Bayesian Rationality, Social Epistemology, and the Choreographer

Herbert Gintis*

September 26, 2009

Abstract

This paper suggests that a social norm is better explained as a *choreographer*—a correlating device with causal effectivity for a correlated equilibrium of an underlying stage game—rather than a Nash equilibrium of the stage game. Whereas the epistemological requirements for rational agents playing Nash equilibria are very stringent and usually implausible, the requirements for a correlated equilibrium amount to the existence of *common priors*, which we interpret as induced by the cultural system of the society in question. In this view, human beings may be modeled as rational agents with special neural circuitry dedicated to reacting to, evaluating, and sustaining social norms.

When the choreographer has at least as much information as the players, we need in addition only to posit that individuals obey the social norm when it is costless to do so. When players have some information that is not available to the choreographer (i.e., not all social roles can be fully incentivized), obedience to the social norm requires that individuals have a predisposition to follow the norm even when it is costly to do so. The latter case explains why social norms are associated with other-regarding preferences and provides a basis for a general analysis of corruption in business and government.

Social norms are thus not explained in terms of game theory and Bayesian rationality, but rather are an *emergent property* of human society, which is a complex adaptive system guided by natural selection. Social norms provide a dimension of causal efficacy to social theory, whereas game theory alone recognizes no causal efficacy above the level of individual choice behavior.

Because of the independent causal effectivity of social norms, the standard methodological individualism of classical game theory is untenable. In particular, social norms are predicated upon certain mental predispositions,

*Santa Fe Institute and Central European University. Some of this material is adapted from my book *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences (Princeton, 2009).*

a *social epistemology*, that is also a product of natural selection. This social epistemology fosters the interpersonal sharing of mental concepts, and justifies the assumption of common priors upon which the identification of Bayesian rationality with correlated equilibrium rests.

Keywords: Nash equilibrium, correlated equilibrium, social norm, social epistemology, correlating device, honesty, corruption, Bayesian rationality.

1 Introduction

The coordination of social life is effected by *social norms* that indicate the appropriate behavior of the individual in particular social roles. There is a long tradition in social theory of treating social norms as Nash equilibria of non-cooperative games played by rational agents (Lewis 1969, Taylor 1976, 1982, 1987, Sugden 1986, 1989, Bicchieri 1992, 1999, 2006, Binmore 1993, 1998, 2005). The insight underlying these contributions is that if agents play a game $\mathcal{G}$ with several Nash equilibria, a social norm can serve to choose among these equilibria, thus coordinating expectations and behavior, as well as possibly avoiding socially inefficient equilibria. While this insight applies to some social conventions, such as driving on the right side of the road, it does not apply to most important cases of human cooperation (e.g., norms concerning proper behavior in particular social roles, such as physician or police officer), and it cannot account for the salience of other-regarding preferences as part of the human behavioral repertoire, because the Nash equilibrium criterion presupposes that agents choose best responses given the game's payoffs. More generally, a social norm takes the form of adding a correlating device, so that the resulting behavior is a correlated equilibrium of the original stage name.

For instance, in the automobile traffic case, correlating devices such as lights and stop signs augment the traffic conventions and effect a correlated equilibrium. Similarly, in the hawk-dove game, adding the so-called "bourgeois" social norm, which says that the player who first occupies a desired location behaves like a hawk and the player who confronts such an incumbent acts like a dove leads to social efficiency. In these cases, we may assume players are self-regarding but share a common prior as to the nature of the correlating device. A third example is the social norm "a police officer must not accept a bribe from a driver in deciding whether to issue a traffic ticket." Assuming the driver's behavior is known only to the driver the police officer in question, the social norm is effective only if the officer is willing to sacrifice personal gain in favor of complying with the norm. The correlating device that maintains the reputation of the police force in this case is the social norm of refusing bribes, compliance with which may be high when each officer considers the norm legitimate, perhaps because the norm is socially

beneficial and others officers appear to comply with the norm.

In this paper I will suggested that a social norm is a *choreographer* of a supergame $\mathcal{G}^+$ of $\mathcal{G}$. By the term 'choreographer' I mean an correlating device that is common knowledge to the players, and that is causally effective in implementing a correlated equilibrium of $\mathcal{G}$ in which all agents play best responses to the choreographer's signals, given their personal (perhaps partially other-regarding) preferences and a common prior.

The social norm as choreographer has three attractive properties lacking in the social norm as Nash equilibrium. First, the conditions under which rational agents play Nash equilibria are generally complex and implausible (e.g., that all players share a common conjecture concerning the strategy choice of each player), whereas rational agents with a common prior canonically play a correlated equilibrium.

Second, the social norms as Nash equilibria approach cannot explain why compliance with social norms is often based on other-regarding moral preferences in which agents choose to sacrifice some personal gain to comply with a social norm. We can explain this association between norms and morality in terms of the incomplete information possessed by the choreographer. Morality, in this view, is doing the right thing even if no one is looking—as long as the cost of so doing is not excessive. Because the motivation to behave morally depends on cost and the general level of compliance, moral behavior may characterize some but not all social equilibria.

Finally, there are many more correlated equilibria than Nash equilibria to most games. Some of these equilibria are Pareto-superior to all Nash equilibria and/or allow a distribution of payoffs among players that is unavailable with Nash equilibria.

2 Epistemic Games

Bayesian rational players have beliefs concerning the behavior of the other players, and they maximize their expected utility by choosing best responses given these beliefs. Thus, to investigate the relationship between Bayesian rationality and strategic behavior we must incorporate beliefs into the description of the game.

An *epistemic game* $\mathcal{G}$ consists of a normal form game with players $i = 1, \ldots, n$ and a finite pure-strategy set S_i for each player i, so $S = \prod_{i=1}^{n} S_i$ is the set of pure-strategy profiles for $\mathcal{G}$, with payoffs $\pi_i : S \rightarrow \mathbf{R}$. In addition, $\mathcal{G}$ includes a set of possible *states* Ω of the game, a *knowledge partition* $\mathcal{P}_i$ of Ω for each player i, and a *subjective prior* $p_i(\cdot;\omega)$ over Ω that is a function of the current state $\omega \in \Omega$. A state ω specifies, possibly among other aspects of the game, the strategy profile s used in the game. We write this $s = \mathbf{s}(\omega)$. Similarly, we write $s_i = \mathbf{s}_i(\omega)$ and

$s_{-i} = \mathbf{s}_{-i}(\omega)$.

The subjective prior $p_i(\cdot;\omega)$ represents i's beliefs concerning the state of the game, including the choices of the other players, when the actual state is ω. Thus, $p_i(\omega';\omega)$ is the probability i places on the current state being ω' when the actual state is ω. We write the cell of the partition $\mathcal{P}_i$ containing state ω as $\mathbf{P}_i\omega$, and we interpret $\mathbf{P}_i\omega \in \mathcal{P}_i$ as the set of states that i considers possible (i.e., that have strictly positive probability) when the actual state is ω. Therefore, we require that $\mathbf{P}_i\omega = \{\omega' \in \Omega | p_i(\omega';\omega) > 0\}$. Because i cannot condition behavior on a particular state in the cell $\mathbf{P}_i\omega$ of the knowledge partition $\mathcal{P}_i$, i's subjective prior must satisfy $p_i(\omega'';\omega) = p_i(\omega'';\omega')$ for all $\omega'' \in \Omega$ and all $\omega' \in \mathbf{P}_i\omega$. Moreover, we assume a player believes the actual state is possible, so $p_i(\omega;\omega) > 0$ for all $\omega \in \Omega$.

If $\psi(\omega)$ is a proposition that is true or false at ω for each $\omega \in \Omega$, we write $[\psi] = \{\omega \in \Omega | \psi(\omega) = \text{true}\}$; i.e., $[\psi]$ is the set of states for which ψ is true.

We call a set $E \subseteq \Omega$ an *event*, and we say that player i *knows* the event E at state ω if $\mathbf{P}_i\omega \subseteq E$; i.e., $\omega' \in E$ for all states ω' that i considers possible at ω. We write $\mathbf{K}_i E$ for the event that i knows E.

Given a possibility operator $\mathbf{P}_i$, we define the *knowledge operator* $\mathbf{K}_i$ by

$$\mathbf{K}_i E = \{\omega | \mathbf{P}_i\omega \subseteq E\}.$$

The most important property of the knowledge operator is $\mathbf{K}_i E \subseteq E$; i.e., if an agent knows an event E in state ω (i.e., $\omega \in \mathbf{K}_i E$), then E is true in state ω (i.e., $\omega \in E$). This follows directly from $\omega \in \mathbf{P}_i\omega$ for all $\omega \in \Omega$. We can recover the possibility operator $\mathbf{P}_i\omega$ for an individual from his knowledge operator $\mathbf{K}_i$, because

$$\mathbf{P}_i\omega = \bigcap \{E | \omega \in \mathbf{K}_i E\}.$$

Since each state ω in epistemic game $\mathcal{G}$ specifies the players' pure strategy choices $\mathbf{s}(\omega) = (\mathbf{s}_1(\omega), \ldots, \mathbf{s}_n(\omega)) \in S$, the players' subjective priors must specify their beliefs $\phi_1^\omega, \ldots, \phi_n^\omega$ concerning the choices of the other players. We have $\phi_i^\omega \in \Delta S_{-i}$, where ΔS is the set of probability distributions over set S, which allows i to assume other players' choices are correlated. This is because, while the other players choose independently, they may have communalities in beliefs that lead them independently to choose objectively correlated strategies.

We call ϕ_i^ω player i's *conjecture* concerning the behavior of the other players at ω. Player i's conjecture is derived from i's *subjective prior* by noting that $[s_{-i}] =_{\text{def}} [\mathbf{s}_{-i}(\omega) = s_{-i}]$ is an event, so we define $\phi_i^\omega(s_{-i}) = p_i([s_{-i}];\omega)$, where $[s_{-i}] \subset \Omega$ is the event that the other players choose strategy profile s_{-i}. Thus, at state ω, each player i takes the action $\mathbf{s}_i(\omega) \in S_i$ and has the subjective prior

probability distribution ϕ_i^ω over S_{-i}. A player i is deemed *Bayesian rational* at ω if $\mathbf{s}_i(\omega)$ maximizes $\pi_i(s_i, \phi_i^\omega)$, where

$$\pi_i(s_i, \phi_i^\omega) =_{\text{def}} \sum_{s_{-i} \in S_{-i}} \phi_i^\omega(s_{-i}) \pi_i(s_i, s_{-i}). \tag{1}$$

In other words, player i is Bayesian rational in epistemic game $\mathcal{G}$ if his pure-strategy choice $\mathbf{s}_i(\omega) \in S_i$ for every state $\omega \in \Omega$ satisfies

$$\pi_i(\mathbf{s}_i(\omega), \phi_i^\omega) \geq \pi_i(s_i, \phi_i^\omega) \qquad \text{for } s_i \in S_i. \tag{2}$$

3 The Epistemic Conditions for Nash Equilibrium

Suppose that rational agents know one another's conjectures in state ω, so that for all i and $j \neq i$, if $\phi_i^\omega(s_{-i}) > 0$ and $s_j \in S_j$ is player j's pure strategy in s_{-i}, then s_j is a best response to his conjecture ϕ_j^ω. We then have a genuine "equilibrium in conjectures," as now no agent has an incentive to change his pure strategy choice s_i, given his conjectures.

We say a Nash equilibrium in conjectures $(\phi_1^\omega, \ldots, \phi_n^\omega)$ occurs at ω if for each player i, $\mathbf{s}_i(\omega)$ is a best response to ϕ_i^ω. We then have the following theorem (Aumann and Brandenburger 1995):

THEOREM 1. *Let $\mathcal{G}$ be an epistemic game with $n > 2$ players, and let $\phi^\omega = \phi_1^\omega, \ldots, \phi_n^\omega$ be a set of conjectures. Suppose the players have a common prior p that assigns positive probability to it being mutually known that the game is $\mathcal{G}$, it is mutually known that all players are rational at $\omega \in \Omega$, and it is commonly known at ω that ϕ^ω is the set of conjectures for the game. Then, for each $j = 1, \ldots, n$, all $i \neq j$ induce the same conjecture $\sigma_j(\omega)$ about j's action, and $(\sigma_1(\omega), \ldots, \sigma_n(\omega))$ form a Nash equilibrium of $\mathcal{G}$.*

These conditions are not necessary, but they are strict; i.e, relaxing any one renders the conclusion untrue. The condition that the conjectures be commonly known is especially stringent, and is highly implausible except possibly for games that are extremely simple, have a unique Nash equilibrium, this equilibrium uses pure strategies, and it can be found by the elimination of strictly dominated strategies (Gintis 2009, Ch. 4). Moreover, unless all conjectures are themselves pure strategies, there is no incentive for any player actually to conform to the conjectured play. This is because all of the mixed strategies that occur in a player's best response have equal payoffs against the conjectures of the other players, so all mixed strategies in the support of the best response are equally good candidates for play. Moreover, because this is common knowledge, no player has grounds for believing the other players will play their conjectures either. It follows that no player

has grounds for playing a best response to the conjectured behavior of the other players.

The concept of a social norm as a Nash equilibrium can in principle render plausible the preconditions of this theorem. The cultural forms supporting the social norm may generate a common prior (nothing in the standard model of Bayesian rationality prohibits an individual's subjective prior being generated by culture and tradition), there may be social cues that lead to common knowledge of the game that is played, and with some effort, we might even envisage a social norms supplying each player with an accurate conjecture concerning the strategy choice of all other players. It remains that self-regarding agents have no incentive to play their conjectures if there are mixed strategies, so the practical applicability of the theorem is restricted to pure strategy equilibria.

The major problem with this attempt to define social norms as a choice of Nash equilibrium, however, is that, as we explain below, the conditions for a correlated equilibrium are much weaker, and there are many more correlated equilibria than there are Nash equilibria of most games. For instance, if there is a perfect public signal indicating who occupied a territory first in the hawk-dove game, we can interpret the bourgeois strategy as a Nash equilibrium of the game by simply including the signal in the current state ω of all players. However, if the signal is private and noisy, there will be many costly hawk-hawk confrontations, so the bourgeois strategy will be inefficient and may fail altogether. Suppose, however, that there is a correlating device that in every confrontation of two players, can access both signals and rule that one or the other contestants is the incumbent. This correlating device produces an efficient correlated equilibrium that is unattainable as a Nash equilibrium.

An additional attraction of the correlated equilibrium is that any convex combination of Nash equilibria is a correlated equilibrium. For instance, consider a Battle of the Sexes in which Bob and Alice get 9 and 1 respectively by playing N, 1 and 9 respectively if both play S, and otherwise get nothing. The only symmetric equilibrium of this game has payoff 0.9 to both players. However, a correlating device that signals N to both players with probability 1/2 and signals S otherwise, leads to an expected payoff of 5 to both players and is a correlated equilibrium.

4 The Epistemic Conditions for Correlated Equilibria

A correlated equilibrium of an epistemic game $\mathcal{G}$ is a Nash equilibrium of a game $\mathcal{G}^+$, which is $\mathcal{G}$ augmented by an initial move by a correlating device, who observes a random variable γ on a probability space (Γ, p) and issues a directive $f_i(\gamma) \in S_i$ to each player i as to which pure strategy to choose. Following the device's

directive is a best response, if other players also follow the devices's directives, provided players have the *common prior* p.

Formally, a *correlated strategy* of epistemic game $\mathcal{G}$ consists of a finite probability space (Γ, p), where $p \in \Delta\Gamma$, and a function $f : \Gamma \to S$. If we think of a correlating device that observes $\gamma \in \Gamma$ and directs players to choose strategy profile $f(\gamma)$, then we can identify a correlated strategy with a probability distribution $\tilde{p} \in \Delta S$, where, for $s \in S$, $\tilde{p}(s) = p([f(\gamma) = s])$ is the probability that the correlating device chooses s. We call $\tilde{p}$ the *distribution* of the correlated strategy. Any probability distribution on S that is the distribution of some correlated strategy f is called a *correlated distribution*.

If f is a correlated strategy, then $\pi_i \circ f$ is a real-valued random variable on (Γ, p) with an expected value $\mathbf{E}_i[\pi_i \circ f]$, the expectation taken with respect to p. We say a function $g_i : \Gamma \to S_i$ is *measurable with respect to* f_i if $f_i(\gamma) = f_i(\gamma')$, then $g_i(\gamma) = g_i(\gamma')$. Clearly, player i can choose to follow $g_i(\gamma)$ when he knows $f_i(\gamma)$ iff g_i is measurable with respect to f_i. We say that a correlated strategy f is a *correlated equilibrium* if for each player i and any $g_i : \Gamma \to S_i$ that is measurable with respect to f_i, we have

$$\mathbf{E}_i[\pi_i \circ f] \geq \mathbf{E}_i[\pi_i \circ (f_{-i}, g_i)].$$

A correlated equilibrium induces a *correlated equilibrium probability distribution* on S, whose weight for any strategy profile $s \in S$ is the probability that s will be chosen by the correlating device. Note that a correlated equilibrium of $\mathcal{G}$ is a Nash equilibrium of the game generated from $\mathcal{G}$ by adding the correlating device, whose move at the beginning of the game is to observe the state of the world $\gamma \in \Gamma$, and to indicate a move $f_i(\gamma)$ for each player i such that no player has an incentive to do other than comply with the devices's recommendation, provided that the other players comply as well. We then have (Aumann 1987):

THEOREM 2. *If the players in epistemic game $\mathcal{G}$ are Bayesian rational at ω, have a common prior $p(\cdot; \omega)$ in state ω, and each player i chooses $\mathbf{s}_i(\omega) \in S_i$ in state ω, then the distribution of $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is a correlated equilibrium distribution given by correlating device f on probability space (Ω, p), where $f(\omega) = \mathbf{s}(\omega)$ for all $\omega \in \Omega$.*

The simplicity and elegance of this characterization of a correlated equilibrium contrasts sharply with the Aumann-Brandenburger characterization of Nash equilibrium in Theorem 1. First the conditions in Theorem 1 are strict sufficient conditions, whereas the conditions in Theorem 2 are necessary and sufficient. Second, no assumptions concerning conjectures are needed for Theorem 2. On the other hand, there are generally many more correlated equilibria than there are Nash

equilibria to a game. Indeed, even a game with a unique Nash equilibrium, such as the hawk-dove game or battle of the sexes has an infinite number of correlated equilibria.

Of course, Theorem 2 shares with Theorem 1 the complete lack of motivation as to why players would have common priors and why they would choose any particular equilibrium over any other. Nevertheless, shifting from a Nash to a correlated equilibrium perspective greatly simplifies the task of providing a plausible nexus of social and psychological forces that give the equilibrium concept its explanatory power. Indeed, Theorem 2 requires only that we explain the origin of common priors or, what amounts to the same thing, we explain the choice of one among a myriad of potential correlated equilibria. I do not propose to provide an explanation in this short paper. Indeed, explaining the actual distribution of correlated equilibria in society is coextensive with explaining social organization and culture are what they are when society is in a social equilibrium.

5 The Choreographer Makes His Appearance

Consider a game played by Alice and Bob, with normal form matrix shown to the right. There are two Pareto-efficient pure-strategy equilibria: (1,5) and (5,1). There is also a mixed-strategy equilibrium with payoffs (2.5,2.5), in which Alice plays u with probability 0.5 and Bob plays l with probability 0.5.

Alice	Bob l	Bob r
u	5,1	0,0
d	4,4	1,5

If the players can jointly observe a correlating device that signals ul and dr, each with probability 1/2, Alice and Bob can then achieve the payoff (3,3) by obeying the correlating device; i.e. by playing (u, l) if they see ul and playing (d, r) if they see dr. Note that this is a Nash equilibrium of a larger game in which the correlating device moves first.

A more general correlated equilibrium for this game can be constructed as developed below. This equilibrium is sufficiently sophisticated that I will drop the charade that there just happens to exist a highly complex correlating device to which both Alice and Bob just happen to be attuned and jointly motivated to follow. Rather, I will stress that such a device must be causally instantiated in society, and the players must have the appropriate psychological machinery to recognize that the characteristics of this device is common knowledge between them. To dramatize this, I will call the correlating device a *choreographer*.

Consider a choreographer who would like to direct Alice to play d and Bob to play l so the joint payoff $(4, 4)$ could be realized. The problem is that if Alice obeys the choreographer, then Bob has an incentive to choose r, giving him a payoff of 5 instead of 4. Similarly, if Bob obeys the choreographer, then Alice has

an incentive to choose u, giving her a payoff of 5 instead of 4. The choreographer must therefore be more sophisticated.

Suppose the choreographer has three states. In ω_1, which occurs with probability α_1, he advises Alice to play u and Bob to play l. In ω_2, which occurs with probability α_2, the choreographer advises Alice to play d and Bob to play l. In ω_3, which occurs with probability α_3, the choreographer advises Alice to play d and Bob to play r. We assume Alice and Bob know α_1, α_2, and $\alpha_3 = 1-\alpha_1-\alpha_2$, and it is common knowledge that both have a *normative predisposition* to obey the choreographer unless they can do better by deviating. However, neither Alice nor Bob can observe the state ω of the choreographer, and each hears only what the choreographer tells them, not what the choreographer tells the other player. We will find the values of α_1, α_2, and α_3 for which the resulting game has a Pareto-efficient correlated equilibrium.

Note that Alice has *knowledge partition* $[\{\omega_1\}, \{\omega_2, \omega_3\}]$, meaning that she knows when ω_1 occurs but cannot tell whether the state is ω_2 or ω_3. This is because she is told to move u only in state ω_1 but to move d in both states ω_2 and ω_3. The conditional probability of ω_2 for Alice given $\{\omega_2, \omega_3\}$ is $p_A(\omega_2) = \alpha_2/(\alpha_2 + \alpha_3)$, and similarly $p_A(\omega_3) = \alpha_3/(\alpha_2+\alpha_3)$. Note also that Bob has knowledge partition $[\{\omega_3\}, \{\omega_1, \omega_2\}]$ because he is told to move r only at ω_3 but to move l at both ω_1 and ω_2. The conditional probability of ω_1 for Bob given $\{\omega_1, \omega_2\}$ is $p_B(\omega_1) = \alpha_1/(\alpha_1 + \alpha_2)$, and similarly $p_B(\omega_2) = \alpha_2/(\alpha_1 + \alpha_3)$.

When ω_1 occurs, Alice knows that Bob plays l, to which Alice's best response is u. When ω_2 or ω_3 occurs, Alice knows that Bob is told l by the choreographer with probability $p_A(\omega_2)$ and is told r with probability $p_A(\omega_3)$. Thus, despite the fact that Bob plays only pure strategies, Alice knows she effectively faces the mixed strategy l played with probability $\alpha_2/(\alpha_2+\alpha_3)$ and r played with probability $\alpha_3/(\alpha_2+\alpha_3)$. The payoff to u in this case is $5\alpha_2/(\alpha_2+\alpha_3)$, and the payoff to d is $4\alpha_2/(\alpha_2+\alpha_3)+\alpha_3/(\alpha_2 + \alpha_3)$. If d is to be a best response, we must thus have $\alpha_1 + 2\alpha_2 \leq 1$.

Turning to the conditions for Bob, when ω_3 occurs, Alice plays d so Bob's best response is r. When ω_1 or ω_2 occurs, Alice plays u with probability $p_B(\omega_1)$ and d with probability $p_B(\omega_2)$. Bob chooses l when $\alpha_1 + 4\alpha_2 \geq 5\alpha_2$. Thus, any α_1 and α_2 that satisfy $1 \geq \alpha_1 + 2\alpha_2$ and $\alpha_1 \geq \alpha_2$ permit a correlated equilibrium. Another characterization is $1 - 2\alpha_2 \geq \alpha_1 \geq \alpha_2 \geq 0$.

What are the Pareto-optimal choices of α_1 and α_2? Because the correlated equilibrium involves $\omega_1 \to (u, l)$, $\omega_2 \to (d, l)$, and $\omega_3 \to (d, r)$, the payoffs to (α_1, α_2) are $\alpha_1(5, 1) + \alpha_2(4, 4) + (1 - \alpha_1 - \alpha_2)(1, 5)$, which simplifies to $(1 + 4\alpha_1 + 3\alpha_2, 5 - 4\alpha_1 - \alpha_2)$, where $1 - 2\alpha_2 \geq \alpha_1 \geq \alpha_2 \geq 0$. This is a linear programming problem. It is easy to see that either $\alpha_1 = 1 - 2\alpha_2$ or $\alpha_1 = \alpha_2$ and $0 \leq \alpha_2 \leq 1/3$. The solution is shown in figure 1.

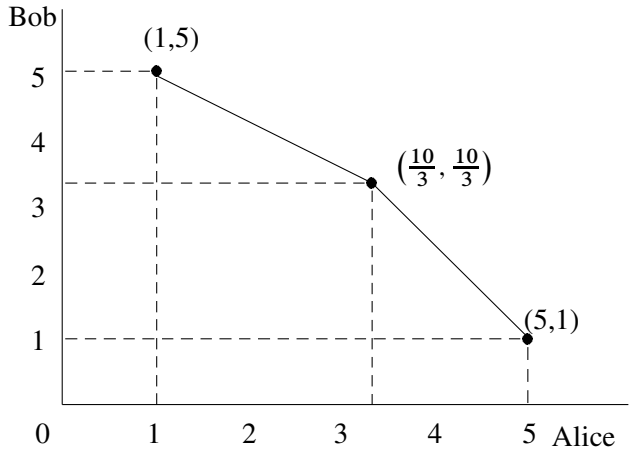

Figure 1: Alice, Bob, and the choreographer

The pair of straight lines connecting (1,5) to (10/3,10/3) to (5,1) is the set of Pareto-optimal points. Note that the symmetric point (10/3,10/3) corresponds to $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$.

6 Moral Behavior And the Fallible Choreographer

The Theorem 2 not only requires that the players have a common prior over the state space Ω, but also that the choreographer know everything that each of the players knows. This requirement is implicit in the requirement that $f(\omega) = \mathbf{s}(\omega)$ for all $\omega \in \Omega$. The most plausible causal structure for a choreographed equilibrium is that the social norm has been instantiated and is commonly known. The cultural system of the players together with their social epistemology, including the fact that social norms are common knowledge, then gives rise to common priors, and Bayesian rationality then leads players to implement the social norm.

When the assumption of choreographer omniscience fails, a correlated equilibrium may still obtain, provided the players have sufficiently strong prosocial preferences. There is strong evidence supporting the notion that rational individuals may have other-regarding preferences and/or may recognize certain moral virtues that lead them to conform voluntarily to a social norm in a situation where a self-regarding and amoral agent would not. In such cases, the choreographer may be obeyed even at a cost to the players, provided of course that the cost of doing so is not excessive (Gintis 2009, Ch. 3).

For instance, each agent's payoff might consist of a *public component* that is known to the choreographer and a *private component* that reflects the idiosyn-

crasies of the agent and is unknown to the choreographer. Suppose the most the private component in any state for an agent can be is α, but the agent's inclination to follow the choreographer has strength greater than α. Then, the agent continues to follow the choreographer's directions whatever the state of his private information. Formally, we say an individual has an *α-normative predisposition* towards conforming to the social norm if he strictly prefers to play his assigned strategy so long as all his pure strategies have payoffs no more than α greater than when following the choreographer. We call an α-normative predisposition a *social preference* because it facilitates social coordination but violates self-regarding preferences for $\alpha > 0$. There are evolutionary reasons for believing that humans have evolved such social preferences for fairly high levels of α in a large fraction of the population through gene-culture coevolution (Gintis 2003, Ch. 3).

Suppose, for example, that police in a certain town are supposed to apprehend criminals, where it costs police officer i a variable amount $f_i(\omega)$ to file a criminal report. If the identified perpetrator is in the same ethnic group as i, or if the perpetrator offers a bribe to be released, $f_i(\omega)$ might be very high, whereas an offender from a different ethnic group, or one who does not offer a bribe, might entail a low value of $f_i(\omega)$. How can this society erect incentives to induce the police to act in a non-corrupt manner?

Assuming police officer i is self-regarding and amoral, i will report a crime only if $f_i(\omega) \leq w$, where w is the reward for filing an accurate criminal report (assuming accuracy can be guaranteed by fact-checking). A social norm equilibrium that requires that all apprehended criminals be prosecuted cannot then be sustained because all officers for whom $f_i(\omega) > w$ with positive probability will at least at times behave corruptly. Suppose however officers have a *normative predisposition* to behave honestly, in the form of a police culture favoring honesty that is internalized (i.e., integrated into the preference ordering) by all officers. If $f_i(\omega) < w + \alpha$ with probability one for all officers i, where α is the strength of police culture, the social norm equilibrium can be sustained, despite the fact that the choreographer has incomplete information concerning events in which criminal behavior is detected.

7 Where Does Common Knowledge Come From?

Suppose we have a set of n agents, each of whom has a knowledge operator $\mathbf{K}_i$, $i = 1, \ldots, n$. We say an event E is *self-evident* to agent i if $\mathbf{K}_i E = E$; i.e., in every state in which E occurs, i knows that E occurs. It is obvious that E is self-evident to i if and only if E is a union of cells in i's knowledge partition $\mathcal{P}_i$.

We say $E \subseteq \Omega$ is a *public event* if E is self-evident for all $i = 1, \ldots, n$.

Clearly Ω is a public event, and if E and F are public events, so is $E \cap F$. Hence, for any $\omega \in \Omega$, there is a minimal public event $\mathbf{P}_*\omega$ containing ω; namely the intersection of all public events containing ω. It is easy to see that the public event operator $\mathbf{P}_*\omega$ corresponds to a partition $\mathcal{P}_*$ of Ω, namely the finest partition of Ω that is a common coarsening of the individual knowledge partitions $\mathbf{P}_i$.

We may define a *public event* operator $\mathbf{K}_*$ as the knowledge operator corresponding to $\mathbf{P}_*$, so $\mathbf{K}_* E = \{\omega | \mathbf{P}_*\omega \subseteq E\}$. We then see that an event E is a public event exactly when $\mathbf{K}_* E = E$. Thus, E is a public event if and only if E is self-evident to all players at each $\omega \in E$. Also, E is a public event if and only if E is the union of minimal public events of the form $\mathbf{P}_*\omega$. Moreover, if E is a public event, then at every $\omega \in E$ everyone knows that E is a public event at ω.

In the standard treatment (Lewis 1969, Aumann 1976), an event is *common knowledge* if everyone knows E, everyone knows that everyone knows E, and so on. We then have

THEOREM 3. *An event $E \subseteq \Omega$ is common knowledge if and only if it is a public event.*

To see this, suppose E is a public event. Then, for any $i, j, k \in \{1, \ldots, n\}$, $\mathbf{K}_i E = E$, $\mathbf{K}_j \mathbf{K}_i E = \mathbf{K}_j E = E$, $\mathbf{K}_k \mathbf{K}_j \mathbf{K}_i E = \mathbf{K}_k E = E$, and so on. Thus, all events of the form $\mathbf{K}_k \mathbf{K}_j \ldots \mathbf{K}_i E$ are self-evident for k, so E is common knowledge. Conversely, suppose that for any sequence $i, j, \ldots, k = 1, \ldots, n$, $\mathbf{K}_i \mathbf{K}_j \ldots \mathbf{K}_k E \subseteq E$. We define $\mathbf{P}_i^1 \omega = \cup_{j=1}^n \mathbf{P}_j \omega$, which is the set of states that are possible for at least one agent at ω. Given $\mathbf{P}_i^k \omega$, we then define $\mathbf{P}_i^{k+1} \omega = \cup_{j=1}^n \mathbf{P}_j^k \omega$. It is easy to see that $\mathbf{P}_*\omega = \cup_{k=1}^\infty \mathbf{P}_*^k \omega$. Then, for any $\omega \in E$, because $\mathbf{P}_i \omega \subseteq E$, we have $\mathbf{P}_*^1 \omega \subseteq E$. We also have $\mathbf{K}_i \mathbf{P}_*^1 \omega \subseteq E$ because $\mathbf{K}_i \mathbf{K}_j E \subseteq E$ for $i, j = 1, \ldots, n$, so $\mathbf{P}_*^2 \omega \subseteq E$. We now see that $\mathbf{P}_*^k \omega \subseteq E$ for all k, so $\mathbf{P}_*\omega \subseteq E$. Therefore E is the union of public events and hence is a public event.

We have defined a public event as an event that is self-evident to all players. We then showed that an event E is public if and only if it is common knowledge. It appears, then, that at a public event there is a perfect *commonality of knowledge*: players know a great deal about what other players know. Where does this knowledge come from? The answer is that we have tacitly assumed that each $\mathcal{P}_i$ is known to all, not in the formal sense of a knowledge operator but rather in the sense that an expression of the form $\mathbf{K}_i \mathbf{K}_j E$ makes sense and means "i knows that j knows that E." Formally, to say that i knows that j knows E at ω means that at every state $\omega' \in \mathbf{P}_j \omega$, $\mathbf{P}_i \omega' \subseteq E$. But i knows that this is the case only if he knows $\mathbf{P}_j \omega$, which allows him to test $\mathbf{K}_i \omega' \subseteq E$ for each $\omega' \in \mathbf{P}_j \omega$.

For example, suppose Alice, Bob, and Carole meet yearly on a certain date at a certain time to play a game $\mathcal{G}$. Suppose, by chance, all three happen to be in Dallas,

Texas, the day before, and although they do not see each other, each witnesses the same highly unusual event x. We define the universe $\Omega = \{\omega, \omega'\}$, where the unusual even occurs in ω but not in ω'. Then, $\mathbf{P}_A\omega = \mathbf{P}_B\omega = \mathbf{P}_C\omega = \{\omega\}$, and hence $\mathbf{K}_A\omega = \mathbf{K}_B\omega = \mathbf{K}_C\omega = \{\omega\}$. Thus ω is self-evident to all three individuals, and hence ω is a public event. Therefore at ω, Alice knows that Bob knows that Carole knows ω, and so on. But, of course, this is not the case. Indeed, none of the three individuals is aware that the others know the event x.

The problem is that we have misspecified the universe. Suppose an event ω is a four-vector, the first entry of which is either x or $\neg x$ (meaning "not x") and the other three are "true" or "false," depending on whether Alice, Bob, and Carole, respectively, knows or does not know whether x occurred. The universe Ω now has 16 distinct states, and the state ω that actually occurred is $\omega = [x$, true, true, true]. However, now $\mathbf{P}_A\omega = \{\omega' \in \Omega | \omega'[1] = x \wedge \omega'[2] = \text{true}\}$. Therefore, the state ω is now *not* self-evident for Alice. Indeed, the smallest self-evident event $\mathbf{P}_A\omega$ for Alice at ω in this case is Ω itself!

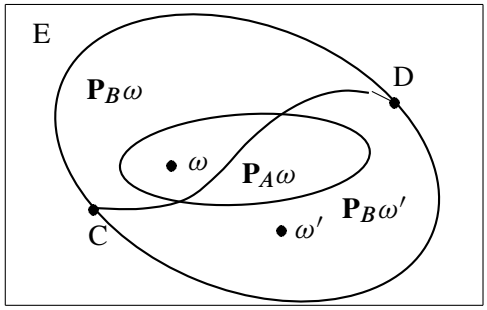

Figure 2: The case where, at ω, Bob knows that Alice knows E.

In fact, the standard epistemic model is reasonable only if all agents know the knowledge partitions of all other agents. This fact is illustrated in figure 2, which shows the situation where Alice knows E at ω, because her minimal self-evident event $\mathbf{P}_A\omega$ at ω (the smaller ellipse) lies within E. Moreover $\mathbf{P}_A\omega$ intersects two of Bob's minimal self-evident events, $\mathbf{P}_B\omega$ and $\mathbf{P}_B\omega'$ (the two parts of the larger ellipse separated by the wavy line CD). Because both of $\mathbf{P}_B\omega$ and $\mathbf{P}_B\omega'$ lie within E, Bob knows that Alice knows that E at ω (and at every other state in $\mathbf{P}_A\omega$).

This line of reasoning reveals a central *lacuna* in epistemic game theory: its semantic model of common knowledge does not examine the social bases of common knowledge. Economists have been misled by the elegant Theorem 3 that says mutual self-evidence implies common knowledge into believing that the ax-

ioms of rational choice imply something substantive concerning the commonality of knowledge across agents. They do not. Indeed, there is no formal principle specifying conditions under which distinct individuals attribute the same truth value to a proposition p with empirical content (we can assume rational agents all agree on mathematical and logical tautologies) or have a mental representation of the fact that others attribute truth value to p. We address this below by sketching the attributes of what we have termed *mutually accessible* events.

8 The Tactful Ladies

This section presents The Tactful Ladies, a well-known application of the standard, so-called *semantic* epistemic model outlined in previous sections. In the following section, I develop an alternative 'syntactic' epistemic model due to Aumann (1999) that clearly reveals the tacit assumptions concerning the sharing of mental constructs that are obscured in the semantic model. Following this, I re-analyze The Tactful Ladies using the syntactic model.

While walking in a garden, Alice, Bonnie, and Carole encountered a violent thunderstorm and were obliged to duck hastily into a restaurant for tea. Carole notices that Alice and Bonnie have dirty foreheads, although each is unaware of this fact. Carole is too tactful to mention this embarrassing situation, which would surely lead them to blush, but she observes that, like her, each of the two ladies knows that someone has a dirty forehead but is also too tactful to mention this fact. The thought occurs to Carole that she also might have a dirty forehead, but there are no mirrors or other detection devices handy that might help resolve her uncertainty.

At this point, a little boy walks by the three young ladies' table and exclaims, "I see a dirty forehead!" After a few moments of awkward silence, Carole realizes that she has a dirty forehead and blushes.

How is this feat of logical deduction possible? Certainly, it is mutually known among the ladies that at least one of them has a dirty forehead, so the little boy did not inform any of them of this fact. Moreover, each lady can see that each of the other ladies sees at least one dirty forehead, so it is mutually known that each lady knew the content of the little boy's message before he delivered it. However, the little boy's remark does inform each lady that they all know that they all know that one of them has a dirty forehead. This is something that none of the ladies knew before the little boy's announcement. For instance, Alice and Bonnie each knows she might not have a dirty forehead, so Alice knows that Bonnie might believe that Carole sees two clean foreheads, in which case Alice and Bonnie know that Carole might not know that there is at least one dirty forehead. Following the little boy's

announcement, however, and assuming the other ladies are logical thinkers (which they must be if they are Bayesian decision makers), Carole's inference concerning the state of her forehead is unavoidable.

To see why, suppose Carole does not have a dirty forehead. Carole then knows that Alice sees one dirty forehead (Bonnie's), so Alice has learned nothing from the little boy's remark. But Carole knows that Bonnie sees that Carole's forehead is not dirty, so if Bonnie's forehead is not dirty, then Alice would see two clean foreheads, and the little boy's remark would have implied that Alice knows that she is the unfortunate possessor of a dirty forehead. Because Alice did not blush, Carole knows that Bonnie would have concluded that she herself must have a dirty forehead and would have blushed. Because Bonnie did no such thing, Carole knows that her assumption that she has a clean forehead is false.

To analyze this problem formally, suppose Ω consists of eight states of the form $\omega = xyz$, where $x, y, z \in \{d, c\}$ are the states of Alice, Bonnie, and Carole, respectively, and where d and c stand for "dirty forehead" and "clean forehead," respectively. Thus, for instance, $\omega = ccd$ is the state of the world where Carole has a dirty forehead but Alice and Bonnie both have clean foreheads. When Carole sits down to tea, she knows $E_C = \{ddc, ddd\}$, meaning she sees that Alice and Bonnie have dirty foreheads, but her own forehead could be either clean or dirty. Similarly, Alice knows $E_A = \{cdd, ddd\}$ and Bonnie knows $E_B = \{dcd, ddd\}$. Clearly, no lady knows her own state. What does Bonnie know about Alice's knowledge? Because Bonnie does not know the state of her own forehead, she knows that Alice knows the event "Carole has a dirty forehead," which is $E_{BA} = \{cdd, ddd, ccd, dcd\}$. Similarly, Carole knows that Bonnie knows that Alice knows $E_{CBA} = \{cdd, ddd, ccd, dcd, cdc, ddc, ccc, dcc\} = \Omega$. Assuming Carole has a clean forehead, she knows that Bonnie knows that Alice knows $E'_{CBA} = \{cdc, ddc, dcc, ccc\}$. After the little boy's announcement, Carole then knows that Bonnie knows that Alice knows $E''_{CBA} = \{cdc, ddc, dcc\}$, so if Bonnie did not have a dirty forehead, she would know that Alice knows $E''_{BA} = \{dcc\}$, so Bonnie would conclude that Alice would blush. Thus, Bonnie's assumption that she herself has a clean forehead would be incorrect, and she would blush. Because Bonnie does not blush, Carole knows that her assumption that she herself has a clean forehead is incorrect.

9 A Syntactic Model of Distributed and Shared Knowledge

Following Aumann (1999), suppose we have n individuals, and a set of letters from an alphabet $\mathcal{X} = \{x, y, z, \ldots\}$, symbols $\vee$, $\neg$, $k_1, \ldots, k_n$ and left, '(' and right, ')' parentheses. We think of the letters as representing sentences that can be either

true or false in different states of the world. We interpret $k_i\,x$ as asserting that agent i knows that x is true. *Formulas* are constructed recursively as follows:

a. Every letter is a formula.

b. If f and g are formulas, so are $(f) \vee (g)$, $\neg(f)$, and $k_i\,(f)$ for each i.

We abbreviate $(\neg f) \vee g$ as $f \Rightarrow g$, $\neg(\neg f \vee \neg g)$ as $f \wedge g$, $(f \Rightarrow g) \wedge (g \Rightarrow f)$ as $f \iff g$, and we drop parentheses where no ambiguity results, assuming the usual precedence ordering of the propositional calculus, and assigning the highest precedence to the knowledge symbols k_i. The above conditions ensure that every tautology of the propositional calculus based on $\mathcal{X}$ is a formula (Hedman 2004).

A *list* $\mathcal{L}$ is a set of formulas. A formula is a *tautology* if it is a tautology of the propositional calculus, or it has one of the following forms, where f and g are formulas:

$$k_i\,f \Rightarrow f \tag{3}$$

$$k_i\,f \Rightarrow k_i\,k_i\,f \tag{4}$$

$$k_i\,f \wedge k_i\,(f \Rightarrow g) \Rightarrow k_i\,g \tag{5}$$

$$\neg k_i\,f \Rightarrow k_i\,\neg k_i\,f. \tag{6}$$

$$f \in \mathcal{T} \Rightarrow k_i\,f \in \mathcal{T} \tag{7}$$

Note that the knowledge operators $\mathbf{K}_i$ from the standard semantic model has each of these properties. Formula (6), called the *axiom of transparency*, is required to ensure that the semantic realization of the syntactic system has a partition structure. Note that (5), which says that the knowledge operator satisfies *modus ponens*, is equivalent to $k_i\,(f \Rightarrow g) \Rightarrow (k_i\,f \Rightarrow k_i\,g)$.

We call a system consisting of the alphabet $\mathcal{X}$, the formulas and the tautologies a *syntactic system* $\mathcal{S}$. The set of tautologies $\mathcal{T}$ is closed under *modus ponens* (i.e., $f, (f \Rightarrow g) \in \mathcal{T}$ implies $g \in \mathcal{T}$) and the knowledge operator (i.e., $f \in \mathcal{T}$ implies $k_i\,f \in \mathcal{T}$).

A *state* ω is list that is closed under *modus ponens*, and for every formula f, exactly one of f and $\neg f$ is in ω. It is easy to see that if ω is a state, then $\mathcal{T} \subset \omega$. For otherwise ω would contain a false formula from the propositional calculus, which by *modus ponens* implies the ω contains all formulas, which is false by construction. Moreover, every state ω is a *complete* list of the formulas that are true in that state; i.e., we cannot add another formula to ω without violating the list property.

It is easy to see that a state ω includes the truth values of all $x \in \mathcal{X}$, as well as the information each player needs to ascertain what other players know in that state.

10 The Tactful Ladies and the Commonality of Knowledge

The Tactful Ladies Problem involves many unstated epistemological assertions going far beyond the common knowledge of rationality involved in the conclusion that Carole knows the state of her forehead. Let us see exactly what they are using the syntactic model developed in the previous section.

Let x_i be the condition that i has a dirty forehead and let k_i be the knowledge operator for i, where $i = A, B, C$, standing for Alice, Bonnie, and Carole, respectively. When we write i, we mean any $i = A, B, C$, and when we write i, j, we mean any $i, j = A, B, C$, with $j \neq i$, and when we write i, j, m we mean $i, j, m = A, B, C$ and $i \neq j \neq m \neq i$. Let y_i be the condition that i blushes. The six symbols x_i and y_i represent the possible states of affairs in a state space Ω. Let E be the event prior to the little boy's exclamation $b = x_A \vee x_B \vee x_C$.

The statement of the problem tells us that $x_i \in E$ and $k_i x_j \in E$; i.e., each lady sees the forehead of the other two ladies, but not her own. The problem also asserts that $k_i x_i \Rightarrow y_i \in E$ (a lady who knows she has a dirty forehead will blush), and $y_i \Rightarrow k_j y_i \in E$ (i.e., each lady knows when one of the others blushes). It is easy to check that these conditions are compatible with $\neg k_i x_i \in E$; i.e., no lady knows the state of her own forehead at event E. These conditions also imply that $k_i b \in E$ (each lady knows the little boy's statement is true).

While the problem intends that $k_i x_j \Rightarrow k_i k_m x_j \in E$ (i.e., if i knows that j has a dirty forehead, she then knows that m knows this as well), this implication does not follow from any principle of rationality, so we must include it as a new principle. The concept needed is that of a *mutually accessible natural occurrence*. The mutual accessibility of x_i to j and m may appear to be a weak assumption, but in fact it is the *first time* in this section that we have made a substantive assertion that one agent knows that another agent knows something. With this assumption, which we explore in the next section, it follows that $k_i k_j b \in E$—each lady knows the others know b holds in E (recall that b is the little boy's statement that ccc is false). To see this, note that $k_i x_j \Rightarrow k_i k_m x_j \Rightarrow k_i k_m b$, which is true for all i and $m \neq i$.

Let E' be the state of knowledge following the exclamation $b = x_A \vee x_B \vee x_C$, which we assume is common knowledge. To prove that in E' one of the ladies (e.g., Carole) blushes, we will assume that y_i is mutually accessible to j, m, and j (i.e., when one lady knows some y_i, she knows that the others know y_i, and is a symmetric reasoner with respect to m concerning event y_i (i.e. if i knows z and knows that j has the information to deduce z, then $k_i k_j z$).

The reasoning following the little boy's statement can be summarized as follows. We will show that if Carole assumes $\neg x_C$ at any state in E', she will arrive at

a contradiction. Assuming $\neg x_C$ is true and b is common knowledge, and writing

$$p = (\neg x_B \Rightarrow k_A \neg x_B \Rightarrow k_A(\neg x_B \wedge \neg x_C \wedge b) \Rightarrow k_A x_A \Rightarrow y_A),$$

we have by symmetric reasoning, $k_C\, p \Rightarrow k_C k_B\, p \Rightarrow k_C k_B\, y_A \Rightarrow k_C\, y_A$, which is false in E'. Thus in E', $k_C k_B x_B \Rightarrow k_C\, y_B$, which is not true at any state in E'. Hence x_C is true in E', and since Carole knows the current state is in E', $k_C\, x_C$, which implies y_C; i.e., Carole blushes.

11 From Natural Occurrence to Common Knowledge

The question as to how rational individuals come to share knowledge and know that this is the case is one of the classic problems of philosophy. Because philosophers have not come to agreement concerning the solution to this problem, it might be thought that epistemic game theory has little to go on in developing a model of knowledge sharing. In this section I show that this is incorrect by exhibiting some ubiquitous forms of knowledge sharing where the issues are sufficiently simple to resolve to our satisfaction.

There is a basis for the formation of common priors to the extent that the event in question is what we may call a *natural occurrence*, such as "the ball is yellow," that can be inferred from first-order sense data. Thus in the previous section, we treated blushing as a natural occurrence. We say a natural occurrence is *mutually accessible* to a group of agents when this first-order sense data is accessible to all members of the group, so that a member who knows N then knows that all the other members know N. For instance, if i and j are both looking at the same yellow ball, if each sees the other looking at the ball, and if each knows the other has normal vision and is not delusional, then the ball's color is mutually accessible: i knows that j knows that the ball is yellow, and conversely. In short, we can assume that a social situation involving a set of individuals can share an attentive state concerning a natural occurrence such that, the natural occurrence is mutually accessible (Tomasello 1999, Lorini et al. 2005).

Higher-order epistemic constructs, such as beliefs concerning the intentions, beliefs, and prospective actions of other individuals, beliefs about the natural world that cannot be assessed through individual experience, as well as beliefs about suprasensory reality, are not natural occurrences and are not mutually accessible (Morris 1995, Gul 1998, Dekel and Gul 1997). How, then, do such higher-order constructs become commonly known?

The answer is that members of our species have the capacity to conceive that other members have minds and respond to experience in a manner parallel to themselves—a capacity that is extremely rare and may be possessed by humans

alone (Premack and Woodruff 1978, Adolphs 2009). Thus, if agent i believes something, and if i knows that he shares certain environmental experiences with agent j, then i knows that j believes this thing as well. In particular, humans have cultural systems that provide natural occurrences that serve as *symbolic cues* for higher-order beliefs and expectations. Common priors, then, are the product of common culture.

The neuropsychological literature on how minds know other minds deals with mirror neurons, the human prefrontal lobe, and other brain mechanisms that facilitate the sharing of knowledge and beliefs (Iacoboni 2009). From the viewpoint of modeling human behavior, these facilitating mechanisms must be translated into axiomatic principles of strategic interaction.

Many events are defined in part by the mental representations of the individuals involved. For instance, an individual may behave very differently if he construes an encounter as an impersonal exchange as opposed to a comradely encounter. Mental events fail to be mutually accessible because they are inherently private signals. Nevertheless, there are mutually accessible events N that reliably *indicate* social events E that include the states of mind of individuals in the sense that for any individual i, if i knows N, then i knows E (Lewis 1969, Cubitt and Sugden 2003).

For instance, if I wave my hand at a passing taxi in a large city, both I and the driver of the taxi will consider this an event of the form "hailing a taxi." When the driver stops to pick me up, I am expected to enter the taxi, give the driver an address, and pay the fare at the end of the trip. Any other behavior would be considered bizarre.

By an *indicator* we mean a mutually accessible event N that specifies a social event (not a natural occurrence) E to all individuals in a group; i.e., for any individual i, $\mathbf{K}_i N \Rightarrow \mathbf{K}_i E$. Indicators are generally learned by group members through acculturation processes. When one encounters a novel community, one undergoes a process of learning the various indicators of a social event specific to that community. In behavioral game theory an indicator is often called a *frame* of the social event it indicates, and then the *framing effect* includes the behavioral implications of expectations cued by the experimental protocols themselves.

We define individual i as a *symmetric reasoner* with respect to individual j for an indicator N of event E if, whenever i knows N, and i knows that j knows N, then i knows that j knows E; i.e., $\mathbf{K}_i N \wedge \mathbf{K}_i \mathbf{K}_j N \Rightarrow \mathbf{K}_i \mathbf{K}_j E$ (Vanderschraaf and Sillari 2007). We say the individuals in the group are symmetric reasoners if, for each i, j in the group, i is a symmetric reasoner with respect to j.

Like mutual accessibility, joint attentive states, and indicators, symmetric reasoning is an addition to Bayesian rationality that serves as a basis for the concordance of beliefs. Indeed, one may speculate that our capacity for symmetric

reasoning is derived by analogy from our recognition of mutual accessibility. For instance, I may consider it just as clear that I am hailing a taxi as that the vehicle in question is colored yellow, and has a lighted sign saying "taxi" on the roof.

THEOREM 4. *Suppose individuals in a group are Bayesian rational symmetric reasoners with respect to the mutually accessible indicator N of E. If it is mutual knowledge that the current state $\omega \in N$, then E is common knowledge at ω.*

Proof: Suppose $\mathbf{P}_i\omega \subseteq N$ for all i. Then, for all i, $\mathbf{P}_i\omega \subseteq E$ because N indicates E. For any i, j, because N is mutually accessible, $\omega \in \mathbf{K}_i\mathbf{K}_j N$, and because i is a symmetric reasoner with respect to j, $\omega \in \mathbf{K}_i\mathbf{K}_j E$. Thus, we have $\mathbf{P}_i\omega \subseteq \mathbf{K}_j E$ for all i, j (the case $i = j$ holding trivially). Thus, N is an indicator of $\mathbf{K}_j E$ for all j. Applying the above reasoning to indicator $\mathbf{K}_k E$, we see that $\omega \in \mathbf{K}_i\mathbf{K}_j\mathbf{K}_k E$ for all i, j, and k. All higher levels of mutual knowledge are obtained similarly, proving common knowledge. ■

COROLLARY 1. *Suppose in state ω that N is a mutually accessible natural occurrence for a group of Bayesian rational symmetric reasoners. Then N is common knowledge in state ω.*

Proof: When $\omega \in N$ occurs, N is mutually known since N is a natural occurrence. Obviously, N indicates itself, so the assertion follows from theorem 4.

Note that we have adduced common knowledge of an event from simpler epistemic assumptions, thus affording us some confidence that the common knowledge condition has some chance of realization in the real world. This is in contrast to common knowledge of rationality, which is taken as primitive data and hence has little plausibility (Gintis 2009, Gintis in press). Communality of knowledge should always be derived from more elementary psychological and social regularities.

12 The Failure of Methodological Individualism

There is a tacit understanding among classical game theorists that no information other than the rationality of the agents should be relevant to analyzing how they play a game. This understanding is *methodological individualism*, a doctrine that holds that nothing beyond the characteristics of individuals is needed to model social behavior, so that in particular, higher-level social constructs such as social norms and institutions are explicable in terms of the interaction of individuals.

The most prominent early proponent of methodological individualism was Austrian school economist and philosopher Ludwig von Mises, in his book *Human Action*, first published in 1949. While most of Austrian school economic theory has

not stood the test of time, methodological individualism has, if anything, grown in stature among economists. "Nobody ventures to deny," writes von Mises, "that nations, states, municipalities, parties, religious communities, are real factors determining the course of human events." He continues: "Methodological individualism, far from contesting the significance of such collective wholes, considers it as one of its main tasks to describe and to analyze their becoming and their disappearing, their changing structures, and their operation.... a social collective has no existence and reality outside of the individual members' actions. ... the way to a cognition of collective wholes is through an analysis of the individuals' actions." (p. 42).

A passing familiarity with levels of scientific explanation shows that this argument is not necessarily well-founded. A computer, for instance, is composed of a myriad of solid-state and other electrical and mechanical devices, but stating that one can successfully model the operation of a computer using only facts about the behavior of these underlying parts is false. Similarly, eukaryotic cells are composed of a myriad of organic chemicals, yet organic chemistry does not supply all the tools for modeling cell dynamics.

We learn from modern complexity theory that there are many levels of physical existence on earth, from elementary particles to human beings, each level solidly grounded in the interaction of entities at a lower level, yet having emergent properties that are ineluctably associated with the dynamic interaction of its lower-level constituents, yet are incapable of being explained on a lower level. The panoramic history of life synthesis of biologists Maynard Smith and Szathmáry (1997) elaborates this theme that every major transition in evolution has taken the form of a higher level of biological organization exhibiting properties that cannot be deduced from its constituent parts. Morowitz (2002) extends the analysis to emergence in physical systems. Indeed, the point should not be mystifying because there is nothing preventing the most economical model of a phenomenon from being the model itself (Chaitin 2004). Adding emergent properties as fundamental entities in the higher-level model thus may permit the otherwise impossible: the explanation of complex phenomena.

Epistemic game theory suggests that the conditions ensuring that individuals play an equilibrium are not limited to their *personal* characteristics but rather include their *common* characteristics, in the form of common priors and common knowledge. We saw (theorem 4) that both individual characteristics and collective understandings, the latter being irreducible to individual characteristics, are needed to explain common knowledge. It is for this reason that methodological individualism is incorrect when applied to the analysis of social life.

The material presented here suggests the fruitfulness of dropping methodological individualist ideology but carefully articulating the analytical linkages between

individually rational behavior and the social institutions that align the beliefs and expectations of individuals, making possible effective social intercourse.

13 Social Epistemology: Public Indicators and Social Frames

Let G be the event that the current social situation is a game $\mathcal{G}$. G is not a natural occurrence and hence cannot be mutually accessible to the players of $\mathcal{G}$. However, mutual knowledge that $\mathcal{G}$ is being played is a condition for Nash equilibrium according to Theorem B of Aumann and Brandenburger (1995). How does G become mutually known? There may be a mutually accessible event F that reliably *indicates* that G is the case, in the sense that for any individual i, $\mathbf{K}_i F \subseteq \mathbf{K}_i G$ (Lewis 1969, Cubitt and Sugden 2003, Vanderschraaf and Sillari 2007). We think of G as representing the game that is socially appropriate when the "frame" F occurs. For instance, returning to the taxi narrative, both I and the driver of the taxi will consider my arm-waving to be an event of the form "hailing a taxi." The underlying mutually accessible natural occurrences F constituting a frame for this game include the color of the automobile (yellow), the writing on the side of the automobile ("Joe's Taxi"), and my frantic arm movements while looking at the automobile. When the driver stops to pick me up, I am expected to enter the taxi, give the driver an address, and pay the fare at the end of the trip. Any other behavior would be considered bizarre and perhaps suspicious. For instance if, instead of giving the driver an address, I invited the taxi driver to have a beer, or asked him to lend me money, or sought advice concerning a marital problem, the driver would consider the situation to be egregiously out of order.

In many social encounters, there are mutually accessible cues F that serve as a frame indicating that a specific game G is being played, or is to be played. These frames are learned by individuals through a social acculturation processes. When one encounters a novel community, one undergoes a process of learning the particular mutually accessible indicators of social frames in that community. Stories of misunderstanding such indicators, and hence misconstruing the nature of a social frame is the common subject of amusing anecdotes and tales.[1]

We may summarize these concepts by defining a frame $F \subseteq \Omega$ as a *public indicator* of G for n individuals if F indicates G for all agents, and F is mutually accessible for all pairs of agents. We then have

[1] I am reminded of such an event that I experienced in an unfamiliar city, Shanghai. At rush hour, I went through our usual motions to hail a taxi, with no success—several available taxis simply passed on by. A stranger motioned to us to stand at a certain spot along the street and hail from there. Although this spot looked no different to me than any other spot on the street, a taxi pulled over almost immediately.

THEOREM 5. *Suppose F is a public indicator of G and F is mutually accessible to all agents $i = 1, \ldots n$. Then G is mutually known for all $\omega \in F$.*

14 Homo Ludens: Rules as Mutually Accessible Conditions

Humans are not the only species that play games. Dogs chase and wrestle without causing harm. They are playing and learning the rules of their games (Bekoff 2008). In many mammalian species, animals signal, learn fair play, and punish others who do not play fair, and apologize when caught violating the rules.

However, there is no non-human animal that is capable of playing a game using new rules that are not part of its natural repertoire. This is why there is no experimental data illustrating how non-human species play the Ultimatum game or the Prisoner's Dilemma. Of course, one can formally place two ravens in an game-theoretic situation, but there is no evidence that either participant realizes that the other is obeying a set of rules imposed by the experimenter.

The fact that game-playing is a deep feature of human culture was stressed by Huizinga (1955[1938]), although in a context uninformed by epistemic game theory. Humans can play games not only because they have a level of cognitive ability that permits learning the rules of the game, but also because they are symmetric reasoners: if A learns the rules of a game and knows that B and C have experienced the same social process through which such learning occurs, then A knows that B and C know the rules, that B and C know that A knows the rules, that A knows that B and C know that A knows the rules, and so on. In game theory, the assumption of common knowledge of the rules of the game is rarely even mentioned, much less justified through an explicit epistemic argument. Yet, from the point of view of the evolved psychology of our species, this is a most remarkable, and virtually unexplored, human capacity.

15 Conclusion

There are important implications of the fact that a social norm is the choreographer of a correlated equilibrium rather than a Nash equilibrium selection device. A simple game $\mathcal{G}$ may have many qualitatively distinct correlated extensions $\mathcal{G}^+$, which implies that life based on social norms can be qualitatively richer than the simple underlying games that they choreograph. The correlated equilibrium concept thus indicates that social theory goes beyond game theory to the extent that it supplies dynamical and equilibrium mechanisms for the constitution and transformation of social norms. At the same time, the power of the correlated equilibrium interpretation of social norms indicates that social theory that rejects game theory is likely

to be significantly handicapped.

In a fundamental sense the correlated equilibrium is more basic than the Nash equilibrium. The epistemic conditions under which rational agents will play a Nash equilibrium are extremely confining and cannot be expected to hold in any but a small subset of even the simplest games, such as games with very few strategies per player that are solvable by the iterated elimination of strongly dominated strategies (Aumann and Brandenburger 1995, Basu 1994, Gintis 2009). By contrast, Bayesian rationality is effectively isomorphic with correlated equilibrium in the presence of common priors.

The epistemic game theoretic analysis of social norms can serve as the theoretical core for a general social theory of human strategic interaction that renders harmonious the approaches of economics, psychology, and sociology (Gintis 2009, Ch. 12). This analysis shows precisely where classical game theory goes wrong: it focuses on Nash as opposed to correlated equilibria, and its adherence to methodological individualism leads it to ignore the rich social fabric of potential conditioning devices (Ω, f), each corresponding to a distinct social structure of interaction. Moreover, the theory renders salient the epistemic conditions for the existence of a social norm, conditions that are fulfilled only in an idealized, fully equilibrated, social system. In general, social norms will be contested and only partially implemented, and the passage from one choreographed equilibrium to another will be mediated by forms of collective action and individual heroism that cannot be currently explicated in game theoretic terms.

REFERENCES

Adolphs, Ralph, "The Social Brain: Neural Basis of Social Knowledge," *Annual Review of Psychology* 60 (2009):693–716.

Aumann, Robert J., "Agreeing to Disagree," *The Annals of Statistics* 4,6 (1976):1236–1239.

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

— , "Interactive Epistemology I: Knowledge," *International Journal of Game Theory* 28 (1999):264–300.

— and Adam Brandenburger, "Epistemic Conditions for Nash Equilibrium," *Econometrica* 65,5 (September 1995):1161–1180.

Basu, Kaushik, "The Traveler's Dilemma: Paradoxes of Rationality in Game Theory," *American Economic Review* 84,2 (May 1994):391–395.

Bekoff, Marc, *Animals at Play: Rules of the Game* (Philadelphia: Temple University Press, 2008).

Binmore, Kenneth G., *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).

— , *Natural Justice* (Oxford: Oxford University Press, 2005).

Chaitin, Gregory, *Algorithmic Information Theory* (Cambridge: Cambridge University Press, 2004).

Cubitt, Robin P. and Robert Sugden, "Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory," *Economics and Philosophy* 19 (2003):175–210.

Dekel, Eddie and Faruk Gul, "Rationality and Knowledge in Game Theory," in David M. Kreps and K. F. Wallis (eds.) *Advances in Economics and Econometrics, Vol. I* (Cambridge: Cambridge University Press, 1997) pp. 87–172.

Gintis, Herbert, "The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms," *Journal of Theoretical Biology* 220,4 (2003):407–418.

— , *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences* (Princeton, NJ: Princeton University Press, 2009).

— , "Common Knowledge and Rationality," *Rationality and Society* (in press).

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

Hedman, Shawn, *A First Course in Logic: An Introduction to Model Theory, Proof Theory, Computability, and Complexity* (Oxford: Oxford University Press, 2004).

Huizinga, Johan, *Homo Ludens* (Boston: Beacon Press, 1955[1938]).

Iacoboni, Marco, "Imitation, Empathy, and Mirror Neurons," *Annual Review of Psychology* 60 (2009):653–670.

Lewis, David, *Conventions: A Philosophical Study* (Cambridge, MA: Harvard University Press, 1969).

Lorini, Emiliano, Luca Tummolini, and Andreas Herzig, "Establishing Mutual Beliefs by Joint Attention: Towards and Formal Model of Public Events," 2005. Institute of Cognitive Sciences, Rome.

Maynard Smith, John and Eors Szathmáry, *The Major Transitions in Evolution* (Oxford: Oxford University Press, 1997).

Morowitz, Harold, *The Emergence of Everything: How the World Became Complex* (Oxford: Oxford University Press, 2002).

Morris, Stephen, "The Common Prior Assumption in Economic Theory," *Economics and Philosophy* 11 (1995):227–253.

Premack, D. G. and G. Woodruff, "Does the Chimpanzee Have a Theory of Mind?," *Behavioral and Brain Sciences* 1 (1978):515–526.

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

— , "Spontaneous Order," *Journal of Economic Perspectives* 3,4 (Fall 1989):85–97.

Taylor, Michael, *Anarchy and Cooperation* (London: John Wiley and Sons, 1976).

— , *Community, Anarchy, and Liberty* (Cambridge, UK: Cambridge University Press, 1982).

— , *The Possibility of Cooperation* (Cambridge, UK: Cambridge University Press, 1987).

Tomasello, Michael, *The Cultural Origins of Human Cognition* (Cambridge, MA: Harvard University Press, 1999).

Vanderschraaf, Peter and Giacomo Sillari, "Common Knowledge," in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (plato.stanford.edu/archives/spr2007/entries/common-knowledge: Stanford Univerisity, 2007).