}\right) p^*.$$

We can now solve (1) and (2) for p^*, getting

$$p^* = \frac{c(1-\delta)}{\delta(1-\epsilon)^{n-2}\left(1 - \left(\frac{n(b-c)+c}{b}\right)\epsilon + (n-1)\epsilon^2\right)}. \tag{3}$$

This corresponds to an expected average per period payoff given by

$$v^* = b^*\left(1 - \frac{(n-1)c(1-\delta)\epsilon}{b(1+(n-1)\epsilon^2) - (n(b-c)+c)\epsilon}\right). \tag{4}$$

Note that, as per the Fudenberg, Levine, and Maskin Theorem 6.4, given n and ϵ, there is a $\delta < 1$ that renders v^* as close to b^* as one desires. However, (4) implies that for any given discount factor $\delta < 1$, as $(n-1)\epsilon \to \frac{b(1-\epsilon)}{b(1-\epsilon)-\delta c}$, $v^* \to 0$. For small ϵ and $b/c > 2$, this implies that $n\epsilon < 2$. More generally the larger the ratio of c to $b(1-\epsilon)$, the larger the maximum group size. The intuition behind this result is that when cost c is close to benefit $b(1-\epsilon)$, cooperators are hurt very little when the punish, but defectors are hurt a lot, since they still must pay the cost c but receive little benefit. When c is much smaller than $b(1-\epsilon)$, inefficiency is high, since cooperators and defectors are almost equally hurt from punishment.

3 Conditions for Nash Equilibrium

The Folk Theorem proves the existence of an open set of equilibrium payoffs corresponding to sequential equilibria of the repeated public good game with imperfect public signals and with discount factors sufficiently close to unity. It is thus an existence theorem that gives no hint as to how rational agents might choose one among

the continuum of equilibria to play, or even why rational agents might choose a Nash equilibrium at all. In the previous section, I constructed a plausible implementation of a sequential equilibrium that approximates the equal-payoff Pareto-efficient solution for sufficiently patient players, given a fixed group size n and error rate ϵ. However, there are infinitely many other sequential equilibria implementing alternative payoffs near a point on the Pareto frontier of V^*, each favoring some players at the expense of others. No special privilege can reasonably be accorded to the equal-payoff equilibrium in a real-world situation where players are socially heterogeneous along various dimensions.

In some very simple games, there are conditions under which rational agents can "learn" to play a Nash equilibrium (Fudenberg and Levine 1997, Young 2006). These conditions do not obtain for repeated games, however. The strongest arguments in favor of the assertion that a Nash equilibrium will be played come from evolutionary game theory, where it is shown that every stable equilibrium of a dynamical system governed by a monotone dynamic, such as the replicator dynamic (Taylor and Jonker 1978), is a Nash equilibrium of the underlying game (Nachbar 1990, Samuelson and Zhang 1992). However, if there is a continuum of equilibria implementing an $n-1$-dimensional surface of Pareto-efficient equilibria, it is unclear even in principal how an monotone evolutionary dynamic implementing a particular equilibrium might be constructed, much less how realistic such a dynamic might be.

Research in interactive epistemology suggests that the conditions for achieving Nash equilibrium are quite stringent and rarely satisfied, except in the simplest of cases. The problem with achieving a Nash equilibrium is that rational agents may have heterogeneous and incompatible beliefs concerning how other players will behave, and indeed what other players believe concerning one's own behavior. Aumann and Brandenburger (1995) develop a set of conditions for Nash equilibrium that are sufficient and strict—i.e., with a violation of any one, a non-Nash equilibrium counterexample can be found. Consider a game $\mathcal{G}$ with players $i = 1, \ldots, n$, $n > 2$, a finite pure strategy set A_i for each player i, and a payoff function $g_i : A \to \mathbf{R}$ for each player i. A *belief system* for a game consists of

a. for each player i, a set S_i (i's *types)*; and for each type $s_i \in S_i$,

b. a *probability distribution* $p_i(\cdot; s_i)$ on $S = S_1 \times \ldots \times S_n$, and

c. an action $a_i \in A_i$.

Note that S is the set of possible configurations of players that might obtain at the start of the game. We call $s \in S$ a *state of the world*. For any player i, $p_i(\cdot; s_i)$ represents the players subjective probability concerning which state of the world

actually obtains (we assume players know their own types, so $p_i(s; s_i) = 0$ if s_i is not the i^{th} entry in s).

The heart of a belief system is the probability distribution $p_i(\cdot; s_i)$. Since the *type* of the individual includes his beliefs, each type is distinguished by the beliefs its members hold concerning the beliefs of other members. Carrying this one step further, each type includes beliefs that each other type has concerning the beliefs each holds concerning the beliefs of other types. This nested sequence of beliefs concerning beliefs concerning beliefs concerning beliefs continues *ad infinitum*.

A *conjecture* ϕ^i of player i is a probability distribution on $A_{-i} = A_1 \times \ldots A_{i-1} \times A_{i+1} \times \ldots \times A_n$. Player i's belief system $p_i(\cdot; s_i)$ induces a probability distribution over the actions of the other players' actions as follows. Given the player's type s_i, let E be the set of states of the world $s = (s_{-i}, s_i) \in S$ with associated actions $a = (a_1, \ldots, a_n) \in A$. Define $\phi^i(a_{-i}) = p_i(E; s_i)$. We say an individual i is *rational* if his action a_i maximizes his expected payoff $\mathbf{E}[g_i(a)]$, given the conjecture $\phi^i(a_{-i})$ induced by his beliefs $p_i(\cdot; s_i)$. Finally, we say a player *knows* something if he believes it occurs with probability one.

We say players have a *common prior* $P(\cdot)$ over states of the world such that, for each $i = 1, \ldots, n$, and for every subset $E \subset S$, $p_i(E, s_i) = P(E)$. Aumann and Brandenburger (1995) then show:

Theorem 1. *Let $\phi = (\phi^1, \ldots, \phi^n)$ be a set of conjectures for $\mathcal{G}$. Suppose the players have a common prior, it is mutually known that the game is $\mathcal{G}$ and all players are rational, and it is commonly known that ϕ is the set of conjectures for the game. Then $(\sigma_1, \ldots, \sigma_n)$ is a Nash equilibrium of $\mathcal{G}$.*

This theorem suggests that the epistemological requirements for Nash equilibrium in all but the simplest games cannot be deduced from the assumption of rationality alone. This is because when there are multiple Nash equilibria, even the assumption that other players will choose a Nash strategy (an assumption that is itself difficult to justify) is insufficient to ensure common knowledge of conjectures. Rather, there must be a social process lying outside the sphere of individual Bayesian rationality leading to the alignment of conjectures, the transformation of mutual into common knowledge, and the constitution of common priors.

4 Focal Rules and Correlated Equilibria

The idea that rational agents will play a particular Nash equilibrium of a repeated game is based on the informal argument that the players can get together beforehand and come to an agreement as to how each agent is to perform under each possible contingency. Such a justification is uncompelling and indeed bizarre. If there are

multiple equilibria (and by the Folk Theorem, there generally will be for a wide class of repeated games), how do players adjudicate among them? Moreover, if individuals are to "come to an agreement," what are the institutional constraints and strategic interactions that underlie the negotiation process? Indeed, why should they agree at all, and if there is migration, reproduction, error, mutation, and the like, how can an informal agreement hold up? These questions have no general answers.

This argument for Nash equilibrium is traditionally left vague, no doubt because if more structure were attributed to the adjudication process, then that structure would have to be included in the rules of the supergame, which then would become an adjudication game $\mathcal{A}$ followed by the repeated stage game $\mathcal{G}$. But, of course this just pushes the problem of justifying the Nash equilibrium to the preliminary game $\mathcal{A}$. The methodological individualism of game theory, which requires that supra-individual social phenomena be explicable in terms of the behaviors of individual agents, precludes "anchoring" $\mathcal{G}$ in some pre-existing macro-social reality, of which focal rules are an example.[3]

Let Ω be a finite set of public signals, and let p be a probability distribution over Ω. Given the stage game $\mathcal{G}$, we consider a supergame in which Nature first chooses a signal $\omega \in \Omega$, ω is observed by all players, following which each player i choses a strategy $a_i(\omega) \in A_i$ and receives payoff $\pi_i(a(\omega)) = \pi_i(a_1(\omega), \ldots, a_n(\omega))$. The Nash equilibria of this supergame are called the *correlated equilibria* of $\mathcal{G}$ (Aumann 1974, 1987). Conditions for a correlated equilibrium are thus

$$\pi_i(a_i(\omega), a_{-i}(\omega)) \geq \pi_i(a_i', a_{-i}(\omega)) \tag{5}$$

for all $i = 1, \ldots, n$ and all $a_i' \in A_i$. From this is clear that every convex combination of Nash equilibria of $\mathcal{G}$ is a correlated equilibrium, and the set of correlated equilibria form a convex polytope.

Suppose $a : \Omega \to A$ is a correlated equilibrium, and define

$$p(a) = \sum_{\omega \in \Omega, a(\omega) = a} p(\omega). \tag{6}$$

We assume $p : \Omega \to A$ is common knowledge, so the induced *correlated equilibrium distribution* $p \in \Delta A$ defined by (6) is also common knowledge. Equation 5 now becomes

$$\sum_{a_{-i} \in A_{-i}} (\pi_i(a) - \pi_i(a_i', a_{-i}))p(a) \geq 0 \tag{7}$$

[3]It is somewhat inconsistent that methodological individualism apparently allows the rules of the game to be defined in terms not reducible to properties of individuals, just as in general equilibrium theory, methodological individualism appears to permit prices and markets to have an existence not reducible to the properties of individuals. However we account for these apparent exceptions, clearly methodological individual tolerates only a strict minimum of such supra-individual entities.

for all $i = 1, \ldots, n$ and all $a_i' \in A_i$, where $a = (a_i, a_{-i})$.

If players have trouble coordinating on a Nash equilibrium of $\mathcal{G}$, it is not clear why they might have less trouble coordinating on a correlated equilibrium. Suppose, however, we change the emphasis and assume a public authority is vested with the task of choosing $\omega \in \Omega$ according to the probability distribution p, and informing each player i the choice $a_i(\omega)$. Since $p \in \Delta A$ is common knowledge, each player can verify that the strategy $a_i \in A_i$ announced by the public authority satisfies (7), and hence maximizes his expected payoff. If $(p \in \Omega, a : \Omega \to A)$ is a correlated equilibrium, we call (p, a) *focal rules* for $\mathcal{G}$. Focal rules act as a coordinating device by specifying how players are both *instructed* and *expected* to behave in every relevant social state $\omega \in \Omega$.

It might be claimed that by positing a public authority, I am simply redefining the problem, which now becomes that of explaining why and how particular public authorities come into being. This is a correct observation, but this in no way invalidates the above argument. The public authority has a *proximate existence* that may or may not be capable of explanation in terms of some *ultimate social dynamic*, such as social evolution (Binmore 1993, 1998, 2005; Parsons, 1964; Bowles 2005). Moreover, we may need to understand the substance of "group rationality" of members of our species, by which I mean our tendency to form common priors and common understandings based on common group membership (Bacharach 2006). Doubtless this aspect of the human psyche, absent in most if not all other animals, derives from our evolutionary history as highly social hunter-gatherers (Boyd and Richerson 2004).

5 Where do Common Priors Come From?

Since Nash (1950), the Nash equilibrium has been considered the appropriate solution concept for rational agents. There is little analytical support for this opinion. Aumann (1987), by contrast, observed that the sort of objective correlated equilibrium developed above is virtually synonymous with Bayesian rationality if knowledge is modeled appropriately. Given a game $\mathcal{G}$, let Ω be a finite set of "states of the world" relevant to the play of the game, including but not necessarily limited to the moves taken by the various players. We represent a player's information as a partition $\mathcal{P}_i$ of Ω such that if $P_i \in \mathcal{P}_i$ then and the current state of the world is $\omega \in \Omega$, then i knows only that $\omega \in P_i$, where $p(P_i) > 0$. We assume all players share a probability distribution p over the states in Ω, so if player i knows the current state is in $P_i \in \mathcal{P}_i$, then $p(\omega) = 0$ for $\omega \notin P_i$, and for $\omega \in P_i$, $p(\omega|P_i) = p(\omega)/p(P_i)$. Let $a_i(\omega) \in A_i$ be the pure strategy chosen by player i in state $\omega \in \Omega$. We say i is Bayesian rational if, for each $\omega \in \Omega$, $a_i(\omega)$ maximizes i's payoff conditional on

$\omega \in P_i$. Aumann observes that if each player is Bayesian rational, then $a : \Omega \to A$ is a correlated equilibrium, using the players' common prior $p \in \Delta A$.

In the interactive epistemological tradition to which Aumann (1987) is a contribution, it is generally argued that if the partition structure $\{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$ and the subjective priors $p_1, \ldots, p_n$ are not common knowledge, then the game is misspecified. This argument is plausible and widely accepted. However, the assumption of common priors, which means $p_1 = \ldots = p_n$, has generated some controversy. Aumann's position in his 1987 paper is not that the assumption is justified on rationality grounds, but rather it a reasonable assumption and it is commonly assumed in the vast majority of economic models. This view is disputed by Gul (1998), among others, who observes that the assumption of common priors is reasonable only if the model explicitly includes a state prior to game play in which all agents are symmetrically situated and possess the same information. Our use of common priors in this paper is not subject to this criticism because we take the probability distribution $p \in \Omega$ as socially constituted as part of the construction of a set of focal rules. Players accept a common prior for the same reason they accept as common knowledge that the focal rules form a correlated equilibrium: it is a social custom to do so. To illustrate this point, we offer the following example.

		Bob a	Bob b
Alice	a	2,1	0,0
	b	0,0	1,2

Figure 2: Battle of the Sexes

Consider the two-player game shown in Figure 2, known as the Battle of the Sexes. Alice and Bob must meet to conclude a business deal, but Alice prefers to meet at her office (a), while Bob prefers to meet at his office (b). The decisions to meet must be made independently and simultaneously. The payoffs are as shown in the figure. If this is all the information available to us, there is a natural information structure to apply to the problem. Let $\Omega = \{\text{bb}, \text{ba}, \text{ab}, \text{aa}\}$, where bb means both player choose b, ba means Alice chooses b and Bob chooses a, and so forth. Let p be a correlated equilibrium distribution over Ω, so $p(\text{bb}) + p(\text{ba}) + p(\text{ab}) + p(\text{aa}) = 1$.

The set of correlated equilibria must then satisfy

$$(\pi_a(a, b) - \pi_a(b, b))p(\text{ab}) + (\pi_a(a, a) - \pi_a(b, a))p(\text{aa}) \geq 0 \tag{8}$$

$$(\pi_b(b, a) - \pi_b(b, b))p(\text{ba}) + (\pi_b(a, a) - \pi_b(a, b))p(\text{aa}) \geq 0 \tag{9}$$

$$(\pi_a(b,a) - \pi_a(a,a))p(\text{ba}) + (\pi_a(b,b) - \pi_a(a,b))p(\text{bb}) \geq 0 \qquad (10)$$

$$(\pi_b(a,b) - \pi_b(a,a))p(\text{ab}) + (\pi_b(b,b) - \pi_b(b,a))p(\text{bb}) \geq 0, \qquad (11)$$

where $\pi_x(y,z)$ is the payoff to player $x = a, b$ when Alice plays y and Bob plays z. To derive the first equation, note that if Alice is asked to chose a, she knows that Bob was asked to choose a with probability $p_b(a|a) = p(\text{aa})/(p(\text{aa}) + p(\text{ab}))$ and to choose b with probability $p_b(b|a) = p(\text{ab})/(p(\text{aa}) + p(\text{ab}))$. Alice's payoff to choosing a is thus $p_b(a|a)\pi_a(a,a) + p_b(b|a)\pi_a(a,b)$. Subtracting her payoff from choosing b and multiplying out the common denominator, we get the first equation, which says that when Alice is asked to choose a, it is Bayesian rational to do so. The other three equations are derived similarly.

Simplifying these inequalities we get

$$2p(\text{aa}) \geq p(\text{ab}) \qquad (12)$$

$$p(\text{aa}) \geq 2p(\text{ba}) \qquad (13)$$

$$2p(\text{bb}) \geq p(\text{ab}) \qquad (14)$$

$$p(\text{bb}) \geq 2p(\text{ba}). \qquad (15)$$

This convex polytope includes the Nash equilibrium with payoffs (2,1) for $p(\text{aa}) = 1$, the Nash equilibrium with payoffs (1,2) for $p(\text{bb}) = 1$, and the mixed strategy equilibrium when all four are equalities, so $p(\text{aa}) = p(\text{bb}) = 2/9$, $p(\text{ab}) = 4/9$, and $p(\text{ba}) = 1/9$. The payoffs to the players is $(2p(\text{aa}) + p(\text{bb}),\ p(\text{aa}) + 2p(\text{bb}))$, which is Pareto-efficient when $p(\text{ab}) = p(\text{ba}) = 0$.

Our interpretation of the assumption of common priors is illustrated by the Bayesian rationality inequalities (12-15). The only possible argument from rationality is to draw on symmetry to treat all conditions as equalities, in which case we arrive at the mixed strategy equilibrium. Alternatively, we can cite efficiency criteria to set $p(\text{ab}) = p(\text{ba}) = 0$, but there is no reason for the two player to agree on $p(\text{aa})$. If Alice takes $p(\text{aa}) > 1/3$ and Bob takes $p(\text{aa}) < 1/3$, they will make non-equilibrium choices. Common priors, when they exist, are a product of social dynamics, not Bayesian updating.

In a Bob-dominated society, we would expect $p(\text{aa}) < 1/3$, and in a Alice-dominated society, we would expect $p(\text{aa}) > 1/3$. In the repeated game, in a society with a balance of $\alpha/(1-\alpha)$ in favor of Alices, we might expect the focal rule to take the form of emitting a signal $y \in \{\text{a,b}\}$, where $\text{P}[y = a] = \alpha$, such that $y =$x "means" that, for that period, both Alice and Bob are instructed to choose x$\in$a,b. In short, common priors are the expression of a social equilibrium.

This interpretation of common priors sheds light on Aumann's (1975) famous observation that when priors are common knowledge, if posteriors are also common knowledge, then they must be equal. In other words, new information cannot lead

rational agents to disagree. Since there is a huge volume of trade each day in financial securities, this theorem suggests that either there is a large amount of private information in the economy, or priors are not common. On our interpretation of common priors, since there is no organized social process leading to a specification of financial instrument pricing, the absence of common priors does not indicate either a failure of rationality or the presence of massive private information in financial markets.

6 Focal Rules as Choreographer: Enabling Nash Equilibrium

To illustrate the power of focal rules in aligning the choices of rational agents, consider an n-player game in which each player can choose an integer in the range $I = \{1, \ldots, 10\}$. Nature chooses an integer $k \in I$, and if all n players also choose k, each has payoff 1. Otherwise, each has payoff 0. Nature also supplies any agent who inquires (one sample per agent) a noisy signal that equals k with probability p and another integer in i with probability $1 - p$. A best response for each player is to sample the signal and choose a number equal to the signal received. The payoff is p^n. For a correlated equilibrium, suppose the focal rule is for the public authority to sample the noisy signal and specify that each player choose the integer resulting from the sample. The payoff of each player is now p.

This example shows that there may be huge gains to groups that develop focal rules. It is thus not surprising that the coevolution of genes and culture in humans (Cavalli-Sforza and Feldman 1973, Boyd and Richerson 1985, Dunbar 1993) has favored both the evolution of focal rules and of human predispositions to embrace common priors and recognize common knowledge. Nor is this process limited to humans, as the study of territoriality in various nonhuman species makes clear (Gintis 2006).

The methodological individualist tradition in interactive epistemology stresses the importance of common knowledge, but provides not a single instance of a principle of rational cognition that entails common knowledge. This is because accepting information as common knowledge is a human disposition without logical foundation. To see this, suppose Alice and Carol are sipping wine and Bob enters the room, announcing "Dinner will be ready in five minutes." Bob's statement will immediately be common knowledge to Alice and Carol. A more insightful way of describing this situation is to define an event $E \subseteq \Omega$ to be a *public event* if, whenever any $\omega \in E$ occurs, everyone knows it occurs; i.e., ω is *mutual knowledge* (Milgrom 1981). Then, an event is common knowledge if and only if it is a union of public events. Bob's announcement is a serious candidate for being a public event. The empirical question is: when in general do humans consider an event to

be public? An answer to this an related questions will provide a bridge between interactive epistemology and game theory on the one hand, and the theory of the evolution and dynamics of human society on the other.

7 Cooperation with Private Signaling

Repeated game models with private signals, including Sekiguchi (1997), Bhaskar and Obara (2002), Ely and Välimäki (2002), and Piccione (2002), are subject to the critique of the previous sections, but private signaling models are complicated by the fact that no sequential equilibrium can support full cooperation in any period. To see this, consider the first period. If each player uses the full cooperation strategy, then if a player receives a defection signal from another player, with probability one this represents a bad signal rather than an intentional defection. Thus, with very high probability, no other member received a defection signal. Therefore no player will react to a defect signal by defecting, and hence the always defect strategy will have a higher payoff that the always cooperate strategy. To deal with this problem, all players defect with positive probability in the first period. A similar analysis applies to all future periods.

In any Nash equilibrium, the payoff to any two pure strategies that are used with positive probability by a player must have equal payoffs against the equilibrium strategies of the other players. Therefore, the probability of defecting must be chosen so that each player is indifferent between cooperating and defecting on each round. Bhaskar and Obara (2002) accomplish this by showing that there exists a structure of agent beliefs that supports a repeated game sequential equilibrium. The problem with this solution is that there is no guarantee that agents will have these beliefs, and Bhaskar and Obara (2002) offer no reason why they should. Perhaps such beliefs could be promulgated as focal rules, but if this were the case, we should observe complex belief-generating focal rules in groups that manage to cooperate on the basis of private signals. There are no accounts of this in the sociological or anthropological literature.

Ely and Välimäki (2002) have developed a different approach to the problem, following Piccione (2002), who showed how to achieve coordination in a repeated game with private information without the need for belief updating. They construct a sequential equilibrium in which at every stage, each player is indifferent between cooperating and defecting no matter what other players do. Such an individual is thus willing to follow an arbitrary mixed strategy in each period, and the authors show that there exists such a strategy for each player that ensures close to perfect cooperation, provided individuals are sufficiently patient and the errors are small.

The weakness of this approach in explaining real-world cooperation, is one

shared by the mixed strategy Nash equilibria of many games. In a general game, if a player's mixed strategy is part of a Nash equilibrium, then the payoffs to all the pure strategies used with positive probability must be equal. Hence no player has an incentive to calculate and use the mixed strategy at all, since he does equally well by simply choosing among the pure strategies occurring in the support of the mixed strategy in question. If there are costs to computing and randomizing, which is necessarily the case by basic information theory and elementary quantum mechanics, choosing the most convenient pure strategy will be strictly preferred to computing and playing the mixed strategy.

This problem is shared by all mixed strategy equilibria, so is a general problem of classical game theory. There have been two major approaches to dealing with this problem. The first, developed by Harsanyi (1973) treats mixed strategy equilibria as limit cases of slightly perturbed "purified" games with pure strategy equilibria. The second approach, which uses interactive epistemology to define knowledge structures representing subjective degrees of uncertainty, is due to Robert Aumann and his coworkers. Harsanyi's purification theorem does not hold for sequential equilibria of repeated games, however, and the Aumann approach does not predict how agents will actually play, since it determines only the *conjectures* each player has of the *other* players' strategies. These conjectures may imply that I can do no better than play the mixed strategy equilibrium of the game, but since any pure strategy in the support of the equilibrium strategy has equal payoff, it does not dictate anything beyond choosing a strategy from the support of the equilibrium strategy.

Implementing the sequential equilibria solutions to cooperation with private signals is therefore more challenging than in the imperfect public signal case. Focal rules suggest pure, not mixed, strategies, because there is no reason for a rational agent to follow a mixed strategy recommendation, even if a public randomizing device is available (which will very rarely be the case). The only plausible recourse in the case of private signals involves the public authority devising rules for the transformation of private into public information. Candidates include courts, tribunals, and perhaps gossip, although the latter is difficult to implement with self-regarding agents who place no intrinsic value on honesty. Such institutions must also be publicly observable, and their study might well allow the formulation of game-theoretic models of greater explanatory power.

8 Conclusion

The individualist and rationalist methodology historically associated with game theory leads practitioners to be content with solutions in which complexly chore-

ographed sequential equilibria materialize quite without the need for a choreographer. There are no propositions from game theory or anywhere else that suggest that such equilibria have any explanatory value. Methodological individualism, at least of the type common in game theory, is erroneous from a purely scientific viewpoint. There are gains to be made by exploring the joint determination of strategic behavior by social institutions that foster correlated equilibria in the form of focal rules that govern a repeated game based on a stage game $\mathcal{G}$. The inclusion of such institutions in game theory have the added attraction of making game-theoretic predictions more empirically testable, since focal rules are public information for players, rather than being difficult to access subjective beliefs and expectations.

References

Arrow, Kenneth J. and Gerard Debreu, "Existence of an Equilibrium for a Competitive Economy," *Econometrica* 22,3 (1954):265–290.

Aumann, Robert J., "Subjectivity and Correlation in Randomizing Strategies," *Journal of Mathematical Economics* 1 (1974):67–96.

— , "Agreeing to Disagree," *The Annals of Statistics* 4,6 (1975):1236–1239.

— , "Correlated Equilibrium and an Expression of Bayesian Rationality," *Econometrica* 55 (1987):1–18.

— and Adam Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica* 65,5 (September 1995):1161–80.

Bacharach, Michael, *Beyond Individual Choice: Teams and Games in Game Theory* (Princeton, NJ: Princeton University Press, 2006). Natalie Gold and Robert Sugden (eds.).

Bhaskar, V. and Ichiro Obara, "Belief-Based Equilibria the Repeated Prisoner's Dilemma with Private Monitoring," *Journal of Economic Theory* 102 (2002):40–69.

Binmore, Ken, *Game Theory and the Social Contract: Playing Fair* (Cambridge, MA: MIT Press, 1993).

— , *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998).

Binmore, Kenneth G., *Natural Justice* (Oxford: Oxford University Press, 2005).

Bowles, Samuel, *Microeconomics: Behavior, Institutions, and Evolution* (Princeton: Princeton University Press, 2004).

Boyd, Robert and Peter J. Richerson, *Culture and the Evolutionary Process* (Chicago: University of Chicago Press, 1985).

— and — , *The Nature of Cultures* (Chicago, IL: University of Chicago Press, 2004).

Cavalli-Sforza, L. and M. W. Feldman, "Models for Cultural Inheritance: Within Group Variation," *Theoretical Population Biology* 42,4 (1973):42–55.

Dunbar, R. I. M., "Coevolution of Neocortical Size, Group Size and Language in Humans," *Behavioral and Brain Sciences* 16,4 (1993):681–735.

Durkheim, Emile, *De La Division du Travail Social* (Paris: Presses Universitaires de France, 1967[1902]).

Ely, Jeffrey C. and Juuso Välimäki, "A Robust Folk Theorem for the Prisoner's Dilemma," *Journal of Economic Theory* 102 (2002):84–105.

Fudenberg, Drew and David K. Levine, *The Theory of Learning in Games* (Cambridge: The MIT Press, 1997).

— , — , and Eric Maskin, "The Folk Theorem with Imperfect Public Information," *Econometrica* 62 (1994):997–1039.

Gintis, Herbert, "The Nature of the Labor Exchange and the Theory of Capitalist Production," *Review of Radical Political Economics* 8,2 (Summer 1976):36–54.

— , "The Evolution of Private Property," *Journal of Economic Behavior and Organization (*2006).

Gul, Faruk, "A Comment on Aumann's Bayesian View," *Econometrica* 66,4 (1998):923–928.

Harsanyi, John C., "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points," *International Journal of Game Theory* 2 (1973):1–23.

Klein, Benjamin and Keith Leffler, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy* 89 (August 1981):615–641.

Milgrom, Paul, "An Axiomatic Characterization of Common Knowledge," *Econometrica* 49 (1981):219–222.

Nachbar, John H., "Evolutionary' Selection Dynamics in Games: Convergence and Limit Properties," *International Journal of Game Theory* 19 (1990):59–89.

Nash, John F., "Equilibrium Points in n-Person Games," *Proceedings of the National Academy of Sciences* 36 (1950):48–49.

Parsons, Talcott, "Evolutionary Universals in Society," *American Sociological Review* 29,3 (June 1964):339–357.

Piccione, Michele, "The Repeated Prisoner's Dilemma with Imperfect Private Monitoring," *Journal of Economic Theory* 102 (2002):70–83.

Samuelson, Larry and Jianbo Zhang, "Evolutionary Stability in Asymmetric Games," *Journal of Economic Theory* 57,2 (1992):363–391.

Sekiguchi, Tadashi, "Efficiency in Repeated Prisoner's Dilemma with Private Monitoring," *Journal of Economic Theory* 76 (1997):345–361.

Smith, Adam, *The Wealth of Nations* (New York: Modern Library, 1937[1776]).

Stiglitz, Joseph, "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature* 25 (March 1987):1–48.

Taylor, Peter and Leo Jonker, "Evolutionarily Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40 (1978):145–156.

Weber, Max, *Economy and Society* (Berkeley: University of California Press, 1978[1914]). G. Roth and C. Wittich (eds.).

Young, Peyton, *Strategic Learning and its Limits* (Oxford: Oxford University Press, 2006).