Rationality and Common Knowledge

Herbert Gintis*

September 10, 2009

1 Introduction

Interactive epistemology is the study of the distribution of knowledge among rational agents, using modal logic in the tradition of Hintikka (1962) and Kripke (1963), and agent rationality based on the rational actor model of economic theory, in the tradition of Von Neumann and Morgenstern (1944) and Savage (1954). *Epistemic game theory*, which is interactive epistemology adjoined to classical game theory (Aumann, 1987, 1995), has demonstrated an intimate relationship between rationality and correlated equilibrium (Aumann 1987, Brandenburger and Dekel 1987), and has permitted the rigorous specification of the conditions under which rational agents will play a Nash equilibrium (Aumann and Brandenburger 1995).

A central finding in this research is that rational agents use the strategies suggested by game-theoretic equilibrium concepts when there is a *communality of knowledge* in the form of common probability distributions over the stochastic variables that arise in the play of the game (so-called *common priors*), and *common knowledge* of key aspects of the strategic interaction. We say an event E is *common knowledge* for agents $i = 1, \ldots, n$ if the each agent i knows E, each i knows that each agent j knows E, each agent k knows that each j knows that each i knows E, and so on (Lewis 1969, Aumann 1976).

Specifying when individuals share knowledge and when they know that they share knowledge is among the most challenging of epistemological issues. Contemporary psychological research has shown that these issues cannot be resolved by analyzing the general features of high-level cognitive functioning alone, but in fact concern the particular organization of the human brain. Humans have a *theory*

*Santa Fe Institute and Central European University. To appear in *Rationality and Society*. Thanks to the Hungarian Scientific Research Fund (OTKA) and the Massachusetts Research Foundation for financial support. Some of the material in this paper is adapted from Herbert Gintis, *The Bounds of Reason* (Princeton University Press, 2009).

of mind that is possessed in extremely rudimentary form by other primate species and not at all by most other species, even those with complex forms of social organization (Premack and Woodruff 1978, Heyes 1998, Tomasello et al. 2005). This mental capacity permits us to attribute mental states to others, and to assess when other share our beliefs, intentions, and goals (Baron-Cohen 1991, 1995).

Epistemic game theory has the potential to be a major theoretical instrument in modeling and collecting data concerning social epistemology. However, the standard knowledge framework of epistemic game theory glosses over the general conditions of knowledge sharing. My goal in this paper is to clarify the standard approach to interactive epistemology in a manner that makes explicit the fundamental epistemological assumptions behind knowledge sharing.

The standard semantic model (Aumann 1976) derives collective representations of the state of affairs from axioms of epistemic logic, in the form of a theorem asserting that if an event E is *self-evident* (i.e., is known when it is true) for each of a set of agents, then E is necessarily common knowledge among these agents. Clearly, deep epistemological assumptions must be buried in the structure of the standard semantic model that allows us to pass from what each agent knows to what each agent knows concerning the knowledge of the other agents. Certainly no principle of Bayesian rationality permits us to assert that two rational agents share beliefs and that they know that this is the case. I will show exactly what these buried assumptions are and how they can be rendered salient for analytical treatment.

This analysis does not imply that the semantic model must be discarded. In fact, if we have independent epistemic arguments justifying common knowledge of those events for which common knowledge holds in the semantic model, then we are justified in using the semantic model. The argument, however, goes from the justification of common knowledge to the use of the semantic model, not the other way around. By contrast, common knowledge of rationality, I will argue is never an acceptable epistemic assumption.

The framework proposed in this paper for generating propositions concerning common knowledge from basic epistemic assumptions employs three inferential stages. First, there are some events E with the property that if i knows E, then i knows that any other agent j also knows E. For instance, consider a *natural occurrence* N, such as “the ball is yellow,” or “it is raining in Paris” that are immediately perceived by individuals as first-order sensory experience. Under some conditions these natural occurrences are *mutually accessible* to members of a group, meaning 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, the ball’s color may be mutually accessible: i knows that j knows that the ball is yellow. Similarly, there are *rules* R such that in a group of players, if one player i knows that R is a rule, then i knows that any other player j knows the rule as well, presumably by virtue of a social process parallel to that through which i obtained this knowledge.

Second, there are higher-order socially defined events which we may call *games* $\mathcal{G}$, which specify the type of strategic interaction appropriate to the social situation at hand. Games are not mutually accessible, but social conventions may specify that a mutually accessible event F *indicates* $\mathcal{G}$. We call F a *frame*, we write $\mathcal{G} = \mathcal{F}(F)$, and we think of the relation "F indicates $\mathcal{G}$ to agent i" as asserting that when i knows F, he proceeds through a series of mental steps involving the consideration of known social regularities, such as norms and conventions, at the conclusion of which i knows $\mathcal{G}$ (Lewis 1969, Cubitt and Sugden 2003). Knowing $\mathcal{G}$ entails knowing the rules of $\mathcal{G}$, which then become common knowledge by virtue of the reasoning in the previous paragraph. Assuming F is a *public indicator* of $\mathcal{G}$, and that the individuals involved are *symmetric reasoners* (precise definitions are left for later), then $\mathcal{G}$ will be common knowledge.

Third, given epistemic game $\mathcal{G}$, certain social processes that transform private into public information may justify the assumption of a *common prior* for $\mathcal{G}$, which in turn determines a correlated equilibrium of $\mathcal{G}$. The correlating device is a social norm or convention $\mathcal{N} = \mathcal{C}(\mathcal{G})$, which specifies exactly how the game will be played by rational agents. Note that the common prior assumption is extremely demanding, because it requires not only that individuals have the same priors, but that this fact is common knowledge.

Of course, in the real world, at any stage there may be irregularities, lacunae, and clashes that produce non-equilibrium outcomes, and become the object of a more dynamic analysis of strategic interaction when the rules, the payoffs, and the games themselves and their social cues are the object of cultural evolution and strategic intervention.

2 The Standard Semantic Model and Common Knowledge

Let Ω be a set of possible states that a social system can assume. We call Ω the *universe* and each $\omega \in \Omega$ is a possible *state* of the universe. Suppose there are $n > 1$ agents, $i = 1, \dots n$. We assume that in each state ω, agent i knows only that he is in some subset $\mathbf{P}_i\omega \subseteq \Omega$ of states. We say $\omega' \in \mathbf{P}_i\omega$ is *possible* for i at ω. We assume $\omega \in \mathbf{P}_i\omega$ (i.e., i thinks his current state is possible), and $\omega' \in \mathbf{P}_i\omega \Longrightarrow \mathbf{P}_i\omega' = \mathbf{P}_i\omega$ (i.e., if two of i's possibility sets intersect, they are identical). This implies that the sets $\{\mathbf{P}_i\omega | \omega \in \Omega\}$ form a *partition* of Ω (i.e., every state is in exactly one of the $\{\mathbf{P}_i\omega\}$). We call this partition the *knowledge partition* $\mathcal{P}_i$ of the universe Ω for agent i.

An *event* E is an arbitrary subset of Ω. We say the event E *occurs* when the system is in state $\omega \in \Omega$ if $\omega \in E \subseteq \Omega$. We say that agent i *knows* the event E when the state is actually ω if $\mathbf{P}_i\omega \subseteq E$. Note that since $\omega \in \mathbf{P}_i\omega$, when i knows event E in state ω, then E actually is true at ω: knowledge implies truth.

We define $\mathbf{K}_i E \subseteq \Omega$ as the event that i knows event E. It is clear that $\omega \in \mathbf{K}_i E$ if and only if $\mathbf{P}_i\omega \in E$, so $\mathbf{K}_i E$ is the union of all cells $\mathbf{P}_i \in \mathcal{P}_i$ that lie completely in E. If agent i knows E for all $\omega \in E$, we say E is *self-evident* for i. An event E is thus self-evident for i if i knows that E is true in every state where E is in fact true. Note that E is self-evident for i if and only if $\mathbf{K}_i E = E$.

It is easy to see that a self-evident event for i is the union of cells of the knowledge partition $\mathcal{P}_i$. Moreover, the union of any number of events that are self-evident for i is also self-evident for i. Hence, for any event E, $\mathbf{K}_i E$ is the largest self-evident event contained in E.

We say an event E is *common knowledge* for agents $i = 1, \ldots, n$ if, for any state $\omega \in E$, and for any finite sequence of subscripts $l_1, \ldots, l_r$, where each $l_j \in \{1, \ldots n\}$, agent l_1 knows that agent l_2 knows that ... knows that agent l_r knows E in state ω, or more succinctly, $\mathbf{K}_{l_1}\mathbf{K}_{l_2} \ldots \mathbf{K}_{l_r} E = E$.

Suppose an event E is self-evident for all $i = 1, \ldots, n$ agents. We will prove that E is common knowledge. Consider the event $\mathbf{K}_{l_1}\mathbf{K}_{l_2} \ldots \mathbf{K}_{l_r} E$. We must show that this event is simply E. Because E is self-evident to agent l_r, we have $\mathbf{K}_{l_r} E = E$. Therefore, we have

$$\mathbf{K}_{l_1}\mathbf{K}_{l_2} \ldots \mathbf{K}_{l_r} E = \mathbf{K}_{l_1}\mathbf{K}_{l_2} \ldots \mathbf{K}_{l_{r-1}} E.$$

We can proceed similarly for the remaining agents, proving the theorem, which is due to Aumann (1976).

The conclusion that when an event is self-evident to all agents it is common knowledge is quite striking, for it asserts that self-evidence, which appears to describe the epistemic position of an isolated agent, when shared among agents, permits agents to know the content of the minds of other agents.

In fact, the background assumption that gives rise to the assertion that mutual self-evidence implies common knowledge is that each agent "knows" the information partitions of the other agents.[1] For a simple example of this, suppose $\mathbf{P}_i\omega \subset \mathbf{P}_j\omega \subset E$ for two agents i and j. Then in state ω, j knows that E is true in every state that i considers possible, and hence j knows that i knows that E in state ω.

Aumann (1976) defends the semantic model by asserting that no additional assumptions are involved:

[1] The quotation marks are to signal that this is an informal notion of knowledge not captured in the model itself.

> The implicit assumption that the information partitions...are themselves common knowledge...constitutes no loss of generality. Included in the full description of a state of the world is the manner in which information is imparted to the two persons. (p. 1273)

In fact, in the standard presentations of the epistemic model found in the literature, states of the world do not include such full descriptions. Moreover, as we shall see, expanding the model to allow states to include such full descriptions does not lead to a model in which one can assert that individuals know the beliefs of others, unless substantive principles are added to the usual axioms of the modal logic of knowledge.

Aumann (1987) expands on his reasoning as follows:

> Because the specification of each ω includes a complete description of the state of the world, it includes also a list of those other states ω' of the world that are, for Player 2, indistinguishable from ω.... Therefore the very description of the ω's implies the structure of [the partition].... The description of the ω's involves no 'real' knowledge; it is only a kind of code book or dictionary. (p. 9)

However, there is nothing in the definition of common knowledge that specifies what one agent knows about the partition of another agent. Aumann asserts that one can prove in complete generality that using the "tautology" of partition structures, mutual self-evidence implies each agent knows what other agents know. There is surely something amiss here. As we shall see, when the knowledge system "involves no real knowledge," a non-trivial common knowledge condition does not obtain.

For instance, consider a game with two players, Alice and Bob, in which there are two cards, one labeled 'h' and the other labeled 'l'. At the start of the game each player is given one card. The semantic model for this game has two states, which we can label 'hl' and 'lh' corresponding to the two possible distribution of cards. It is easy to see that the event $E =$ "Bob holds the h" is self-evident to both Alice and Bob. Hence, E is common knowledge. This argument does not *prove* that E is common knowledge; rather, our assumptions concerning the mental capacities and social relations between Alice and Bob lead us to believe that E is common knowledge. If, for instance, Bob lacked a "theory of mind" (Baron-Cohen 1991), he might not be able to deduce from the fact that he holds the h that Alice knows, by virtue of the fact that she holds the l, that Bob holds the h. Indeed, if the species involved were any other than humans, we would not expect the semantic model to reflect the commonality of knowledge between the two players (Premack and Woodruff 1978). Note that there is nothing in Bayesian rationality that suggests that rational individuals possess such a "theory of mind."

If the problem were simply the failure of the semantic model to apply to nonhumans, a deeper analysis might not be worth the trouble for those of us who confine our studies to the behavior of our own species. However, the problem applies also to humans in real-world contexts where there is no experimenter laying down the rules of the game, but rather agents must infer the structure of the game and its epistemic properties from possibly ambiguous social cues. In such cases, an explicit social epistemology will be needed to conclude that the semantic model is an appropriate representation of the game.

3 A Syntactic Model of Distributed and Shared Knowledge

Aumann (1999) elaborates on his defense of the notion that the partition structure is purely a formalism, developing a syntactic model of common knowledge in which partitions of the universe are not employed, and shows that it has a canonical relationship with a semantic model that employs the partition machinery. This model lays bare the epistemological presumptions of the standard semantic model of common knowledge and reveals the presuppositions that permit common knowledge to be inferred. In fact, we will see that Aumann's construction shows the *substantive* rather than the *tautological* nature of the partition construction in the standard semantic model.

My exposition follows Aumann (1999), except that I omit most proofs. Suppose we have $n > 1$ individuals, and a set of letters from an alphabet $\mathcal{X} = \{x, y, z, \ldots\}$ which we think of as events, logical symbols $\vee$, $\neg$, $k_1, \ldots, k_n$ and left, '(' and right, ')' parentheses. *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 $f \Longrightarrow g$, $\neg(\neg f \vee \neg g)$ as $f \wedge g$, $(f \Longrightarrow g) \wedge (g \Longrightarrow f)$ as $f \Leftrightarrow g$, and we drop parentheses where no ambiguity results, assuming the usual precedence ordering of the propositional calculus, and assigning the highest order 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:

$$k_i f \Longrightarrow f \tag{1}$$

$$k_i f \Longrightarrow k_i k_i f \tag{2}$$

$$k_i(f \Longrightarrow g) \Longrightarrow (k_i f \Longrightarrow k_i g) \tag{3}$$

$$\neg k_i f \implies k_i \neg k_i f. \tag{4}$$

Equation (1) mirrors the semantic property $\mathbf{K}_i E \subseteq E$ for any event E. Equation (2) mirrors the fact that $\mathbf{K}_i E$ is self-evident for any event E in the semantic model. Equation (3) asserts that *modus ponens* applies to the knowledge operator. Finally, (4), called the *axiom of transparency*, is required to ensure that the semantic realization of the syntactic system has a partition structure. We call a system consisting of the alphabet $\mathcal{X}$, the formulas and the tautologies a *syntactic system* $\mathcal{S}$. We generate the set of tautologies $\mathcal{T}$ by assuming *modus ponens* (i.e., $f, (f \implies g) \in \mathcal{T}$ implies $g \in \mathcal{T}$) and agents know all the tautologies (i.e., $f \in \mathcal{T}$ implies $k_i f \in \mathcal{T}$).

A *state* ω of the syntactic system $\mathcal{S}$ is list that is closed under *modus ponens*, and for every formula f, exactly one of f and $\neg f$ is in ω. In other words, a state is a maximally consistent set of formulas, and so represents a possible state of the system and the epistemic state of the agents concerning the system. For instance, suppose there is one letter, x, and let ω be a state of $\mathcal{S}$. Then, either $x \in \omega$ or $\neg x \in \omega$. Moreover, $x \vee \neg x \in \omega$. To see this, suppose $x \vee \neg x \notin \omega$. Then $\neg(x \vee \neg x) \in \omega$. This can be rewritten as $\neg x \wedge x \in \omega$. But the formula $f \wedge g \implies f$ is a tautology, so if $f \wedge g \in \omega$, then $f \in \omega$ by *modus ponens*. Similarly, $g \in \omega$. Applying this to $\neg x \wedge x \in \omega$, we see that $x \in \omega$ and $\neg x \in \omega$, which is impossible.

It is easy to see that if ω is a state, then every tautology is in ω (i.e., $\mathcal{T} \subset \omega$). 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 non-equivalent formula to ω without violating the list property.

Moving to the knowledge operators, suppose $x \in \omega$. Then, for each agent i, either $k_i x \in \omega$ or $\neg k_i x \in \omega$. Moreover, we must have $\neg k_i \neg x \in \omega$ for each i. This is because if $x \in \omega$, then $k_i \neg x \in \omega$ would imply, by (1), that $\neg x \in \omega$, which is a contradiction. We often abbreviate $\neg k_i \neg f$ as $p_i f$, and we say "i considers f possible." Similarly, for any formula $f \in \omega$, for each agent i, either $k_i f \in \omega$ or $\neg k_i f \in \omega$, and $p_i f \in \omega$. Thus, every state ω of $\mathcal{S}$ includes arbitrarily long finite strings of the form 'i knows that j considers it possible that k knows that..." terminating in either x or $\neg x$.

States of $\mathcal{S}$ are thus very large sets, containing much redundant material. For instance, if $f \in \omega$, then $f \vee g \in \omega$ for any formula g whatever. Moreover, if f is a tautology, then any string of the form $k_i k_j \dots k_r f \in \omega$. Nevertheless, a state ω does capture all of the epistemic relations among the agents and the underlying letters $x, y, z \dots$.

For any formulas f and g and any state ω, we have:

$$\neg f \in \omega \text{ iff } f \notin \omega \tag{5}$$

$$f \vee g \in \omega \text{ iff } f \in \omega \text{ or } g \in \omega \tag{6}$$

$$f \wedge g \in \omega \text{ iff } f \in \omega \text{ and } g \in \omega \tag{7}$$

$$f \Longrightarrow g \in \omega \text{ iff } f \in \omega \text{ implies } g \in \omega \tag{8}$$

$$f \Leftrightarrow g \in \omega \text{ iff } f \in \omega \text{ iff } g \in \omega. \tag{9}$$

4 The Semantic Interpretation of Syntactic System $\mathcal{S}$

Let Ω^* be the set of all states derived from the syntactic system $\mathcal{S}$ using the above construction. We define a semantic knowledge system on universe Ω^* as follows. A set $E \subseteq \Omega^*$ is called an *event*. Let $\kappa_i(\omega)$, $\omega \in \Omega^*$ be the set of all formulas in ω of the form $k_i f$ for some formula f. The cells of the partition $\mathcal{P}_i^*$ of Ω^* for player i are defined by $\mathbf{P}_i^*\omega = \{\omega' \in \Omega^* | \kappa_i(\omega) = \kappa_i(\omega')\}$. Thus, an agent can distinguish between two states in Ω^* if and only if he has some knowledge in one state that he does not have in the other. The knowledge operator $\mathbf{K}_i^*$ is defined in the usual manner by $\mathbf{K}_i^* E = \{\omega | \mathbf{P}_i^*\omega \subseteq E\}$. We call the resulting system the *canonical semantic knowledge system* Ω^* corresponding to $\mathcal{S}$.

To construct a map from $\mathcal{S}$ to Ω^*, we define

$$E_f = \{\omega \in \Omega^* | f \in \omega\}; \tag{10}$$

i.e., the event E_f is the set of states in which f is true. The most important implications of this mapping are

$$\mathbf{K}_i E_f = E_{k_i f} \tag{11}$$

$$E_f \subseteq E_g \text{ iff } (f \Longrightarrow g) \in \mathcal{T}; \tag{12}$$

$$\mathbf{K}_i E = \left\{\omega | \cap_{k_i f \in \omega} E_f \subseteq E\right\}. \tag{13}$$

These properties imply that the syntactic operator k_i does in fact correspond to the knowledge operator $\mathbf{K}_i$, and logical implication in the syntactic system is the counterpart of set inclusion in the semantic system. Moreover, what an agent knows, given an event E, is whatever follows logically from the formulas that the agent knows, given E.

In terms of Ω^*, the theorem that an event is common knowledge if and only if it is self-evident for all agents is of course true. However, it is easy to see that no event in Ω^* that is in the image of $\mathcal{S}$ can be self-evident to all agents. For if E is in the image of $\mathcal{S}$ under the above mapping, the $E = E_f$ for some formula f. If f is a tautology, the $E = \Omega^*$, which is trivially common knowledge. Otherwise, if f

has no knowledge operators, then $\neg k_i f \in E$ holds for all i, so E is not common knowledge. If f begins with k_i, then $\neg k_j f \in E$ for $j \neq i$, so E is not common knowledge. Finally, if f begins with $\neg k_i$, the E is not common knowledge by definition. We conclude that no event in the image of $\mathcal{S}$ is common knowledge except the trivial event Ω^* itself.

However, there are common knowledge events in Ω^*, which we can characterize by forming a set X consisting of subset of the letters $x, y, z, \ldots$, some of which may be prefixed by $\neg$, and including a state ω in an event E if and only if ω includes all formulas in X, as well as all formulas of the form $k^*_{l_1} k^*_{l_2} \ldots k^*_{l_r} f$, where $f \in X$ and k^* is either k or p. To ensure that ω is a state, we must also include $\neg k_i y$ for letters y not included in X, and filling out ω with consistent formulas terminating in $\neg k_i y$. Each such event may be called a *common knowledge subuniverse* of Ω^*.

5 The Universal Property of Semantic Model Ω^*

The map $f \rightarrow E_f$ from the syntactic system $\mathcal{S}$ to the semantic system Ω^* is injective in the sense that $E_f = E_g$ for formulas f and g if and only if $f \Leftrightarrow g$ is a tautology of the syntactic system. Moreover, (11), (12), and (13) show that the mapping preserves the epistemic properties of the syntactic system. While all nontrivial common knowledge events fail to be in the image of this map, the semantic system Ω^* is *universal* in the sense that all semantic models derivable from $\mathcal{S}$ are common knowledge subuniverses of Ω^*.

To see this, let Ω^+ be an arbitrary semantic system with n agents and a set of possibility operators $\mathbf{P}_i^+$. For every letter $x, y, z, \ldots$ in the syntactic system $\mathcal{S}$, let $\phi(x)$, $\phi(y)$, $\phi(z)$ be arbitrary subsets of Ω^+. We interpret $\phi(x)$ as the set of states of Ω^+ at which x is true. We can extend the mapping ϕ to all of $\mathcal{S}$ using the following rules:

$$\phi(\neg f) = \{\omega \in \Omega^+ | \omega \notin \phi(f)\}; \tag{14}$$

$$\phi(f \vee g) = \phi(f) \cup \phi(g); \tag{15}$$

$$\phi(k_i f) = \mathbf{K}_i^+ \phi(f), \tag{16}$$

where $\mathbf{K}_i^+$ is the knowledge operator in Ω^+ derived from the $\mathbf{P}_i^+$.

In this manner, we can map the syntactic system $\mathcal{S}$ into an arbitrary semantic system Ω^+ representing some specification of the meaning of the letters $x, y, z, \ldots$ of the syntactic system. The key property of this association is that if we define $\psi(\omega^+)$ for any state $\omega^+ \in \Omega^+$ by

$$\psi(\omega^+) = \{f \in \mathcal{S} | \omega^+ \in \phi(f)\}; \tag{17}$$

that is, $\psi(\omega^+)$ is the set of formulas of the syntactic system that are interpreted as true in state $\omega^+ \in \Omega^+$. It is straightforward to show that $\omega = \psi(\omega^+)$ is closed under *modus ponens* and for every formula f, ω contains exactly one of f and $\neg f$. Hence ω is a state of the semantic system Ω^*. It follows that ψ is a mapping from Ω^+ to Ω^*.

Let Ω^ϕ be the image of Ω^+ under the mapping ψ, and let Γ^ϕ be the smallest common knowledge subuniverse of Ω containing Ω^ϕ. It is straightforward to show that the following diagram commutes:

$$\begin{tikzcd}
& \mathcal{S} \arrow[dl, "\phi^*"'] \arrow[dr, "\phi"] & \\
\Gamma^\phi & & \Omega^+ \arrow[ll, "\psi"']
\end{tikzcd}$$

where ϕ^* is the mapping $f \to E_f \cap \Omega^\phi$. This shows that every semantic model is isomorphic to a common knowledge subuniverse of the canonical semantic model associated with an appropriate syntactic model (Aumann 1999).

6 A Semantic Model with no Nontrivial Common Knowledge

It is possible to construct a finite semantic epistemic model Ω in which states directly incorporate the distribution of knowledge. This model shows that without additional substantive assumptions, we cannot conclude that any agent knows that another agent knows something.

Suppose the syntactic system $\mathcal{S}$ has agents $i = 1, \ldots, n$ and a single letter, so $\mathcal{X} = \{x\}$. We define a semantic system Ω^+ in which each state is of the form $\omega = \{\omega_0, \omega_1, \ldots, \omega_n\}$, where $\omega_0 \in \{x, \neg x\}$, and $\omega_i \in \{t, f\}$ for $i = 1, \ldots, n$. We interpret ω_0 as the "underlying" state of the system, and $\omega_i = t$ if i knows this underlying state, and $\omega_i = f$ otherwise.

For agent i, we define $\mathbf{P}_i\omega = \{\omega' \in \Omega^+ | \omega'_0 = \omega_0 \wedge \omega_i = \omega'_i\}$ if $\omega_i = t$, and $\mathbf{P}_i\omega = \Omega^+$ if $\omega_i = f$. It is easy to check that if event $E = \mathbf{P}_i\omega$ where $\omega_i = t$, then $K_i E = E \subset \Omega^+$ and for $j \neq i$, $K_j K_i E = \emptyset$. In other words, in this semantic system, no agent knows whether other agents know the underlying state of the system. It follows that the only event that is common knowledge is the trivial event Ω^+.

We can define a natural mapping from $\mathcal{S}$ to Ω^+ by defining $\phi(x) = \{\omega \in \Omega^+ | \omega_0 = x\}$, and extending ϕ to $\mathcal{S}$ using (14-16). It is easy to show that for

$z \in \{x, \neg x\}$ and $j \neq i$, $\phi(k_j k_i z) = \emptyset$ and $\phi(\neg k_j k_i z) = \Omega^+$. The common knowledge subuniverse Γ^ϕ corresponding to Ω^+ is then the common knowledge subuniverse generated by E_f, where the syntactic formula f is give by

$$f = \bigwedge_{\substack{i,j=1 \\ i \neq j}}^{n} \neg k_j k_i x \wedge \neg k_j k_i \neg x.$$

Clearly, in this subuniverse, no agent knows which other agents know the underlying state of the system.

7 The Three Tactful Ladies: a Syntactic Analysis

While walking in the garden, Alice, Bonnie and Carole encounter a violent thunderstorm and duck hastily into a nearby salon for tea. Carole notices that Alice and Bonnie have dirty foreheads, although each is unaware of her own condition. Carole is too tactful to mention this embarrassing situation, which would surely lead them to blush, but she observes that, like herself, each of the two ladies knows that someone has a dirty forehead but is also too tactful to mention this fact.

At this point, a little boy walks by the three ladies' table and exclaims "I see a dirty forehead!" Alice comments, "I can't say that I have a dirty forehead." "Nor can I," says Bonnie. Carole realizes that she has a dirty forehead, and blushes.

The problem is interesting because the three ladies already knew what the little boy told them, and Alice and Bonnie would have said the same thing before hearing the little boy's announcement. It thus appears mysterious that Carole can deduce anything after the little boy's exclamation that she could not have decided before.

The standard semantic model for this problem is to suppose Ω consists of eight states of the form $\omega = xyz$, where $x, y, z \in \{d, c\}$ describe 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 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, so she blushes.

This is a very famous problem that puts the semantic model through its paces. But, there are many unstated epistemological assertions involved in the conclusion that Carole knows the state of her forehead. We can see exactly what they are by working within the syntactic model for the problem.

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 without qualification, we mean any $i = A, B, C$, and when we write i, j, we mean any $i, j = A, B, C$ with $j \neq i$. Finally, let y_i be the condition that i blushes. The six symbols x_i and y_i represent the letters of a syntactic structure $\mathcal{S}$, with universe Ω. Let E be the event prior to the little boy's exclamation. The statement of the problem tells us that for all $\omega \in E$, $x_i \in \omega$, and $k_i x_j \in \omega$, for $i \neq j$; i.e., each lady sees the forehead of the other two ladies, but not her own. It is easy to check that these conditions are compatible with $\neg k_i x_i \in \omega \in E$; i.e., no lady knows the state of her own forehead at event E.

The problem also asserts that when $k_i x_i$ occurs, y_i also occurs. Moreover, the problem presumes that y_i is *mutually visible* in the sense that when y_i occurs, so does $k_j y_i$. In fact, we will need the stronger statement that when y_i occurs, so does $k_l k_j y_i$ for any agents i, j, l. This is because, in the above analysis, Carole infers that Bonnie knows that Alice does not blush from the fact that Alice does not blush. This may appear to be a weak assumption, but in fact it is the *first time* we have made a substantive assertion of the form $k_i k_j f$. We say that a natural occurrence z is *public* to order r if the occurrence of z entails the occurrence of $k_i k_j \ldots k_l z$ for all atomic formulas starting with r k-operators; i.e., $z \in \omega$ implies $k_i k_j \ldots k_l z \in \omega$. In our problem, we must assume y_i is public of order two. *Publicity is the second mechanism we have encountered for asserting that agents share beliefs.*

The public property is actually the conjunction of two properties worthy of separate definition. We say natural occurrence z is *external* at event E if z implies $k_i z$; i.e., $z \in \omega \in E$ implies $\kappa_i z \in \omega$. We say that z is *mutually accessible* at E if $k_i z \in \omega \in E$ implies $k_i k_j z \in \omega$. Both terms are defined with respect to all agents, or to a particular subset of agents. An event is public order two if the event is external and mutually accessible.

The problem also asserts that $k_i x_i \in \omega \in E$ implies $y_i \in \omega$. The problem does not assert the equally important fact that $k_j(k_i x_i \implies y_i) \in \omega$, but this fact is surely needed to solve the problem, especially in the contrapositive form $k_j(\neg y_i \implies \neg k_i x_i) \in \omega$; i.e., if a lady does not blush, the other ladies know that she does not know she has a dirty forehead. This characteristic of y_i as a public natural occurrence gives us a third manner of inferring second-order epistemic statements: y_i is external, but entails knowledge of the "internal" event $\neg k_i x_i$. An external natural occurrence z that entails $k_j k_i w$ for some w may be called an *external indicator* of $k_i w$. If z is also mutually accessible, we may call z a *public indicator* of w. In particular, $\neg y_i$ is a public indicator of $\neg k_i x_i$ in our problem. We actually never use the assumption that y_i implies $k_i x_i$; many different mental states can be associated with bluffing without altering the logic of the argument. However, Carole's reasoning is logically sound only if she knows that Bonnie knows that Alice's failure to blush means Carole knows that Bonnie knows that Alice does not know the state of her forehead. That is, we require that y_i be mutually accessible to order three.

Note that the fact than a natural occurrence z is public says something about how agents know that they share the external world around them and turn perceptions in a parallel manner into knowledges. The concept of publicity is the closely associated with the fact the agents know that they share certain species-level sensory and information processing systems, and certain types of mental events are reliably connected to sensory experiences across individuals. The fact that z is an indicator of an internal state may also be rooted in elementary physiology and sense perception. For instance, there are reliable, universal external events indicating fear, pain, joy, sleepiness, anger, and other primary emotions (i.e., emotions we share with many other vertebrate species). However, the property of being an indicator may also be socially specific. Blushing is a human universal (Brown 1991), but the mental states that lead to blushing are highly socially specific. Ladies blush when they have dirty forehead in some societies, but not in others. Moreover, little boys generally do not blush when their foreheads are dirty.

Where does the little boy's exclamation enters the analysis? Let E' be the state of knowledge following the exclamation $p = x_A \vee x_B \vee x_C$? We must assume p is public to some appropriate order, or the problem cannot be solved. Moreover, $k_i p$ is true in E because x_i is external to $j \neq i$. Assuming x_i is mutually accessible to $j, l \neq i$, $k_j k_i p$ is also true in E. So, if p gives new information, it must be public of order three. We now have the following argument.

The reasoning following the little boy's statement can be summarized as follows. Step 1: Carole assumes $\neg x_C$ and infers $k_A \neg x_C$ and $k_B \neg x_C$; Step 2: Bonnie assume $\neg x_B$ and concludes, using the fact that y_i is a public event for $j \neq i$, that $k_A \neg x_B$. Step 3: Suppose $\neg x_B$ and $\neg x_C$ are mutually accessible. then by assumption $k_B \neg x_B$, so $k_B k_A \neg x_B$, and also by assumption $k_B \neg x_C$, so $k_B \kappa_A \neg x_C$.

Because $k_B k_A(x_A \vee x_B \vee x_C)$, we have

$$k_B(k_A(x_A \vee x_B \vee x_C) \wedge k_A \neg x_B \wedge \neg x_C) \Longrightarrow k_B k_A x_A \Longrightarrow k_B y_A.$$

However, y_A is false, so $k_B y_A$ is false. Thus Bonnie's assumption that $\neg x_B$ is wrong, so she logically concludes x_B, which means $k_B x_B$, and hence y_B. Step 4: Because p is third-order public and assuming x_i are also third order mutually accessible, Carole knows all of the above reasoning, and hence she knows that $\neg x_C$ implies y_B. Because y_B is false, she concludes that x_C, so $k_C x_C$, which implies y_C.

8 Mutually Accessible Events and Symmetric Reasoning

Given the syntactic system $\mathcal{S}$ with agents $i = 1, \ldots, n$, we define a set N of natural occurrences (letters of the alphabet $\mathcal{X}$) to be *mutually accessible* if, for any i, j and any $x \in N$, $k_i x \Longrightarrow k_i k_j x$, We say $x \in N$ is a *public signal* at event E if $x \in \omega \in E$ implies $k_i x \in \omega$ for $i = 1, \ldots, n$. We say event E is *common knowledge* in $\mathcal{S}$ at state ω if ω includes all atomic formulas of the form $k_{i_1} k_{i_2} \ldots k_{i_r} E$ for all orders $r > 0$. Finally, we say i and j are *symmetric reasoners* with respect to a mutually accessible event A if for any event E, $k_i A \Longrightarrow k_i E$ implies $k_i A \Longrightarrow k_i k_j E$.

Before deploying these concepts in proving Theorem 1, it is well to pause to see what they mean. A mutually accessible event must be a natural occurrence providing first-order sensory data to the agents involved. Thus, mental events are not mutually accessible, nor are inferences of mental events by virtue of overt behavior (e.g., "John believes he knows the answer" is not the type of event that could be mutually accessible, although "John looks ill" may be, and "John looks green" is even more likely to be mutually accessible). Moreover, the mutual accessibility of an event depends on the agents involved being aware that they all are receiving the same sensory input, that all possess normal mental capacities, all are attentive to the event, etc. Mutually accessible natural occurrences are the minimum irreducible transmission mechanisms of mental constructs across minds.[2] In humans, only sight and sound give rise to reliably mutually accessible events. It is unclear whether mutually accessible events occur in other species, although there is (disputed) evidence that chimpanzees share the capacity to recognize mutually accessible events (Premack and Woodruff 1978, Heyes 1998, Tomasello 1999, Tomasello et al. 2005).

[2] The relationship between natural occurrences and the 'sense data' of the logical positivists (Wittgenstein 1999[1921]) is close. However, I do not grant natural occurrences the epistemological primacy accorded them in logical positivist thought. Such occurrences are simply the first links in a chain of constructs that jointly explain common knowledge.

The concept of symmetric reasoning allows agent i to infer from the fact that he shares mutually accessible events with agent j, that both i and j will engage in a parallel sequence of mental activities in transforming the data that they share into further knowledges, so they will share certain mental constructs derived from the parallel processing of the same information. For instance, if an event E is mutually accessible, so i knows that j knows E and conversely, i might reason that j knows that i knows E, and the therefore i might know that j knows that they share the knowledge that i knows that j knows E. Of course, the symmetric reasoning assumption is very strong and will often fail to apply, because it may be the case that several people witness the same natural occurrences and draw extremely heterogeneous conclusions therefrom (the 'Rashamon Effect').[3]

THEOREM 1. *Consider the syntactic system $\mathcal{S}$ on letters $\mathcal{X}$ with agents $i = 1, \ldots, n$. Let $N \subseteq \mathcal{X}$ be a set of natural occurrences that are public signals and mutually accessible at ω. If the agents are symmetric reasonsers with respect to each $x \in N$, then N is common knowledge at ω.*

Proof: Choose an $x \in N$. Then, if $x \in \omega$, ω includes the first-order atomic formulas $k_i x$ because x is mutually known. Because x is mutually accessible, ω contains the second order atomic formulas $\{k_i k_j x\}$ for all i, j. The proof follows by induction. Suppose ω contains all r-order atomic formulas $k_{i_1} k_{i_2} \ldots k_{i_r} x$. For all i, $k_i x \in \omega$ implies $k_i k_{i_2} \ldots k_{i_r} x \in \omega$, so because agents are symmetric reasoners with respect to x, for any j, $k_i x \in \omega$ implies $k_i k_j k_{i_2} \ldots k_{i_r} x \in \omega$. Hence, all $(r+1)$-order atomic formulas of the form $k_i k_j \ldots k_l x \in \omega$. ■

9 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.

[3] The term "symmetric reasoning" is defined in Vanderschraaf and Sillari (2007), who attribute the term to personal communication with Chris Miller and Jarah Evslin. The concept is attributed to Lewis (1969).

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 understand that the rules are *mutually accessible*: 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.

Of course, by assuming that the rules of the game are mutually accessible, an argument similar to our analysis of natural occurences justifies common knowledge of the rules of the game.

10 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 Aumann and Brandenburger's (1995) Theorem B. 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, 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." 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 waving of a hand 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.[4]

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

THEOREM 2. *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 $\omega \in F$.*

11 Common Knowledge of Rationality

A rational agent has an objective function and a subjective prior over events in the world, and act to maximizes the expected payoff associated with various strategies. Game theorists widely accept the notion that common knowledge of rationality (CKR) is a plausible epistemic assumption in a model where in fact all agents are rational. Indeed, the elimination of dominated strategies (Pearce 1984, Bernheim 1984) and the use of backward induction (Aumann 1995) without even mention of the common knowledge of rationality assumption is virtually universal. In fact, however, the common knowledge of rationality assumption is extremely strong (Aumann 1996). Indeed, it is not a defensible epistemic assumption (Gintis 2009).

It is difficult to assess the conditions under which CKR is plausible, because CKR is not derived from more elementary, and themselves plausible, epistemic conditions. Rationality is surely neither a natural occurrence or a rule of the game. Perhaps some symmetric reasoning assumption could adduced to the proof of CKR, but it is easy to show that such an assumption would necessarily depend on the payoffs of the game being played, which shows that CKR cannot be treated as a purely epistemic conditions. In particular, we could not claim that CKR is an "idealized" extension of agent rationality.

To see this, we may use a game in which it is implausible that rational players will eliminate dominated strategies. Consider the following game G_n, known as the *Traveler's Dilemma* (Basu 1994). Two business executives pay bridge tolls while on a trip but do not have receipts. Their superior tells each of them to report

[4] 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.

independently an integral number of dollars between 2 and n on their expense sheets. If they report the same number, each will receive this much back. If they report different numbers, each will get the smaller amount, plus the low reporter will get an additional $2 (for being honest) and the high reporter will lose $2 (for trying to cheat).

	s_2	s_3	s_4	s_5
s_2	$2, 2$	$4, 0$	$4, 0$	$4, 0$
s_3	$0, 4$	$3, 3$	$5, 1$	$5, 1$
s_4	$0, 4$	$1, 5$	$4, 4$	$6, 2$
s_5	$0, 4$	$1, 5$	2.6	$5, 5$

Figure 1: The Traveler's Dilemma

Let s_k be strategy "report k." Figure 1 illustrates the game G_5. Note first that a mixed strategy $\epsilon s_2 + (1 - \epsilon)s_4$ strongly dominates s_5 whenever $1/2 > \epsilon > 0$. When we eliminate s_5 for both players, a mixed strategy $\epsilon s_2 + (1 - \epsilon)s_3$ strongly dominates s_4 for any $\epsilon > 0$. When we eliminate s_4 for both players, s_2 strongly dominates s_3 for both players. Hence (s_2, s_2) is the only strategy pair that survives the iterated elimination of strongly dominated strategies. It follows that s_2 is the only strategy that can be played, assuming CKR.

A similar analysis holds for all games G_n. Suppose $n = 100$. It is not plausible to think that individuals would actually play s_2, s_2 because by playing a number greater than, say, 92, either is assured of at least 90. Experiments confirm that players almost never play the CKR strategy (Basu 1994, Capra et al. 1999).

If CKR is not plausible for this game, it is not clear why it should ever be plausible as an initial epistemic assumption.

12 Common Knowledge of Rationality and Epistemic Blindspots

I have argued that there is no plausible basis for the assumption of CKR. Yet, it is generally considered in economic theory that if something is true and there is no reason to believe that there is asymmetric information concerning this fact, then it is harmless to assume that the truth is common knowledge. There is, in fact, good reason to reject this argument. Given a few basic rules of the modal logic of knowledge, it can easily be shown that a proposition can be true yet cannot be known to be true (Sorensen 1988). Consider, for instance, the Surprise Examination Paradox.

A group of game theorists once took an intensive Monday-through-Friday logic course. After several weeks, the professor announced that there would be a surprise examination one day the following week. Each student thought to himself, “The exam cannot be given next Friday because then it would not be a surprise.” Each then concluded that, for similar reasons, the exam could not be given next Thursday, next Wednesday, next Tuesday, or next Monday. Each student thus concluded that the professor was mistaken. The professor gave the exam the next Tuesday, and all of the students were surprised.

Kaplan and Montague (1960) were the first to notice that there is nothing paradoxical about the professor’s assertion, but adding the assumption that a student *knows* the assertion does produce an inconsistent logical system. I will follow Binkley (1968), who weakened the epistemic conditions under which the contradiction appears. Let us assume there are only two days, Monday and Tuesday. The five-day argument is similar, but longer. We take the case of a single student with knowledge operator k. We assume for any knowledge operator that

A1 $kf \implies \neg k \neg f$
A2 $kf \wedge k(f \implies g) \implies kg$
A3 $kf \implies kkf$

Note that A1 is weaker than the usual assumption $kf \implies f$; i.e., “what is known is possible” is weaker than “what is know is true.” We also assume the student knows all tautologies of the propositional calculus and all axioms.

Let $k_m f$ mean “the student knows f on Monday” and let k_t mean “the student knows f on Tuesday.” Let E_m be the event “the exam is given on Monday,” and let E_t be the event “the exam is given on Tuesday.” We assume

A4 $\neg E_m \implies k_t \neg E_m$
A5 $k_m f \implies k_m k_t f$

A4 says that if the exam is not given on Monday, then on Tuesday the student knows this fact, and A5 says that if the student knows something on Monday, he knows on Monday that he will continue to know it on Tuesday. The professor’s assertion can be written as

$$E = (\neg E_m \iff E_t) \wedge (E_m \implies \neg k_m E_m) \wedge (E_t \implies \neg k_t E_t). \qquad (18)$$

Let us assume $k_m E$. From A4 we have

$$k_m(\neg E_m \implies k_t \neg E_m). \qquad (19)$$

From $k_m E$ we have $k_m(E_t \implies \neg E_m)$, which with (19) gives

$$k_m(E_t \implies k_t \neg E_m). \qquad (20)$$

Now from $k_m E$ and A5, we have $k_m k_t(\neg E_m \implies E_t)$, so

$$k_m(k_t \neg E_m \implies k_t E_t). \qquad (21)$$

From (20) and (21), we have

$$k_m(E_t \Longrightarrow k_t E_t). \tag{22}$$

Now $k_m E$ implies $k_m(E_t \Longrightarrow k_t \neg E_t)$, which together with (22) implies $k_m(\neg E_t)$, and hence $k_m E_m$. This, together with $k_m E$ gives

$$k_m \neg k_m E_m. \tag{23}$$

However, $k_m E_m$ and A3 imply $k_m k_m E_m$, so by A1, we have $\neg k_m \neg k_m E_m$, which contradicts (23). Therefore the original assumption $k_m E$ is false. E is thus true but it is inadmissible to assume therefore that the student knows that E is true.

This example of epistemic blindspot should be a cautionary tale for the assumption of CKR. In the surprise examination case, the student may have good reason to "know" that the exam will actually occur and will be a surprise (e.g., because this has happened in previous years, it is an ironclad school policy, the professor never lies, etc.), but cannot "know" in the formal sense of the admissibility of the assumption $k_m E$. Similarly, in a game-theoretic setting, there may be CKR is some informal sense, but the assumption may contradict other assumptions of the model. The following is an example of this.

13 How to Play the Repeated Prisoner's Dilemma

	C	D
C	R,R	S,T
D	T,S	P,P

In cases where a stage game is repeated a finite but considerable number of times, it is reasonable to assume Bayesian rationality, avoid backward induction, and use decision theory to determine player behavior. Consider, for instance, the Prisoner's Dilemma, the stage game of which is shown to the right with $T > R > P > S$, repeated until one player defects or 100 rounds have been played. Backward induction implies that both players will defect in the very first round, and indeed, this is the only Nash equilibrium of the game, and is implied by CKR.

Suppose Player 1 conjectures that Player 2 will cooperate up to round k and then defect, with probability g_k. Then, Player 1 will choose a round m to defect in that maximizes the expression

$$\begin{aligned}\pi_m = \sum_{i=1}^{m-1}((i-1)R + S)g_i + ((m-1)R + P)g_m \\ + ((m-1)R + T)(1 - G_m),\end{aligned} \tag{24}$$

where $G_m = g_1 + \cdots + g_m$. For plausible conjectures, maximizing this expression suggests cooperating for many rounds. For instance, suppose g_k is uniformly distributed in the rounds $m = 1, \ldots, 99$. Suppose, for concreteness, $(T, R, P, S) = (4, 2, 1, 0)$. Then, it is easy to show that (24) implies it is a best response to cooperate up to round 98. Indeed, if Player 1 conjectures that player 2 will defect in round 1 with probability 0.95 and otherwise defect with equal probability on any round from 2 to 99, then it is still a best response to defect in round 98. Introducing CKR leads to an inconsistent set of assumptions for this model because CKR now implies the players are not rational, given their conjectures.[5]

14 Conclusion

Throughout much of the previous century, economic theory placed little value on the commonality of beliefs. The Walrasian general equilibrium model required no restriction on the heterogeneity of preferences, and admitted no role for individual beliefs. The increased importance of game theory in the last quarter of the century presented serious conceptual issues concerning the relationship between rationality and equilibrium criteria. Interactive epistemology in general, and epistemic game theory in particular, have shown that the commonality of beliefs, in the form of common priors, common knowledge, and correlating devices, is central in explaining how rational actors successfully coordinate their activities.

Yet, economists have avoided grappling with the question of where common priors, common knowledge, and correlating devices come from. Harsanyi's doctrine, asserting that all differences in subjective priors of rational individuals is due to asymmetric information has a degree of plausibility in dealing with what we have called, following Aumann (1999), "natural occurrences," but not otherwise. Semantic knowledge models (Aumann 1987) assert that mutual self-evidence logically implies common knowledge, but as we have seen, they achieve this by making implausible tacit assumptions concerning the structure of knowledge. Using Aumann's (1987) syntactic knowledge model, we have shown that explicit epistemological assumptions concerning the sharing of knowledge among individuals are required to prove common knowledge.

[5] It might be thought that the above conjectures are themselves incompatible with CKR. This is not the case, although the conjectures are incompatible with common knowledge of common conjectures, as the reader can easily check.

REFERENCES

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.

— , “Backward Induction and Common Knowledge of Rationality,” *Games and Economic Behavior* 8 (1995):6–19.

— , “Reply to Binmore,” *Games and Economic Behavior* 17 (1996):138–146.

— , “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.

Baron-Cohen, Simon, “Precursors to a Theory of Mind: Understanding Attention in Others,” in Andrew Whiten (ed.) *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading* (Oxford: Basil Blackwell, 1991) pp. 233–251.

— , *Mindblindness* (Cambridge, MA: MIT Press, 1995).

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).

Bernheim, B. Douglas, “Rationalizable Strategic Behavior,” *Econometrica* 52,4 (July 1984):1007–1028.

Binkley, Robert, “The Surprise Examination in Modal Logic,” *Journal of Philosophy* 65 (1968):127–135.

Brandenburger, Adam and Eddie Dekel, “Rationalizability and Correlated Equilibrium,” *Econometrica* 55 (1987):1391–1402.

Brown, Donald E., *Human Universals* (New York: McGraw-Hill, 1991).

Capra, C. Monica, Jacob K. Goeree, Rosairo Gomez, and Charles A. Holt, “Anomalous Behavior in a Traveler’s Dilemma?,” *American Economic Review* 89,3 (June 1999):678–690.

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.

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

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

Heyes, C. M., "Theory of Mind in Nonhuman Primates," *Behavioral and Brain Sciences* 21,1 (1998):101–034.

Hintikka, Jaakko, *Knowledge and Belief* (Ithica: Cornell University Press, 1962).

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

Kaplan, D. and R. Montague, "A Paradox Regained," *Notre Dame Journal of Formal Logic* 1,3 (1960):79–90.

Kripke, Saul, "Semantical Considerations on Modal Logic," *Acta Philosophica Fennica* 16 (1963):83–94.

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

Pearce, David, "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52 (1984):1029–1050.

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

Savage, Leonard J., *The Foundations of Statistics* (New York: John Wiley & Sons, 1954).

Sorensen, Roy A., *Blindspots* (Oxford: Oxford University Press, 1988).

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

— , Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll, "Understanding and Sharing Intentions: The Origins of Cultural Cognition," *Behavioral and Brain Sciences* 28,5 (2005):675–691.

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).

Von Neumann, John and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press, 1944).

Wittgenstein, Ludwig, *Tractatus Logico-Philosophicus* (New York: Dover, 1999[1921]).

c\Papers\Modal Logic and Social Norms\CommonKnowledge.tex September 10, 2009