
Bayesian Rationality and Social Epistemology

Social life comes from a double source, the likeness of consciences and the division of social labor.

Emile Durkheim

There is no such thing as society. There are individual men and women, and there are families

Margaret Thatcher

At least since Schelling (1960) and Lewis (1969), game theorists have interpreted social norms as Nash equilibria. More recent contributions based upon the idea of social norms as selecting among Nash equilibria include Binmore (2005) and Bicchieri (2006). There are two problems with this approach. The first is that the conditions under which rational individuals play a Nash equilibrium are extremely demanding (theorem §8.4), and are not guaranteed to hold simply because there is a social norm specifying a particular Nash equilibrium. Second, the most important and obvious social norms do not specify Nash equilibria at all, but rather are devices that implement *correlated equilibria* (§2.11, §7.5).

Informally, a correlated equilibrium of an epistemic game \mathcal{G} is a Nash equilibrium of a game \mathcal{G}^+ , in which \mathcal{G} is augmented by an initial move by a new player, whom we call the *choreographer*, 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 choreographer’s directive is a best response for each player, if other players also follow the choreographer’s directives.

This chapter uses epistemic game theory to expand on the notion of social norms as choreographer of a correlated equilibrium, and to elucidate the socio-psychological prerequisites for the notion that social norms implement correlated equilibria.

The correlated equilibrium is a much more natural equilibrium criterion than the Nash equilibrium, because of a famous theorem of Aumann (1987), who showed that Bayesian rational agents in an epistemic game \mathcal{G} with a common subjective prior play a correlated equilibrium of \mathcal{G} (§2.11-2.13). Thus, while rationality and common priors do not imply Nash equilibrium,

these assumptions do imply correlated equilibrium and as we shall see, social norms act not only as choreographer, but also supply the epistemic conditions for common priors.

In a correlated equilibrium, rational players have no incentive to deviate from the instructions of the choreographer, but if the correlated equilibrium involves multiple strategies with equal payoffs, they may have no incentive to follow them either. If a correlated equilibrium can be purified (see Chapter 6), each agent effectively has a strict preference to follow the directives of the choreographer. However, in most complex games purification will fail (§6.3, §6.4), in which case, as we shall see, we must assume that agents have a *normative predisposition* towards following the choreographer's instructions unless they have alternatives with strictly higher payoffs.

The isomorphism between correlated equilibrium and Bayesian rationality with common priors assumes that the choreographer has at least as much information as any player. This means that all information is *public*, an assumption that is violated in many practical cases. 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 strength greater than α . Then, the agent will continue 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 payoff no more than α greater than 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 2003a).

7.1 From Battle to Ballet of the Sexes

Suppose there is a social norm specifying that when choosing between opera and gambling, the male of the pair decides on Monday through Friday, and the female on the weekend. This norm choreographs a correlated equilibrium in which Alfredo and Violetta go to the opera if and only if it

is a weekend. Assuming that their planned meeting occurs equally likely on each day of the week, Alfredo's payoff is $2(5/7) + 1(2/7) = 12/7$ and Violetta's is $1(5/7) + 2(2/7) = 9/7$. This correlated equilibrium is not a Nash equilibrium of the underlying game and, like the pure-strategy Nash equilibria of the game, it is Pareto-efficient.

7.2 The Choreographer Trumps Backward Induction

Suppose Alice and Bob play the 100-round repeated Prisoner's Dilemma under conditions of common knowledge of rationality. They thus defect on every round (§5.12). They then discover a choreographer who chooses a number k , with $1 \leq k \leq 99$, and with probability $1/2$ advises Alice to cooperate up to the k^{th} round and Bob to cooperate up to the $1 + k^{\text{th}}$ round, and with probability $1/2$, reverses the advice to Alice and Bob. Both players are advised to defect forever after the first defection.

Assuming that both Alice and Bob believe that each has probability $\theta(k) = 1/2$ of having the lower number when advised to defect on the k^{th} round, we can show that this is a correlated equilibrium with cooperation up to round $k-1$. Suppose Bob takes the choreographer's advice, cooperating up to the suggested round, and then defecting thereafter. Then, Alice's payoff from cooperating, assuming the payoffs to the prisoner's dilemma stage game are $t > r > p > s$ (corresponding to $4 > 3 > 1 > 0$ in §5.6), is given by

$$\begin{aligned} & \frac{1}{2} [r(k-1) + t + (n-k)p] + \frac{1}{2} [r(k-2) + s + (n-k+1)p] \\ & = r(k-2) + \frac{s+t+p+r}{2} + (n-k)p. \end{aligned}$$

If Alice disobeys the choreographer, she can only possibly gain by defecting either one or two rounds earlier. The payoff to defecting on round $k-1$ is

$$\begin{aligned} & \frac{1}{2} [r(k-2) + t + (n-k+1)p] + \frac{1}{2} [r(k-2) + (n-k+2)p] \\ & = r(k-2) + \frac{t}{2} + (n-k+1)p. \end{aligned}$$

The payoff to obeying the choreographer rather than defecting one round earlier is thus $(r+s-p)/2 > 0$. If Alice defects two rounds earlier, her payoff is $r(k-3) + t + (n-k+2)p$, which is less than obeying, provided

$r - p > (t - s)/4$. Thus, given this inequality (which clearly holds for the game in §5.6), we have a correlated equilibrium. If k is large, this correlated equilibrium has a high payoff, despite CKR.

7.3 The Bourgeois Equilibrium of the Hawk-Dove Game

The hawk-dove game (§2.9) is an inefficient way to allocate property rights, especially if the cost of injury w is not much larger than the value v of the property. To see this, note that players choose hawk with probability v/w , and you can check that the ratio of the payoff to the efficient payoff $v/2$ is

$$1 - \frac{v}{w}.$$

When w is near v , this is close to zero.

The hawk-dove game is thus a beautiful example of the Hobbesian state of nature, where life is nasty, brutish, and short (Hobbes 1968[1651]). However, suppose some members of the population institute a social norm respecting private property, based on the fact that whenever two players have a property dispute, one of them must have gotten there first, and the other must have come later. We may call the former the “incumbent” and the latter the “contester.”

The new strategy, B , called the “bourgeois” strategy, always plays hawk when incumbent and dove when contester. When we add B to the normal form matrix of the game, we get the *hawk, dove, bourgeois game* depicted in figure 7.1. Note that the payoff to bourgeois against bourgeois, $v/2$ is greater than $3v/4 - w/4$, which is the payoff to hawk against bourgeois, and is also greater than $v/4$, which is the payoff to dove against bourgeois. Therefore, bourgeois is a strict Nash equilibrium. It is also efficient, because there is never a hawk-hawk confrontation in the bourgeois equilibrium, so there is never any injury.

The bourgeois strategy is not a Nash equilibrium of the hawk-dove game, but is a correlated equilibrium of the larger social system with the property norm. This example will be elaborated upon in chapter 11.

7.4 Convention as Correlated Equilibrium

The town of Pleasantville has one traffic intersection, one road going north-south and the other east-west. If an east-west car meets a north-south car at the intersection and both stop, one is randomly chosen to go first across

	H	D	B
H	$(v - w)/2$	v	$3v/4 - w/4$
D	0	$v/2$	$v/4$
B	$(v - w)/4$	$3v/4$	$v/2$

Figure 7.1. The hawks-dove-bourgeois game

the intersection and the second follows, with an average loss of time of one second each. If one car stops and the other goes, only the car that stopped will lose a second. If they both go, however, they may crash, so there is an expected loss of $c > 1$ for each.

There is clearly a unique symmetrical Nash equilibrium to this game, in which each car goes with probability $\alpha = 1/c$ and the expected payoff to each player is -1 ; that is, they do no better than both waiting. However, there is an obvious social norm in which one car, say the east-west car, always goes, and north-sound car always waits. This is now a correlated equilibrium that implements an asymmetric Nash equilibrium of the underlying game.

With these examples in mind, we can tackle the underlying theory.

7.5 Correlated Strategies and Correlated Equilibria

We will use epistemic game theory (§6.5) to show that if players are Bayesian rational in an epistemic game \mathcal{G} and have a common prior over Ω , the strategy profiles $\mathbf{s}:\Omega \rightarrow S$ that they play form a correlated equilibrium (Aumann 1987). The converse also holds: for every correlated equilibrium of a game, there is an extension to an epistemic game \mathcal{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 \mathcal{G} is a Nash equilibrium of a game \mathcal{G}^+ , which is \mathcal{G} augmented by an initial move by “Nature,” 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 Nature’s directive is a best response, if other players also follow Nature’s directives, provided players have the common prior p .

The intuition behind the theorem is that in an epistemic game, the state space Ω includes all information concerning the players' actions, so common priors imply that all agents agree as to the probability distributions over the actions they will take. Hence, assuming each agent i has a single best response $s_i(\omega)$ in every state ω (i.e., the equilibrium is a strict correlated equilibrium), the move of each player is known to each other, and because the agents are rational, each must then play a best response to the actions of the others.

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 $f : \Gamma \rightarrow S$. 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 $s \in S$, $\tilde{p}(s) = p(\{\gamma \in \Gamma \mid f(\gamma) = s\})$ 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, \dots, f^k are correlated strategies, and let $\alpha = (\alpha_1, \dots, \alpha_k)$ be a lottery (i.e., $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$). Then $f = \sum_i \alpha_i f^i$ is also a correlated strategy defined on $\{1, \dots, k\} \times \Gamma$. We call such an f a *convex sum* of f^1, \dots, 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 $\sigma = (\sigma_1, \dots, \sigma_n)$ is a Nash equilibrium of a game \mathcal{G} , where for each $i = 1, \dots, n$,

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

where n_i is the number of pure strategies in S_i , and α_{ki} is the weight given by σ_i on the k^{th} 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 associate with \mathcal{G} , which we will call \mathcal{G} as well. To see this, define Γ_i as a set with n_i elements $\{\gamma_{1i}, \dots, \gamma_{n_i i}\}$ and define $p_i \in \Delta S_i$ that places probability α_{ki} on γ_{ki} . Then, for $s = (s_1, \dots, s_n) \in S$, define $p(s) = \prod_{i=1}^n p_i(s_i)$. Now define $\Gamma = \prod_{i=1}^n \Gamma_i$, and let $f : \Gamma \rightarrow S$ be given by $f(\gamma_{k_1 1}, \dots, \gamma_{k_n n}) = (s_{k_1 1}, \dots, s_{k_n n})$. 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 $g_i : \Gamma \rightarrow S_i$ 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 f_i . We say that a correlated strategy f is a *correlated equilibrium* if for each player i , and any $g_i : \Gamma \rightarrow S_i$ 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 \mathcal{G} is a Nash equilibrium of the game generated from \mathcal{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.

7.6 Correlated Equilibrium and Bayesian Rationality

THEOREM 7.1 *If the players in epistemic game \mathcal{G} are Bayesian rational, have a common prior p , and each player i chooses $\mathbf{s}_i(\omega) \in S_i$ in state ω , then the correlated distribution of $\mathbf{s} = (\mathbf{s}_1, \dots, \mathbf{s}_n)$ is a correlated equilibrium distribution given by correlated strategy 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 \mathcal{G} , and the probability distribution on the state space with the common prior p . We then define the correlated strategy $f : \Omega \rightarrow S$ by setting $f(\omega) = (\mathbf{s}_1(\omega), \dots, \mathbf{s}_n(\omega))$, where $\mathbf{s}_i(\omega)$ is i 's choice in state ω (§6.5). Then, for any player i and any function $g_i : \Omega \rightarrow S_i$ that is \mathcal{P}_i -measurable (i.e., that is constant on cells of the partition \mathcal{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 $P \in \mathcal{P}_i$, 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. ■

7.7 The Social Epistemology of Common Priors

“De gustibus,” we are told, “non est disputandum.” The decision theorist does not question the origin, or content of preferences. The Savage axioms, for instance, assume only a few highly general regularities of choice behavior (§1.5). However, we have seen that when we move from individual decision theory to epistemic game theory, this vaunted tolerance is no longer tenable. In its place we require common priors and sometimes, as we shall see, even common knowledge of conjectures (§8.1).

Common priors must be the result of a common process of belief formation. The subjectivist interpretation of probability (di Finetti 1974, Jaynes 2003) is persuasive as a model of human behavior, but is a partial view because it cannot explain why individuals agree on certain probabilities.¹

For standard explanations of common priors, we must turn to the frequency theories of von Mises (1981) and others or the closely related propensity theory of Popper (1959), which interpret probability as the long run frequency of an event, or its propensity to occur at a certain rate. John Harsanyi (1967) has been perhaps the most eloquent proponent of this approach among game theorists, promoting the *Harsanyi doctrine*, which states that all differences in probability assessments among rational individuals must be due to differences in the information they have received. In fact, however, the Harsanyi doctrine applies only under a highly restricted set of circumstances.

Despite lack of agreement concerning the philosophical grounding of probabilistic statements (von Mises 1981, de Laplace 1996, Gillies 2000, Keynes 2004), there is little disagreement concerning the mathematical

¹Savage’s axiom A3 (see p. 15) suggests that there is something supra-individual about probabilities. This axiom says that the probability associated with an event must not depend on the desirability of the payoff contingent upon the event occurring. There is no reason why this should be the case unless there is some supra-individual standard for assessing probabilities.

laws of probability (Kolmogorov 1950). Moreover, modern science is public and objective: except at the very cutting edge of research, there is broad agreement among scientists, however much they differ in creed, culture, or personal predilections.

This line of reasoning suggests that 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. 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 if one member knows N , then he knows that each other member knows N . For instance, if i and j are both looking at the same yellow ball, each sees the other looking at the ball, each knows the other has normal vision and 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, in a joint attentive state, the natural occurrence is mutually accessible (Tomasello 1999, Lorini, Tummolini and Herzig 2005).

When we add to this sense data the possibility of joint attentive states of symmetric reasoners (§7.8), common knowledge of natural occurrences becomes plausible (Theorem 7.2).

But, 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 supra-sensory reality, do not fall into this category (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, *H. sapiens*, 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). Thus, if agent i believes something, and if i knows that he shares certain environmental experiences with agent j , then i knows that j probably 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 neuro-psychological literature on how minds know other minds deals with mirror neurons, the human prefrontal lobe, and other brain mecha-

nisms that facilitate the sharing of knowledge and belief. From the viewpoint of modeling human behavior, these facilitating mechanisms must be translated into axiomatic principles of strategic interaction. This is a huge step to take for game theory, which has never provided a criterion *of any sort* for knowledge or belief.

7.8 The Social Epistemology of Common Knowledge

We have seen that we must add to Bayesian rationality the principle of normative predisposition to have a social epistemology sufficient to assert that rational agents with common priors, in the presence of the appropriate choreographer, will choose to play a correlated equilibrium. We now study the cognitive properties of agents sufficient to assert that social norms can foster common priors.

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 an event N that specifies a social event 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 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, with a light saying “taxi” on the roof.

THEOREM 7.2 *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, $\omega \in \mathbf{K}_i\mathbf{K}_jN$, and because i is a symmetrical reasoner with respect to j , $\omega \in \mathbf{K}_i\mathbf{K}_jE$. Thus we have $\mathbf{P}_i\omega \subseteq \mathbf{K}_jE$ for all i, j (the case $i = j$ holding trivially). Thus, N is an indicator of \mathbf{K}_jE for all j . Applying the above reasoning to indicator \mathbf{K}_kE , we see that $\omega \in \mathbf{K}_i\mathbf{K}_j\mathbf{K}_kE$ for all i, j , and k . All higher levels of mutual knowledge are obtained similarly, proving common knowledge. ■

COROLLARY 7.2.1 *Suppose N is a mutually accessible natural occurrence for a group of Bayesian rational symmetric reasoners. Then N is common knowledge.*

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 7.2. ■

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. Communality of knowledge should always be derived from more elementary psychological and social regularities.

7.9 Social Norms

We say an event E is *norm-governed* if there is a *social norm* $\mathcal{N}(E)$ that specifies *socially appropriate behavior* $\mathcal{N}(E) \subseteq S$, $\mathcal{N}(E) \neq \emptyset$, where S is the strategy profile set for an epistemic game. Note that we allow appropriate behavior to be correlated. How does common knowledge of a social situation E affect the play of a game \mathcal{G} ? The answer is that each player i must associate a particular social norm $\mathcal{N}(E)$ with E that determines appropriate behavior in the game, i must be confident that other players also associate $\mathcal{N}(E)$ with E , i must expect that others will choose to behave appropriately according to $\mathcal{N}(E)$, and behaving appropriately must be a best response for i , given all of the above.

Suppose E indicates $\mathcal{N}(E)$ for players because the players belong to a society in which common culture specifies that when a game \mathcal{G} is played and event E occurs, then appropriate behavior is given by $\mathcal{N}(E)$. Suppose players are symmetric reasoners with respect to E . Then, similar reasoning to Theorem 7.2 shows that $\mathcal{N}(E)$ is common knowledge. We then have

THEOREM 7.3 *Given epistemic game \mathcal{G} with normatively predisposed players who are symmetric reasoners, suppose E is an indicator of social norm $\mathcal{N}(E)$. Then, if appropriate behavior according to $\mathcal{N}(E)$ is a correlated equilibrium for \mathcal{G} , the players will choose the corresponding correlated strategies.*

7.10 Game Theory and the Evolution of Norms

Social norms cannot be explained as the product of the interaction of Bayesian rational agents. Rather, as developed in Chapter 12, social norms are explained by sociobiological models of gene-culture coevolution (Cavalli-Sforza and Feldman 1973, Boyd and Richerson 1985). Humans have evolved psychological predispositions that render social norms effective. Social evolution (Cavalli-Sforza and Feldman 1981, Dunbar 1993, Richerson and Boyd 2004) has favored both the emergence of social norms and of human predispositions to follow social norms, to embrace common priors and recognize common knowledge of many events. Nor is this process limited to humans, as the study of territoriality in various non-human species makes clear (Gintis 2007b). The culmination of this process is a pattern of human attributes that can likely be subjected to axiomatic formulation much as we have done with the Savage axioms.

The notion of a social norm as a choreographer is only the first step in analyzing social norms—the step articulating the linkage between Bayesian rationality and game theory on the one hand, and macro-social institutions and their evolution on the other. We can add the dimension of coercion to the concept of social norm, by attaching rewards and punishments to behaviors based on their relationship to socially approved behavior. We can also treat social norms as strongly prosocial under some conditions, meaning that individuals prefer to follow these norms even when it is not in their personal interest to do so, provided others do the same (α -normative predisposition). Finally, we may be able to use a model of the social norm linkage to develop a theory of the evolution of norms, a task initiated by Binmore (1993, 1998, 2005).

7.11 The Merchants’ Wares

Consider the coordination game \mathcal{G} with normal form matrix shown to the right. There are two pure strategy equilibria: (2,2) and (1,1). There is also a mixed-strategy equilibrium with payoffs (1/3,1/3), in which players choose s_1 with probability 1/3. There is no principle of Bayesian rationality that would lead the players to coordinate on the higher-payoff, or any other, Nash equilibrium.

	s_1	s_2
s_1	2,2	0,0
s_2	0,0	1,1

There are two obvious social norms in this case: “when participating in a pure coordination game, choose the strategy that gives players the maximum (respectively, minimum) common payoff.” The following is a plausible social norm that leads to a convex combination of the two coordinated payoffs.

There are two neighboring tribes whose members produce and trade apples and nuts. Members of one tribe wear long gowns, while members of the other tribe wear short gowns. Individuals indicate a willingness to trade by visually presenting their wares, the quality of which is either 1 or 2, known prior to exchange to the seller but not the buyer. After exchanging goods, both parties must be satisfied or the goods are restored to their original owner and no trade is consummated. The social norm \mathcal{N} that governs exchange is “never try to cheat a member of your own tribe, and always try to cheat a member of the other tribe.”

We assume that, when two individuals meet, the visibility of their wares, F , represents a mutually accessible natural occurrence that is an indicator that the appropriate social norm is \mathcal{N} , conforming to which is a best

response for both parties. If both have long robes or short robes, \mathcal{N} indicates that they each give full value 2, while if their robe styles differ, each attempts to cheat the other by giving partial value 1. Because all individuals have a normative predisposition and this is common knowledge, each follows the norm, as either trade is better than no trade. A trade is thus consummated in either case. The expected payoff to each player is $2p + (1 - p)$, where p is the probability of meeting a trader from one's own tribe.

This analysis illustrates several key points. First, as we will see in §8.1, there are very straightforward sufficient conditions for two rational agents to play an Nash equilibrium (Theorem 8.2): mutual knowledge of rationality, mutual knowledge of the game and its payoffs, and mutual knowledge of conjectures (what the other player will choose).

In The Merchants' Wares problem it is clear that the problem is: where does mutual knowledge of conjectures come from? As I have stressed, there is nothing in the theory of rational choice that permits us to conclude that two rational agents share beliefs concerning each other's beliefs about each other's beliefs. Rather, each player forms conjectures concerning the other's likely behavior, and the other's likely conjectures, from his *social knowledge*—in this case, from the common *mores* of the two tribes, and from mutual knowledge that each knows the *mores* of the two tribes.

The conclusion is that there is no reason to posit that rational agents will choose the Pareto-superior equilibrium, because we have seen that sometimes they will not. It is not *reason* but *humanity* that leads us to believe that the Pareto-superior equilibrium is "obvious." We humans, by virtue of our gene-culture coevolutionary history and our civilized culture, harbor a "default frame" that says "in a coordination game, unless you have some special information that suggests otherwise, conjecture that the other player also considers the frame to be a default frame and reasons as you do, and choose the action that assumes your partner is trying to do well by you." We will return to this point in chapter 12, where we locate it as part of an evolutionary epistemology.