Social norms as choreography

Herbert Gintis

Santa Fe Institute, USA and Central European University, Hungary

abstract   This article shows that social norms are better explained as correlating devices for a correlated equilibrium of the underlying stage game, rather than Nash equilibria. 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. When the correlating device has perfect information, we need in addition only to posit that individuals obey the social norm when it is costless to do so. When the correlating device has incomplete information, the operation of the social norm requires that individuals have a predisposition to follow the norm even when this is costly. The latter case explains why social norms are associated with other-regarding preferences and provides a basis for analyzing honesty and corruption.

keywords   Nash equilibrium, correlated equilibrium, social norm, correlating device, honesty, corruption, Bayesian rationality

1. Introduction

This article extends the seminal contributions of David Lewis, Michael Taylor, Robert Sugden, Cristina Bicchieri, and Ken Binmore in treating social norms as Nash equilibria of noncooperative games played by rational agents.[1] The insight underlying all these contributions is that if agents play a game G with several Nash equilibria, a social norm can serve to choose among these equilibria. While this insight applies to several important social situations, it does not apply to most. In this article, I will suggested a more general principle, according to which a social norm is a *choreographer* of a 'supergame' G^+ of G. By the term 'choreographer' I mean a correlating device that implements a correlated equilibrium of G in which all agents play strictly pure strategies (these terms are defined below).

The social norms as choreographer has two attractive properties lacking in the social norms as Nash equilibria. First, the conditions under which rational agents play Nash equilibria are generally complex and implausible, whereas rational agents in a very natural sense play correlated equilibria (see Section 7). Second, the social norms as Nash equilibria approach cannot explain why compliance with social norms is often based on other-regarding and moral preferences in which agents are willing to sacrifice on behalf of compliance with social norms. 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.

2. From Nash to correlated equilibrium

Consider a society in which men prefer the company of women and vice versa, but when they consort, their two forms of entertainment, *m* and *f*, are favored by men and women, respectively. Their payoffs are described in Figure 1. There are two pure-strategy equilibria and one mixed-strategy equilibrium for this game. Clearly, both sexes would be better off if they stuck to either of their pure-strategy equilibria than by choosing the mixed-strategy equilibrium, in which each plays his or her favorite choice with a probability of one-third, resulting in the payoff of two-thirds to each.

	Alice *m*	*f*
Bob *m*	2,1	0,0
f	0,0	1,2

Figure 1 **The battle of the sexes game**

No principle of rational choice objectively favors any of these three equilibria, but a good case could be made for the mixed-strategy equilibrium, as it conforms to a priori symmetry principles that are likely to hold in the absence of other information. Suppose there is a social norm that says 'The man always gets to decide where to go.' Then, if both men and women believe that this social norm is in effect, each knows the other will choose *m*, and hence each will choose *m*, thus validating the social norm.

For a more complicated, but more realistic example, consider a town with a rectangular north–south, east–west array of streets. In the absence of a social norm, whenever two cars find themselves in a condition of possible collision, both stop and each waits for the other to go first. Obviously, not a lot of driving will get done. So, consider a social norm in which (1) all cars drive on the right, (2) at an intersection both cars stop and the car that arrived first proceeds

forward, and (3) if both cars arrive at an intersection at the same time, the car that sees the other car on its left proceeds forward. This is one of several social norms that will lead to an efficient use of the system of streets, provided there is not too much traffic.

Suppose, however, that there is so much traffic that cars spend much of their time stopping a crossings. We might then prefer the social norm in which we amend the above social norm to say that cars traveling north–south always have the right of way and need not stop at intersections. However, if there is really heavy traffic, east–west drivers may never get a chance to move forward at all. Moreover, if some intersections violate the strict north–south, east–west orientation, it may be unclear who does have the right of way in some cases.

Suppose, then, we erect a set of signals at each intersection that indicate 'Go' or 'Stop' to drivers moving in one direction and another set of 'Go' or 'Stop' signals for drivers moving in the crossing direction. We can then correlate the signals so that when one set of drivers see 'Go', the other set of drivers see 'Stop'. The social norm then says that 'If you see Go, do not stop at the intersection, but if you see Stop, then stop and wait for the signal to change to Go.' We add to the social norm that the system of signals alternates sufficiently rapidly and there is a sufficiently effective surveillance system that no driver has an incentive to disobey the social norm.

This would appear to be a perfect example of a social norm, indeed, a convention. However, the original game does not have a system of signals, and the proposed social norm does not single out a Nash equilibrium of the original game. Indeed, it is easy to see that there is a wide array of payoffs in the original game in which the only Nash equilibrium is for both cars to stop when an encounter occurs.

The system of signals in fact represents what is called a *correlated equilibrium* of G.[2] Basically, a correlated equilibrium adds a new player, whom I shall call the *choreographer*,[3] who has signal set S = {'Go,'Stop} and who views a set Ω of 'states of nature', and for each intersection i in the town and each state of nature $\omega \in \Omega$, chooses a signal $s_{in}(\omega) \in S$ for the north–south drivers at i and a signal $s_{ie}(\omega) \in \Omega$ with $s_{ie}(\omega) \neq s_{in}(\omega)$ for the east–west drivers. For simplicity, we may think of $\omega \in \Omega$ as a time of day or as a time elapsed since the last signal flip. So the choreographer flips the signals to the two groups of drivers according to some time schedule. The social norm is then the strict Nash equilibrium of the expanded game G^+ in which all rational agents obey the traffic laws.

In no sense is this a Nash equilibrium of the original game. Nor is G^+ unique; we can propose many alternative correlating devices, based on different state spaces Ω, that produce substantially different patterns of traffic. For instance, we can include in each $\omega \in \Omega$ a measure of the volume of traffic in the two directions at the intersection, and the choreography can increase the 'Go' time for the drivers that are currently in the more congested direction.

	H	D
H	$\dfrac{v-w}{w},\dfrac{v-w}{w}$	$v,0$
D	$0,v$	$\dfrac{v}{2},\dfrac{v}{2}$

Figure 2 **The Hawk-Dove game**

For another example, consider a society in which agents contest for posses-sion of valuable territory. The game G has two possible strategies. The hawk (H) strategy is to escalate battle until injured or your opponent retreats. The dove (D) strategy is to display hostility, but retreat before sustaining injury if your opponent escalates matters. The payoff matrix is given in Figure 2, where $v > 0$ is the value of the territory, $w > v$ is the cost of injury, and $(v - w)/2$ is the payoff when two hawks meet. The agents can play mixed strategies, but they cannot condition their play on whether they are player 1 or player 2, and hence players cannot condition their behavior on being player 1 or player 2. The payoffs are shown in Figure 2.

The Hawk-Dove game has a unique symmetric equilibrium, determined as fol-lows. Let α be the probability of playing hawk. The payoff to playing hawk is then $\pi_h = \alpha(v - w)/2 + (1 - \alpha)v$ and the payoff to playing dove is $\pi_d = \alpha(0) + (1 - \alpha)v/2$. These two are equal when $\alpha^* = v/w$, so the unique, symmetric Nash equilibrium occurs when $\alpha = \alpha^*$. The payoff to each player is thus

$$\pi_d = (1 - \alpha)\frac{v}{2} = \frac{v}{2}\left(\frac{w - v}{w}\right)$$

Note that when w is close to v, almost all the value of the territory is dissipated in fighting.

Clearly, because there is only one symmetric Nash equilibrium, the only pos-sible social norm associated with a Nash equilibrium is extremely inefficient. Suppose, however, that when two players contest, each knows which of the two happened upon the territory first. We may call the first to the territory the 'incumbent' and the second to arrive the 'contester'. Consider the social norm that signals to the incumbent to play hawk and to the contester to play dove. Following the social norm, which we may call the *property rights* strategy, is not even a strategy of G, but is a third strategy to the augmented game G^+. Note that if all individuals obey the property rights social norm, then there can be no efficiency losses associated with the allocation of property.

To see that we indeed have a correlated equilibrium, it is sufficient to show that if we add the property rights strategy P to the Hawk-Dove game, then P is a strict best response to itself. With this addition, we get the game depicted in Figure 3. Note that the payoff to property against property, $v/2$, is greater than $3v/4 - w/4$,

which is the payoff to hawk against property, and is also greater than $v/4$, which is the payoff to dove against property. Therefore, property is a strict Nash equilibrium. It is also efficient because there is never a hawk-hawk confrontation in the property correlated equilibrium, and so there is never any injury.

	H	D	P
H	$\frac{v-w}{2},\frac{v-w}{2}$	$v,0$	$\frac{3v-w}{4},\frac{v-w}{4}$
D	$0,v$	$\frac{v}{2},\frac{v}{2}$	$\frac{v}{4},\frac{3v}{4}$
P	$\frac{v-w}{4},\frac{3v-w}{4}$	$\frac{3v}{4},\frac{v}{4}$	$\frac{v}{2},\frac{v}{2}$

Figure 3 **The Hawk-Dove property game**

The property equilibrium is a highly efficient correlated equilibrium G^+ of the Hawk-Dove game G, and corresponds to the classical political economy defense of property rights, but it applies as well to nonhuman territorial animals and explains status quo bias and loss aversion in humans.[4]

3. Nash equilibrium and correlated equilibrium

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 G may have many qualitatively distinct correlated extensions G^+, which implies that life based on social norms can be significantly 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.

Indeed, 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.[5] By contrast, Aumann has shown that Bayesian rationality in a game-theoretic setting is effectively isomorphic with correlated equilibrium.[6] I shall here sketch his argument, which is extremely straightforward, once the proper machinery is set up.

4. Epistemic games

An *epistemic game* 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 G, with payoffs $\pi_i : S \to \mathbf{R}$. In addition, G includes a set of possible states Ω of the game, a *knowledge partition* $'P_i$ of Ω for each player i, and a *subjective prior* $p_i(\bullet;\omega)$ over Ω that is a function of the current state ω. A state ω specifies, possibly among other aspects of the game, the strategy profile s used in the game. We write this as $s = \mathbf{s}(\omega)$. Similarly, we write $s_i = \mathbf{s}_i(\omega)$ and $s_{-i} = \mathbf{s}_{-i}(\omega)$.

The subjective prior $p_i(\bullet;\omega)$ 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. A *partition* of a set X is a set of mutually disjoint subsets of X whose union is X. We write the cell of the knowledge partition $'P_i$ containing the state ω as $\mathbf{P}_i\omega$, and we interpret $\mathbf{P}_i\omega \in {}'P_i$ as the set of states that i considers possible (that is, among which i cannot distinguish) 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 distinguish among states in the cell $\mathbf{P}_i\omega$ of his knowledge partition $'P_i$, his 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 $\omega \in \Omega$.

The possibility operator $\mathbf{P}_i$ has the following two properties: for all $\omega,\omega' \in \Omega$,

(P1) $\omega \in \mathbf{P}_i\omega$
(P2) $\omega' \in \mathbf{P}_i\omega \Rightarrow \mathbf{P}_i\omega' = \mathbf{P}_i\omega$

P1 says that the current state is always possible (that is, $p_i(\omega|\omega) > 0$) and P2 follows from the fact that $'P_i$ is a partition: if $\omega' \in \mathbf{P}_i\omega$, then $\mathbf{P}_i\omega'$ and $\mathbf{P}_i\omega$ have a nonempty intersection, and hence must be identical.

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

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

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

The most important property of the knowledge operator is $\mathbf{K}_iE \subseteq E$, that is, if an agent knows an event E in state ω (that is, $\omega \in \mathbf{K}_iE$), then E is true in state ω (that is, $\omega \in E$). This follows directly from P1.

We can recover the possibility operator $\mathbf{P}_i\omega$ for an individual from his knowledge operator $\mathbf{K}_i$, because

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

To verify this equation, note that if $\omega \in \mathbf{K}_i E$, then $\mathbf{P}_i \omega \subseteq E$, so the left-hand side of the equation is contained in the right-hand side. Moreover, if ω' is not in the right-hand side, then $\omega' \notin E$ for some E with $\omega \in \mathbf{K}_i E$, so $\mathbf{P}_i \omega \subseteq E$ and so $\omega' \notin \mathbf{P}_i \omega$. Thus, the right-hand side of the equation is contained in the left.

If $\mathbf{P}_i$ is a possibility operator for i, the sets $\{\mathbf{P}_i \omega \mid \omega \in \Omega\}$ form a partition $'P$ of Ω. Conversely, any partition $'P$ of Ω gives rise to a possibility operator $\mathbf{P}_i$, two states ω and ω' being in the same cell if $\omega' \in \mathbf{P}_i \omega$. Thus, a knowledge structure can be characterized by its knowledge operator $\mathbf{K}_i$, by its possibility operator $\mathbf{P}_i$, by its partition structure $'P$, or even by the subjective priors $p_i(\bullet; \omega)$.

Since each state ω in epistemic game 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 $\phi_i^\omega \in \Delta S_{-i}$, which allows i to assume that 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 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 defining $\phi_i^\omega([s_{-i}] = p_i([s_{-i}]; \omega)$, where $[s_{-i}] \subset \Omega$ is the event in which 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

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

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

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

5. Example: a simple epistemic game

Suppose Alice and Bob each choose heads (h) or tails (t), neither observing the other's choice. We can write the universe as $\Omega = \{hh, ht, th, tt\}$, where xy means Alice chooses x and Bob chooses y. Alice's knowledge partition is then $'P_A = \{\{hh, ht\}, \{th, tt\}\}$ and Bob's knowledge partition is $'P_B = \{\{hh, th\}, \{ht, tt\}\}$. Alice's possibility operator $\mathbf{P}_A$ satisfies $\mathbf{P}_A hh = \mathbf{P}_A ht = \{hh, ht\}$ and $\mathbf{P}_A th = \mathbf{P}_A tt = \{th, tt\}$, whereas Bob's possibility operator $\mathbf{P}_B$ satisfies $\mathbf{P}_B hh = \mathbf{P}_B th = \{hh, th\}$ and $\mathbf{P}_B ht = \mathbf{P}_B tt = \{ht, tt\}$.

In this case, the event 'Alice chooses h' is $E_A^h = \{hh, ht\}$, and because $\mathbf{P}_A hh$, $\mathbf{P}_A ht \subset E$, Alice knows E_A^h whenever E_A^h occurs (that is, $E_A^h = \mathbf{K}_i E_A^h$). The event E_B^h expressing 'Bob chooses h' is $E_B^h = \{hh, th\}$, and Alice does not know E_B^h because at th Alice believes tt is possible, but tt $\notin E_B^h$.

6. Correlated strategies and correlated equilibria

We want to show that if players are Bayesian rational in an epistemic game G and have a common prior over Ω, the strategy profiles $\mathbf{s}:\Omega \to S$ that they play form a correlated equilibrium.[7] The converse also holds: for every correlated equilibrium of a game, there is an extension to an epistemic game G with a common prior $p \in \Omega$ such that in every state ω it is rational for all players to carry out the move indicated by the correlated equilibrium.

Informally, a correlated equilibrium of an epistemic game G is a Nash equilibrium of a game G^+, which is G augmented by an initial move by Nature, which 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 Nature's directive is a best response, if other players also follow Nature's directives, provided players have the *common prior p*.

Formally, a *correlated strategy* of epistemic game 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 choreographer who 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 choreographer 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*.

Suppose $f^1, \ldots, f^k$ are correlated strategies and let $a = (a_1, \ldots, a_k)$ be a lottery (that is, $a_i \geq 0$ and $\Sigma_i a_i = 1$). Then, $f = \Sigma_i a_i f^i$ is also a correlated strategy defined on $\{1, \ldots, k\} \times \Gamma$. We call such an f a *convex sum* of $f^1, \ldots, f^k$. Any convex sum of correlated strategies is clearly a correlated strategy. It follows that any convex sum of correlated distributions is itself a correlated distribution.

Suppose $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a Nash equilibrium of a game G, where for each $i = 1, \ldots n$,

$$\sigma_i = \sum_{k=1}^{n_i} a_{ki} s_{ki}$$

(where n_i is the number of pure strategies in S_i and a_{ki} is the weight given by σ_i on the kth pure strategy $s_{ki} \in S_i$). Note that σ thus defines a probability distribution $\tilde{p}$ on S such that $\tilde{p}(s)$ is the probability that pure strategy profile $s \in S$ will be chosen when mixed strategy profile σ is played. Then, $\tilde{p}$ is a correlated distribution of an epistemic game associated with G, which we will call G as well. To see this, define Γ_i as a set with n_i elements $\{\gamma_{1i}, \ldots, \gamma_{n_i i}\}$ and define $p_i \in \Delta S_i$ such that it places probability a_{ki} on γ_{ki}. Then, for $s = (s_1, \ldots, s_n) \in S$, define $p(s) = \prod_{i=1}^n p_i S_i$. Now, define $\Gamma = \prod_{i=1}^n \Gamma_i$ and let $f : \Gamma \to S$ be given by $f(\gamma_{k_1 1}, \ldots, \gamma_{k_n n}) = (s_{k_1 1}, \ldots, s_{k_n n})$. It is easy to check that f is a correlated strategy with correlated distribution $\tilde{p}$. In short, every Nash equilibrium is a correlated strategy, and hence any convex combination of Nash equilibria is a correlated strategy.

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 $f_i(\gamma)$ if 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 choreographer. Note that a correlated equilibrium of G is a Nash equilibrium of the game generated from G by adding Nature, 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 Nature's recommendation, provided that the other players comply as well.

7. Correlated equilibrium and Bayesian rationality

We now show that if the players in epistemic game G are Bayesian rational at ω and have a common prior $p(\bullet; \omega)$ in state ω and that if 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 $\omega \in \Omega$.

To prove this theorem, we identify the state space for the correlated strategy with the state space Ω of G, and the probability distribution on the state space with the common prior p. We then define the correlated strategy $f: \Omega \to S$ by setting $f(\omega) = (\mathbf{s}_1(\omega), \ldots, \mathbf{s}_n(\omega))$, where $\mathbf{s}_i(\omega)$ is i's choice in state ω. Then, for any player i and any function $g_i: \Omega \to S_i$ that is $'P_i$-measurable (that is, constant on the cells of the partition $'P_i$), because i is Bayesian rational, we have

$$\mathbf{E}[\pi_i \ (\mathbf{s}(\omega)) | \omega] \geq \mathbf{E}[\pi_i(\mathbf{s}_{-i}(\omega), g_i(\omega)) | \omega].$$

Now, multiply both sides of this inequality by $p(P)$ and add over the disjoint cells $P \in 'P_i$, which gives, for any such g_i,

$$\mathbf{E}[\pi_i \ (\mathbf{s}(\omega))] \geq \mathbf{E}[\pi_i(\mathbf{s}_{-i}(\omega), g_i(\omega))].$$

This proves that $(\Omega, f(\omega))$ is a correlated equilibrium. Note that the converse clearly holds as well.

8. Common priors and social norm equilibria

The isomorphism between correlated equilibrium distributions and Bayesian rationality demonstrated in Section 7 highlights an assumption that lies at the heart of a game-theoretic concept of social norms. This is the requirement that the players have a common prior over the state space Ω. If the correlated equilibrium assigns a strict best response to each player, it is clear that some amount of preference heterogeneity will not destroy the equilibrium (the reader is invited to verify this). Moreover, if there are known 'types' of players (for example, optimists and pessimists) whose priors are distinct, but commonly known and the population composition is commonly known, it may be possible to redefine the state space so that there are common priors over the new state space, to which the correlated equilibrium theory then applies. The reader is invited to develop this theme.

However, when common priors are lacking and the actual composition and frequency distribution of priors are not held in common for some suitably enlarged state space, the social norm analysis will fail to apply. Rational agents with fundamental disagreements as to the actual structure of their social life do not dance to a choreographer's instructions.

9. The omniscient choreographer and social preferences

The isomorphism between correlated equilibrium distributions and Bayesian rationality also requires that the choreographer be *omniscient* in the sense of having a knowledge partition that is at least as fine as each of the player's knowledge partition. This latter requirement was not explicitly mentioned in the proof, but is implicit in the requirement that $f(\omega) = \mathbf{s}(\omega)$ for all $\omega \in \Omega$.

When this assumption fails, a correlated equilibrium may still obtain, provided the players have sufficiently strong pro-social preferences. Despite the fact that we have placed no restrictions on preferences other than Bayesian rationality, many modelers of social norms, including Bicchieri,[8] predicate their analysis on the fact that rational individuals may have other-regarding preferences or may value certain moral virtues so that they voluntarily conform to a social norm in a situation where a perfectly self-regarding and amoral agent would not. In such cases, the choreographer may be obeyed even at a cost to the players, provided that the cost of doing so is not excessive.

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 idiosyncrasies of the agent and is unknown to the choreographer. Suppose the maximum size of the private component in any state for an agent is α, but the agent's inclination to follow the choreographer has a 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* toward 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.[9]

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. For instance, 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 (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 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.

The following is a more complex example of a social norm equilibrium that requires a normative predisposition of honesty.

10. A reputational model of honesty and corruption

Consider a society in which sometimes people are needy, and sometimes others help the needy. In the first period, a pair of members is selected randomly, one of the pair being designated 'Needy' and the other 'Giver'. Giver and Needy then play a game G in which if Giver helps, a benefit b is conferred on Needy at a cost c to Giver, where $0 < c < b$. Alternatively, if Giver defects, both players receive zero. In each succeeding period, Needy from the previous period becomes Giver in the current period. Giver is then paired with a new, random Needy, and the game G is played by the new pair. If we assume that helping behavior is common knowledge, there is a Nash equilibrium of the following form, provided the discount factor δ is sufficiently close to unity. At the start of the game, each player is labeled 'in good standing'. In every period Giver helps if and only if his partner Needy is in good standing. Failure to do so puts a player 'in bad standing', where he remains for the rest of the game.

To see that this is a Nash equilibrium in which every Giver helps in every peri-

od for δ sufficiently close to one, let v_c be the present value of the game to a Giver and let v_b be the present value of the game for an individual who is not currently Giver or Needy. Then we have $v_c = -c + \delta v_b$ and $v_b = p(b + \delta v_c) + (1 - p)\delta v_b$, where p is the probability of being chosen as Needy. The first equation reflects the fact that a Giver must pay c now and becomes a candidate for Needy in the next period. The second equation expresses the fact that a candidate for Needy is chosen with probability p and then gets b, plus is Giver in the next period, and with a probability of $1 - p$ remains a candidate for Needy in the next period. If we solve these two equations simultaneously, we find that $v_c > 0$ precisely when $\delta > c/(c + p(b - c))$. Because the right-hand side of this expression is strictly less than one, there is a range of discount factors for which it is a best response for a Giver to help, and thus remain in good standing.

Suppose, however, the informational assumption is that each new Giver knows only whether his partner Needy did or did not help his own partner in the previous period. If Alice is Giver and her partner Needy is Bob, and Bob did not help when he was Giver, it could be because when he was Giver, Carole, his Needy partner, had defected when she was Giver or because Bob failed to help Carole even though she had helped Donald, her previous Needy partner, when she was Giver. Because Alice cannot condition her action on Bob's previous action, Bob's best response is to defect on Carole, no matter what she did. Therefore, Carole will defect on Donald, no matter what he did. Thus, there can be no Nash equilibrium with the pure strategy of helping.

This argument extends to the richer informational structure where a Giver knows the previous k actions for any finite k. Here is the argument for $k = 2$, which the reader is encouraged to generalize. Suppose the last five players are Alice, Bob, Carole, Donald, and Eloise, in that order. Alice can condition her choice on the actions taken by Bob, Carole, and Donald, but not on Eloise's action. Therefore, Bob's best response to Carole will not be conditioned on Eloise's action, and hence Carole's response to Donald will not be conditioned on Eloise's action. So, finally, Donald's response to Eloise will not be conditioned on her action, so her best response is to defect when she is Giver. Thus, there is no helping Nash equilibrium.

Suppose, however, back in the $k = 1$ case, that instead of defecting unconditionally when facing a Needy who has defected improperly, a Giver helps with probability $p = 1 - c/b$ and defects with probability $1 - p$. The gain from helping unconditionally is then $b - c$, while the gain from following this new strategy is $p(b - c) + (1 - p)pb$, where the first term is the probability p of helping times the reward b in the next period if one helps minus the cost c of helping in the current period and where the second term is the probability $1 - p$ of defecting times the probability p that you will be helped anyway when your are Needy times the benefit b. Equating this expression with $b - c$, the cost of helping unconditionally, we get $p = 1 - c/b$, which is a number strictly between zero and one and hence a valid probability.

Consider the following strategy. In each round, Giver helps if his partner helped in the previous period, and otherwise helps with probability p and defects with probability $1 - p$. With this strategy each Giver i is indifferent to helping or defecting, because helping costs i the amount c when he is Giver, but i gains b when he is Needy, for a net gain of $b - c$. However, defecting costs zero when Giver, but gives $bp = b - c$ when he is Needy. Because the two actions have the same payoff, it is incentive compatible for each Giver to help when his partner Needy helped and to defect with probability p otherwise. This strategy thus gives rise to a Nash equilibrium with helping in every period.[10]

However, there is no reason for rational self-regarding players to implement this Nash equilibrium. To see this, note that each player chooses a totally mixed strategy as Giver, and hence is indifferent between helping and defecting. Therefore, Givers have no incentive to play the equilibrium strategy. Erecting a social norm does not improve the situation because a Giver who is instructed by the choreography to help or to defect has no incentive to follow the instruction. However, if we add an $\epsilon > 0$ of normative predisposition, an omniscient choreographer could implement this Nash equilibrium by acting as the appropriate randomizing device. Moreover, suppose Givers have private preferences that, for instance, favor some players (for example, friends or coreligionists) over others (for example, enemies or infidels). In this case, the choreographer's instructions will be followed only if players have a commitment to norm following that is greater than their personal preferences to give or withhold aid to particular individuals.

11. Conclusion

In the first pages of *The Grammar of Society*, Cristina Bicchieri asserts that 'social norms . . . *transform* mixed-motive games into coordination ones'.[11] As we have seen in Section 2, this transformation is not always the case, but Bicchieri's affirmation is generally on the mark, and indeed, as shown in this article, is the key to understanding the relationship between rational choice theory and social norms. Section 7 developed the central principle[12] that every state of an epistemic game G in which players are rational can be implemented as a correlated equilibrium distribution, provided the appropriate epistemic conditions hold (common priors and choreographer omniscience). The associated correlated equilibrium is indeed a Nash equilibrium of an augmented game G^+, in which an additional player, the choreographer (aka social norm) who implements the equilibrium, is added.

I believe that the epistemic game theoretic analysis of social norms presented in this article can serve as the theoretical core for a general social theory of human strategic interaction. This analysis shows precisely where classical game theory goes wrong: it focuses on Nash as opposed to correlated equilibria, and hence ignores 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, con-

ditions 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.

notes

The material is adapted from Herbert Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences* (Princeton, NJ: Princeton University Press, 2009).

1. David Lewis, *Conventions: A Philosophical Study* (Cambridge, MA: Harvard University Press, 1969); Michael Taylor, *Anarchy and Cooperation* (London: John Wiley, 1976); Michael Taylor, *Community, Anarchy, and Liberty* (Cambridge: Cambridge University Press, 1982); Michael Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987); Robert Sugden, *The Economics of Rights, Co-operation and Welfare* (Oxford: Blackwell, 1986); Robert Sugden, 'Spontaneous Order', *Journal of Economic Perspectives* 3–4 (1989): 85–97; Cristina Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge: Cambridge University Press, 2006); Kenneth G. Binmore, *Game Theory and the Social Contract: Playing Fair* (Cambridge, MA: MIT Press, 1993); Kenneth G. Binmore, *Game Theory and the Social Contract: Just Playing* (Cambridge, MA: MIT Press, 1998); Kenneth G. Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005).
2. Robert J. Aumann, 'Subjectivity and Correlation in Randomizing Strategies', *Journal of Mathematical Economics* 1 (1974): 67–96; Robert J. Aumann, 'Correlated Equilibrium and an Expression of Bayesian Rationality', *Econometrica* 55 (1987): 1–18.
3. Herbert Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences* (Princeton, NJ: Princeton University Press, 2009).
4. Herbert Gintis, 'The Evolution of Private Property', *Journal of Economic Behavior and Organization* 64 (2007): 1–16.
5. Robert J. Aumann and Adam Brandenburger, 'Epistemic Conditions for Nash Equilibrium', *Econometrica* 65 (1995): 1161–80; Kaushik Basu, 'The Traveler's Dilemma: Paradoxes of Rationality in Game Theory', *American Economic Review* 84 (1994): 391–5; Gintis, *The Bounds of Reason*.
6. Aumann, 'Correlated Equilibrium and an Expression of Bayesian Rationality'.
7. Ibid.
8. Bicchieri, *The Grammar of Society*.
9. Herbert Gintis, 'The Hitchhiker's Guide to Altruism: Genes, Culture, and the Internalization of Norms', *Journal of Theoretical Biology* 220 (2003): 407–18.
10. V. Bhaskar, 'Noisy Communication and the Evolution of Cooperation', *Journal of Economic Theory* 82 (1998): 110–31.
11. Bicchieri, *The Grammar of Society*.
12. Aumann, 'Correlated Equilibrium and an Expression of Bayesian Rationality'.