Center for Research on Economic and Social Theory
CREST Working Paper

Common Knowledge and Game Theory

Ken Binmore
Adam Brandeburger

July, 1988
89-06

DEPARTMENT OF ECONOMICS
University of Michigan
Ann Arbor, Michigan 48109

July 1988

Common Knowledge and Game Theory

by Ken Binmore
Economics Department
University of Michigan
Ann Arbor MI 48109

and Adam Brandenburger
Harvard Business School
Boston MA 02163

Common Knowledge and Game Theory

by Ken Binmore and Adam Brandenburger*

I know not what I know not.

St. Augustine, *Confessions*

1. Introduction. It is traditional to introduce the subject of common knowledge with a little story. The following version is quoted from Littlewood's [1953] *Mathematical Miscellany.*

> Three ladies A, B and C in a railway carriage have dirty faces and are all laughing. It suddenly flashes on A: why doesn't B realize C is laughing at her?—Heavens! *I* must be laughable.

A more elaborate version of the story[1] concerns three less frivolous ladies called Alice, Bertha and Cora. Each lady blushes if and only if she knows her own face to be dirty. All three have dirty faces but nobody blushes until a clergyman enters the carriage and remarks that there is a lady in the carriage with a dirty face. It is now impossible that nobody will blush.

If it were true that nobody blushes then Alice could reason as follows:

> *Alice*: Suppose my face is clean so that Bertha and Cora can exclude me from the set of dirty-faced ladies. Then Bertha could argue as follows:
>
> > *Bertha*: Suppose my face is clean so that Cora can exclude me from the set of dirty-faced ladies. Then Cora could argue as follows:
> >
> > > *Cora*: The clergyman's statement tells me that the set of dirty-faced ladies is not empty. Neither Alice nor Bertha are in the set. Hence I am in the set and must blush.
> >
> > *Bertha*: My simulation of Cora's reasoning informs me that she should blush if I have a clean face. Since she has not blushed, I have a dirty face and must blush myself.
>
> *Alice*: My simulation of Bertha's reasoning informs me that she should blush if I have a clean face. Since she has not blushed, I have a dirty face and must blush myself.

But the argument began with the hypothesis that nobody blushes. Since this hypothesis has been contradicted, somebody must blush. The same reasoning, of course, applies however many ladies there might be.

---

* This paper is an enlarged and much revised version of Binmore and Brandenburger [1988]

Why did the ladies in the story have to wait for the clergyman's intervention before blushing? After all, he only told them that there was a dirty-faced lady in the carriage and each lady *already knew* that there were at least two dirty-faced ladies in the carriage. But, to carry the argument through, the ladies need also to know *what the other ladies would know* under various hypothetical circumstances. Thus Alice needs to know what Bertha would know if Alice had a clean face. Moreover, Alice needs to know what Bertha would know about what Cora would know if Alice and Bertha had clean faces. And so on through as many levels of knowing as there are ladies in the carriage. It is *this* information that is supplied by the clergyman's announcement.

The clergyman's announcement ensures that it will be *common knowledge* that a dirty-faced lady is present whenever there actually is a dirty-faced lady present. For an event to be common knowledge, the requirement is not only that everybody knows it, but that everybody knows that everybody knows it, and everybody knows that everybody knows it: and so on.

This is an abstruse-looking definition, but it is an important one. The reason is that any equilibrium notion that incorporates some measure of self-prophesying *necessarily* entails common knowledge requirements of some kind, although these are seldom stated explicitly. In equilibrium, agents optimize given their predictions of the future. The future therefore partly depends on the predictions the agents make. Where do these predictions come from? If they are well-founded, they will be based on an agent's knowledge. In the case of only two agents, this means that agent 1 must know at least something about agent 2's knowledge, because the future depends partly on agent 2's predictions. But a relevant part of agent 2's knowledge will be what agent 2 knows about agent 1's knowledge. And so on.

The background intuition for such discussions is that rational agents cannot "agree to disagree" [Aumann, 1976]. Suppose, for example, that two agents agree to disagree about the probability p that heads will result from tossing a weighted coin. Agent 1 insists that p is almost certainly 0.501 and agent 2 insists that p is almost certainly 0.499. The two agents then have the economist's equivalent of a philosopher's stone, which they can use to resolve all conflicts of interest between them, without the need for compromise on either side. They can agree to resolve any issue entirely in favor of agent 1 if the number of heads appearing in 1,000,001 tosses exceeds the number of tails, and entirely in favor of agent 2 if not.

But, if both agents simultaneously proposed such a deal, would they go through with it? If both agents are rational and have reason to believe the other is rational, then there are grounds for supposing

otherwise. After the deal has been proposed, each agent now knows something about what the other agent knows. This *new* information should lead both to revise their estimates of p.

The story of the dirty-faced ladies already embodies the essential point. Mutual observation of agents' behavior can lead to information becoming common knowledge that was not common knowledge before.

Perhaps the most important area in which such problems arise is in the study of rational expectations equilibria in the trading of risky securities. How can there be trade if everybody's willingness to trade means that everybody knows that everybody expects to be a winner? (see Milgrom/Stokey [1982] and Geanakoplos [1988].) Since risky securities are traded on the basis of private information, there must presumably be some "agreeing to disagree" in the real world. But to assess its extent and its implications, one needs to have a precise theory of the norm from which "agreeing to disagree" is seen as a deviation.

The beginnings of such a theory are presented here. Some formalism is necessary in such a presentation because the English language is not geared up to express the appropriate ideas compactly. Without some formalism, it is therefore very easy to get confused. However, nothing requiring any mathematical expertise is to be described.

2. Knowledge. The account begins with a set Ω of possible *states of the world.* To keep things simple Ω will always be assumed to be a *finite* set when formal matters are under discussion. A subset E of Ω is to be identified with a possible *event.*

To discuss what an individual i knows, his *knowledge operator* K is introduced. For each event E, the set KE is the set of states of the world in which individual i knows that E has occurred. Or, more briefly, KE is the event that i knows that E has occurred.

For example, in the story of the dirty-faced ladies, a state space Ω with eight states is required. The eight states of the world will be numbered as indicated in figure 1. The event that Alice's face is

states	1	2	3	4	5	6	7	8
A's face	clean	dirty	clean	clean	dirty	dirty	clean	dirty
B's face	clean	clean	dirty	clean	dirty	clean	dirty	dirty
C's face	clean	clean	clean	dirty	clean	dirty	dirty	dirty

Figure 1

dirty is then $D = \{2,5,6,8\}$. If blushing were not part of the story, she would ***know*** that her face was dirty after the clergyman's announcement only in state 2. Writing K_A for Alice's knowledge operator, it would then be true that $K_A D = \{2\}$.

What should be assumed about the knowledge operator? The properties usually considered are listed below. All but the final property ($K4$) will be taken for granted throughout the paper. Property ($K4$) is more controversial and its use will be postponed until section 5.

$(K0)$	$K\Omega = \Omega$	
$(K1)$	$K(E \cap F) = KE \cap KF$	
$(K2)$	$KE \subseteq E$	(axiom of knowledge)
$(K3)$	$KE \subseteq K^2E$	(axiom of transparency)
$(K4)$	$(\sim K)^2E \subseteq KE$	(axiom of wisdom)

The first two properties are book-keeping assumptions. Notice that it follows from $(K1)$ that, if $E \subseteq F$, then $KE \subseteq KF$. Bacharach [1987] refers to $(K2)$ as the "axiom of knowledge" on the grounds that it expresses the requirement that one can only really ***know*** that something has happened if it actually ***has*** happened. For similar reasons, Geanakoplos [1988] refers to "non-deluded" individuals in this connection. In $(K3)$, K^2E stands for $K(KE)$ and hence is the event that the individual knows that he knows that E has occurred. Thus $(K3)$ means that the individual cannot know that something has happened without knowing that he knows it. This explains Bacharach's [1987] use of the terminology "axiom of transparency" for $(K3)$. Notice that $(K2)$ and $(K3)$ together imply that $KE = K^2E$. In $(K4)$, $\sim KE$ stands for the complement of KE, and hence the condition requires that if the individual does *not* know that he does *not* know something, then he knows it. Discussion of this "axiom of wisdom" is postponed until it is used in section 4.

A *truism* T will be defined to be an event that cannot occur without the individual knowing that it has occurred. This translates into symbols as $T \subseteq KT$. For example, the event that a dirty-faced lady is in the railway carriage is a truism for Alice provided that the clergyman can be relied upon to draw her attention to the fact whenever it happens and never when it does not.

If one thinks of truisms as embodying the essence of what is involved in making a direct observation, then there is a sense in which all knowledge is derived from truisms. The following proposition expresses the idea formally.

Proposition 1. *An event E is known to have occurred ($\omega \in KE$) if and only if a truism T has occurred which implies E ($\omega \in T \subseteq E$).*

The proposition has little content but will serve to illustrate the use of properties $(K0)$ through $(K3)$. Note first that, by $(K2)$, the criterion for a truism may be simplified to

$$T = KT.$$

Since $KE = K^2E$ for all events E, it follows that truisms are just the events of the form KE, where E is any subset of Ω.

Returning to proposition 1, observe that, if $\omega \in KE$, then $T = KE$ is a truism satisfying $\omega \in T \subseteq E$. On the other hand, if $\omega \in T \subseteq E$, then $KT \subseteq KE$ and so $\omega \in T = KT \subseteq KE$. Thus $\omega \in KE$.

3. Common knowledge . The formulation of common knowledge in terms of everybody knowing that everybody knows and so on, was first given by Lewis [1969] in a philosophical study of conventions. Lewis attributes the basic idea to Schelling [1960]. Aumann [1976] came up with the idea independently in a somewhat different context. His formulation is important because it provides a precise characterization of when an event is common knowledge that does not require thinking one's way through an infinite regress.

For the case of three dirty-faced ladies, the (everybody knows) operator is defined by

$$\text{(everybody knows)}E = K_A E \cap K_B E \cap K_C E.$$

The event that everybody knows that everybody knows E is then

$$\text{(everybody knows)}^2 E$$

and the event that everybody knows that everybody knows that everybody knows E is

$$\text{(everybody knows)}^3 E.$$

With these preliminaries out of the way, it is now possible to define the event

$$\text{(everybody knows)}^\infty E$$

to be the intersection of all sets of the form $\text{(everybody knows)}^n E$.

Lewis' [1969] criterion for common knowledge can now be expressed formally. An event E is common knowledge when ω occurs if and only if

$$\omega \in \text{(everybody knows)}^\infty E.$$

Some observations about the (everybody knows) operator will be helpful in getting to Aumann's [1976] criterion. It is being assumed that each individual's knowledge operator (K_A, K_B and K_C in the case of the three dirty-faced ladies) satisfies ($K0$) through ($K3$). It follows that the operator K = (everybody knows) satisfies ($K0$) through ($K2$). (Usually ($K3$) will not be satisfied.) From ($K2$) one may conclude that the sets (everybody knows)$^n E$ are shrinking in that each contains its predecessor. Since Ω is a finite set, a positive shrinkage can only occur a finite number of times. Thus, for some N,

$$\text{(everybody knows)}^n E = \text{(everybody knows)}^\infty E$$

for all $n \geq N$.

This point is made to facilitate checking that the *common* knowledge operator K = (everybody knows)$^\infty$ satisfies not only ($K0$) through ($K2$), but ($K3$) as well. The following analog to proposition 1 must therefore be true.

Proposition 2. *An event E is common knowledge ($\omega \in$ (everybody knows)$^\infty E$) if and only if a common truism T has occurred which implies E ($\omega \in T \subseteq E$).*

A common truism is, of course, an event T that satisfies T = (everybody knows)$^\infty T$. However, the infinite regress in this definition can be eliminated. The criterion

$$T = \text{(everybody knows)}T$$

is equivalent[2] and simpler. This in turn can be replaced by the even simpler criterion of the following proposition.

Proposition 3. *An event T is a common truism if and only if it is a truism for each individual separately.*

A proof will be given for the case of two dirty-faced ladies. If T is a truism for each lady separately, then $T = K_A T$ and $T = K_B T$. Thus $T = T \cap T = K_A T \cap K_B T$ = (everybody knows)T. If T is a common truism, then $T = K_A T \cap K_B T \subseteq K_A T \subseteq T$ (by $K2$). Hence $T = K_A T$ and so T is a truism for Alice and a similar argument shows T to be a truism for Bertha.

(Mathematicians may prefer to express proposition 3 in terms of the set $\mathcal{T}$ of common truisms and the sets $\mathcal{T}_A$, $\mathcal{T}_B$ and $\mathcal{T}_C$ of individual truisms. It then asserts that $\mathcal{T} = \mathcal{T}_A \cap \mathcal{T}_B \cap \mathcal{T}_C$. One may add that all of these sets are topologies, being closed under intersections and unions (recall that Ω is finite). Thus $\mathcal{T}$ is the finest common coarsening of the topologies $\mathcal{T}_A$, $\mathcal{T}_B$ and $\mathcal{T}_C$.)

The importance of what have been called common truisms in the preceding discussion has been emphasized by a number of authors. Monderer and Samet [1988] call them "evidently known events." Geanakoplos [1988] prefers "necessarily known events." Milgrom [1981] speaks of "public events."

4. Possibility. After ω occurs, an individual will know that certain states are impossible. For example, if T is any truism containing ω, then he will know that states outside T are impossible. Thus the set $P(\omega)$ of *possible states* when ω occurs, lies inside each truism T containing ω. But $P(\omega)$ must itself be a truism because the individual cannot evade knowing what he regards as possible. Thus $P(\omega)$ is the *smallest*[3] truism containing ω.

Examples of possibility sets can be found by returning to the story of the dirty-faced ladies. The state space Ω is described in figure 1. Figure 2 below shows the possibility sets for each lady *before* the clergyman provides any information. For example, whatever Alice learns about the faces of the other ladies, it remains possible for Alice that her own face is either clean or dirty. Thus $P_A(1) = P_A(2) = \{1,2\}$.

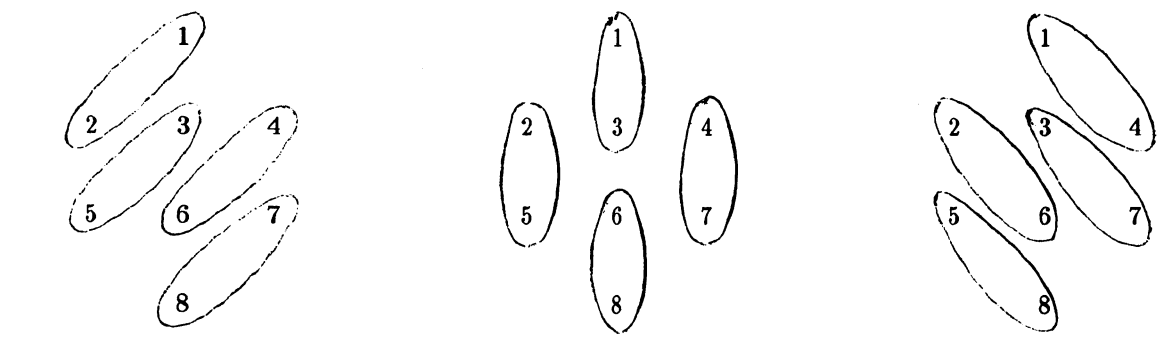


Figure 2.

Figure 3 shows the possibility sets, *after* the clergyman's announcement but *before* any blushing takes place, on the assumption that all are aware that the clergyman *invariably and reliably* comments on the presence of a dirty face. When Alice sees two clean faces, she can now deduce the state of her own face from whether or not the clergyman makes an announcement. Thus $P_A(1) = \{1\}$ and $P_A(2) = \{2\}$.

A more tricky example is obtained by having the clergyman announce that a dirty-faced lady is present if and only if *two* or more dirty-faced ladies are actually present. The ladies know how he will behave *except* in the case when he sees precisely *two* dirty-faced ladies in which case they are *unsure* whether or not he will make an announcement. Figure 4 illustrates the possibility sets.

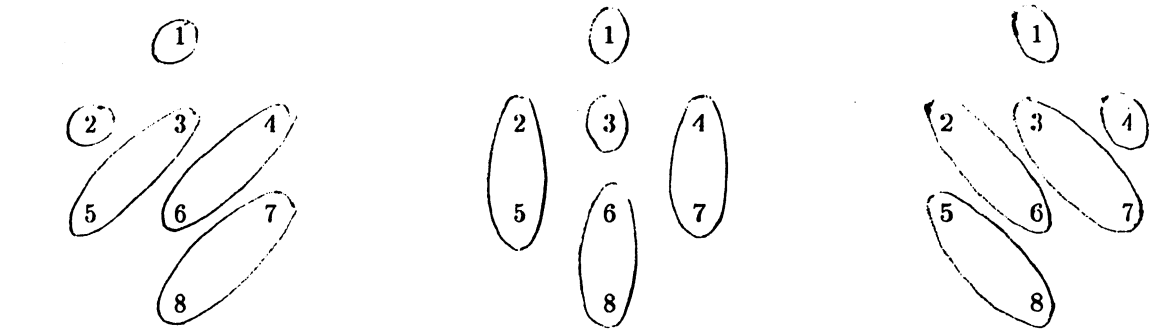


Figure 3.

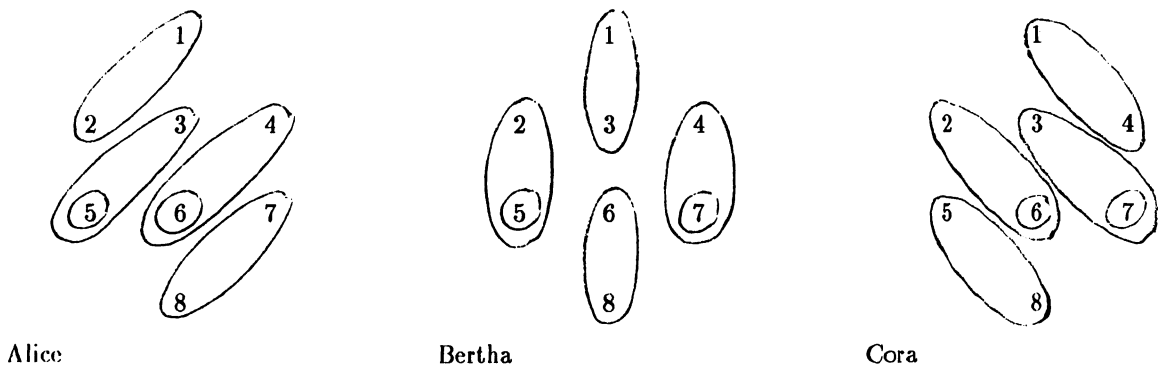


Figure 4.

In the case when Bertha's face is dirty and Cora's is clean, Alice can deduce from the fact that an announcement is made that her face is dirty. Moreover, the clergyman actually will announce that there is a dirty face when Alice's face is dirty. Thus $P_A(\mathfrak{5}) = \{\mathfrak{5}\}$. But $P_A(\mathfrak{3}) = \{3, 5\}$ because, if $\mathfrak{3}$ occurs, then the clergyman will make no announcement and, from this eventuality, Alice can deduce nothing.

Because $P(\omega)$ is the *smallest* truism containing ω, proposition 1 can be rephrased as

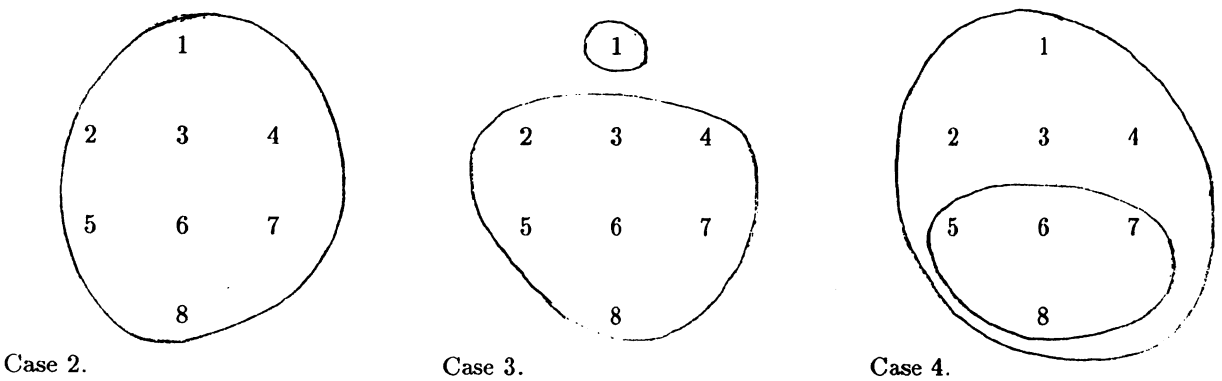
Proposition 4. *When ω occurs, E is known if and only if $P(\omega) \subseteq E$ (i.e. everything possible implies E).*

Proposition 2 can be similarly rephrased. In its rephrased form it is Aumann's [1976] criterion for an event to be common knowledge. Some preliminary explanation is necessary to make this point.

The smallest *common* truism containing ω will be denoted by $M(\omega)$. One may think of this as the set of states deemed to be possible by the community as a whole. It includes not only each state that somebody thinks to be possible, but also each state that somebody thinks somebody else thinks to be possible, and so on.

To find $M(\omega)$, one may use proposition 3 and the fact that truisms are simply unions[4] of possibility sets. To find $M(2)$ in figure 3, for example, look for the smallest set containing 2 which is simultaneously the union of possibility sets belonging to each of Alice, Bertha, and Cora separately. Since $M(2)$ contains 2, it must contain 5 because of Bertha. Hence it must contain 8 because of Cora, and therefore 7 because of Alice. Proceeding in this way one finds that $M(2) = \{2,3,4,5,6,7,8\}$.

Figure 5 illustrates the communal possibility sets for each of the situations illustrated in figures 2, 3 and 4.



Case 2. Case 3. Case 4.

Figure 5.

(Mathematicians may choose to call M the *meet* of P_A, P_B and P_C by analogy with the usage for partitions.)

Proposition 5. [Aumann, 1976] *When ω occurs, E is common knowledge if and only if $M(\omega) \subseteq E$ (i.e. everything communally possible implies E).*

It follows, for example, that in the information set-up described by figure 3, no matter what happens, the set $D = \{2,5,6,8\}$, which represents the event that Alice's face is dirty, never becomes common knowledge.

In the story of the dirty-faced ladies given in section 1, more information was provided than is summarized in figure 3. It was also given that ladies blush when they know their faces to be dirty. Figures 6, 7 and 8 give three different knowledge configurations that are consistent with the story.

The shaded regions indicate states in which a lady has a dirty face. If each state in a possibility set is shaded, a lady will blush if one of these states occurs. Otherwise she will not blush.

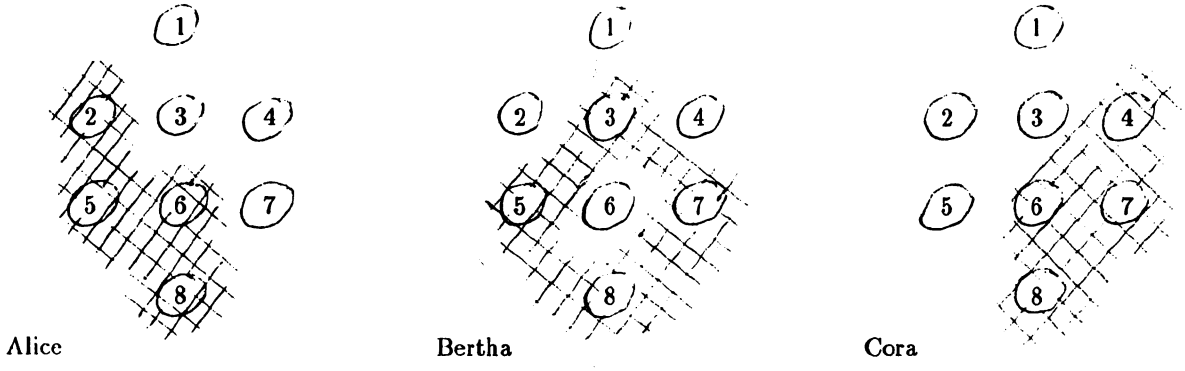


Figure 6.

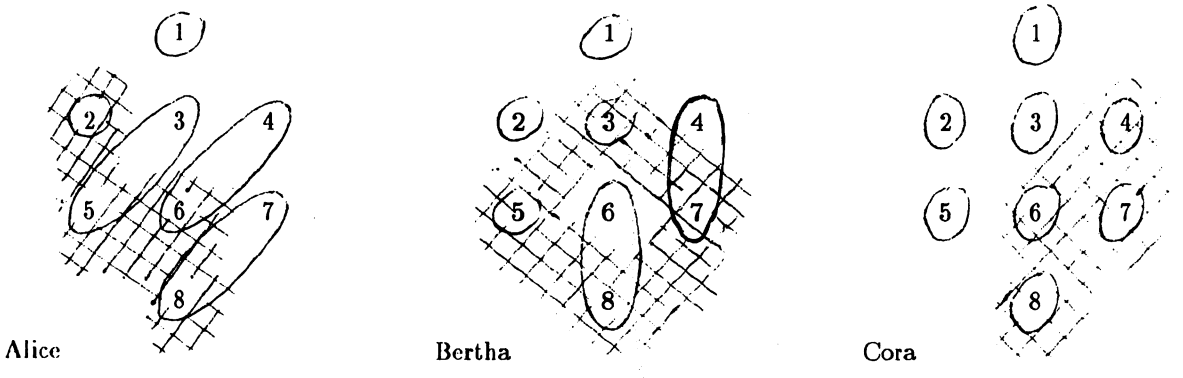


Figure 7.

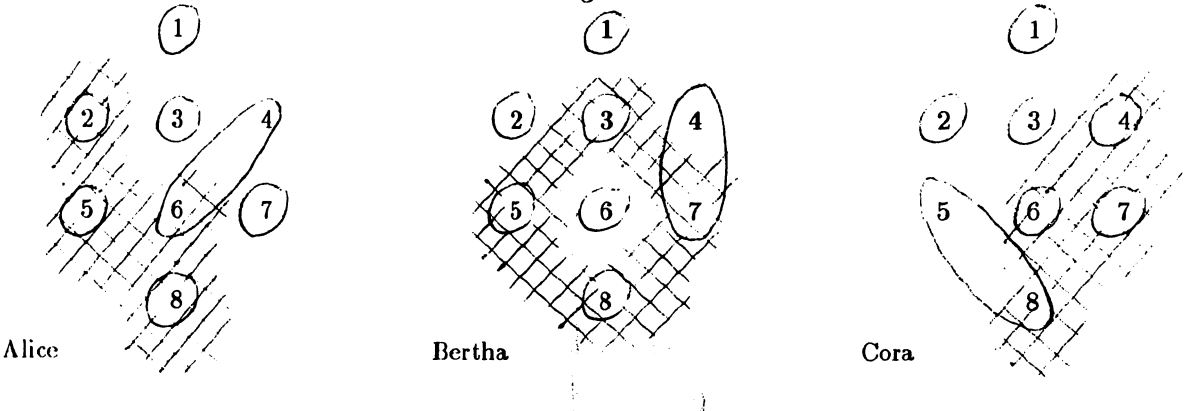


Figure 8.

Consider figure 7 by way of example. It has to be checked that the configuration is consistent with the data given in the story. Suppose, for instance, that event 6 occurs. Cora will then blush, but not Alice or Bertha. But does the fact that Cora blushes not convey information to Alice and Bertha?

Notice that $P_A(6) = \{4, 6\}$. Thus, if 6 occurs, Alice knows that *either* 4 *or* 6 have occurred. But Cora would blush in *both* cases. Hence the fact that Cora blushes does not tell Alice anything. Nor does the fact that Bertha does *not* blush. She would blush neither in case 4 nor in case 6.

In the story, state 8 is what actually occurs. Thus, for the configuration of figure 7, Cora blushes but Alice and Bertha do not. Observe that $M(8) = \{4, 6, 7, 8\}$, which is precisely the event that Cora has a dirty face. This becomes common knowledge whenever it occurs. (It is, of course, the smallest common truism containing 8.)

A final lesson can be extracted from the dirty-faced ladies. It concerns the question of *how* things come to be known. In the story, no information is offered on the *process* by means of which the ladies learn. A conclusion is reached by indirect reasoning. For example, figure 3 is inconsistent with what is given about blushing behavior because, if 2 occurs, then Alice will blush whereas, if 5 occurs, she will not. Thus Bertha can distinguish between 2 and 5 and so it is incorrect to assert that $P_B(2) = P_B(5) = \{2, 5\}$. Only certain knowledge configurations are consistent with the data of the story. Those illustrated in figures 6, 7 and 8 are examples. The argument given in section 1 shows that, for each such example, at least one of the ladies will blush when 8 occurs. (In fact, a lady will blush whenever a dirty-faced lady is actually present.)

How might figure 7 arise? One "explanation" postulates that the opportunity to blush rotates among the ladies, starting with Alice. The knowledge situation then evolves as follows:

1. The ladies observe the faces of their companions. This leads to figure 2 in which $M(8) = \Omega$.
2. The clergyman announces the presence of a dirty face whenever a dirty-faced lady is present. This leads to figure 3 in which $M(8) = \{2, 3, 4, 5, 6, 7, 8\}$.
3. Alice has the opportunity to blush (but not Bertha or Cora). This leads to figure 9 in which $M(8) = \{3, 4, 5, 6, 7, 8\}$.
4. Bertha has the opportunity to blush (but not Cora). This leads to figure 7 in which $M(8) = \{4, 6, 7, 8\}$.
5. Cora has the opportunity to blush and does so when 8 occurs. But figure 7 remains unchanged and continues unchanged even if further rounds of blushing are introduced.

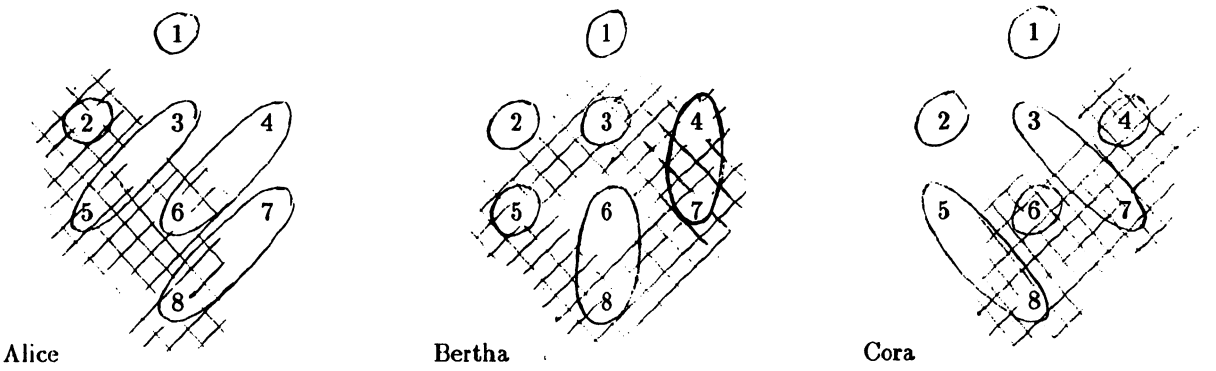


Figure 9.

An "explanation" of figure 6 can be obtained by supposing that all three ladies have the opportunity to blush precisely one second after the clergyman's announcement, and then again precisely two seconds after, and so on. Figure 8 can be "explained" by having Alice and Bertha blush simultaneously in rotation with Cora.

More on the question of how knowledge evolves over time will appear in section 12.

5. Information partitions. In the previous section, the idea of a possibility set was developed taking the knowledge operator K as a primitive. When K satisfies properties $(K0)$ through $(K3)$, possibility sets necessarily satisfy

$$(P2) \quad \omega \in P(\omega)$$

$$(P3) \quad \zeta \in P(\omega) \Rightarrow P(\zeta) \subseteq P(\omega)$$

for all ζ and ω. The first of these simply says that the state that actually occurs is always regarded as possible. The second is a little more complicated. It says that, if something is possible, then anything that would be regarded as possible in that state must also be regarded as possible in the current state.

Section 4 makes it clear that, for practical purposes, the possibility operator P is much more convenient to work with than the knowledge operator K. One can, in fact, pass back and forward between the two without difficulty since $(K0)$ through $(K3)$ hold if and only if $(P2)$ and $(P3)$ hold. If one begins with P, the operator K may be defined by

$$KE = \{\omega : P(\omega) \subseteq E\}$$

In applications, it is usually taken for granted that the possibility operator *partitions* the state space. This means that distinct possibility sets have no points in common. Figures 2 and 3 illustrate six possible partitions of the state space $\Omega = \{1,2,3,4,5,6,7,8\}$. Figure 4 illustrates three situations in

which Ω is not partitioned. (Notice that $P_A(5) = \{5\}$ and $P_A(3) = \{3,5\}$. Thus $P_A(5)$ and $P_A(3)$ are distinct sets but they have the point 5 in common.)

In game theory, possibility sets which partition the state space are called *information sets*, following Von Neumann and Morgenstern. Figure 10 illustrates a simple *game tree*. When player II gets to move, her state space is $\Omega = \{1,2,3\}$. The information set enclosing 2 and 3 indicates that, if player I chooses M or R, then player II will not know which of these were chosen.

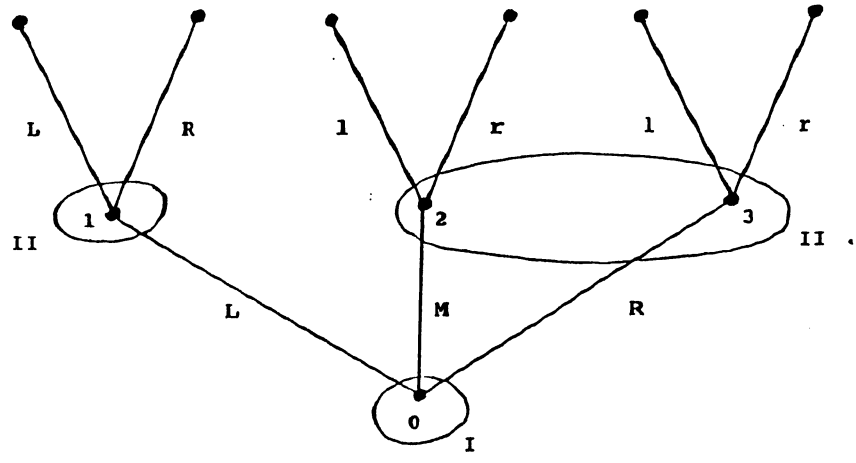


Figure 10

In formal terms, the requirement that the possibility sets partition Ω may be expressed as

$$(P4) \quad \zeta \in P(\omega) \Rightarrow P(\zeta) = P(\omega)$$

for all ζ and ω. Conditions $(P2)$ through $(P4)$ are equivalent to $(K0)$ through $(K4)$. Thus $(K4)$, the controversial "axiom of wisdom," appears on the scene. The implications are discussed in the next section. This section closes with a comment on the common knowledge operator.

If P_A, P_B and P_C partition Ω, then so does their meet M. Consider, for example, cases 2 and 3 in figure 5 which illustrate the meets of the partitions of figures 2 and 3. A consequence is that, if each individual's knowledge operator satisfies $(K0)$ through $(K4)$, then so does the common knowledge operator.

6. Small worlds. There is something paradoxical in the preceding section. The use of information partitions in a situation like the game of figure 10 seems more than harmless: it seems inevitable. On the other hand, using information partitions is equivalent to incorporating ($K4$) into the system. But do we really want to claim that, whenever we don't know that we don't know something, then we know it?

Let us return to Alice in figure 4. The story that goes with this will be told again. The clergyman may be in a good mood or a bad mood. If in a good mood, he announces the presence of a dirty-faced lady if and only if all three ladies have dirty faces. If in a bad mood, he announces the presence of a dirty-faced lady if and only if at least two ladies have dirty faces. In figure 4, the clergyman is in a bad mood but Alice thinks he may either be in a good mood or a bad mood. The result is a system of possibility sets that do not partition Ω.

One reaction is that the state space Ω has been improperly specified since it pays no attention to the mood of the clergyman although this is clearly relevant. If $1, 2, 3, \ldots, 8$ are as previously, but with the extra understanding that the clergyman is in a good mood, while I, II, III,...,VIII are the corresponding states in which the clergyman's mood is bad, then Alice's knowledge configuration in figure 4 may be replaced by that of figure 11.

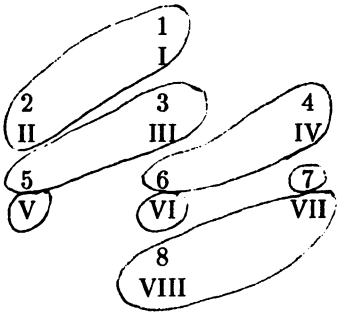


Figure 11

The result is an *information partition* of a *new* state space $\Omega^* = \{1,2,\ldots 8,I,II,\ldots,VIII\}$. One can therefore regard the failure to arrive at an information partition in figure 4 as being a consequence of a wrong modeling judgement being made about the relevant state space. Something which matters, namely the mood of the clergyman, has been omitted from the description of a state. Once this omission has been rectified, by constructing a new state space Ω^* consisting of more fully described states, the problem disappears.

For this kind of reason, Bayesian decision theorists often urge that care be taken to begin with a state space Ω in which the description of a state is *all-inclusive*. This is sound advice if everything that is to be included can be tagged and enumerated at the outset.

To see why this proviso is attached, consider the following argument in defense of information partitions. Recall that the formal requirement for an information partition is $(P4)$ of section 5.

Suppose that an individual finds himself holding the view that only states in $P(\omega)$ are possible. If $P(\zeta) \neq P(\omega)$, he can then deduce that it is impossible that ζ has occurred, because, if ζ had occurred, then his current view of what is possible would differ from the view he actually has. Thus $P(\zeta) \neq P(\omega)$ implies that $\zeta \notin P(\omega)$ or, what is the same thing, $\zeta \in P(\omega)$ implies $P(\zeta) = P(\omega)$. This is the requirement $(P4)$ for an information partition. Now tell the same story but with "possible" replaced everywhere by "conceivable."

From $P(\zeta) \neq P(\omega)$, the conclusion that ζ is inconceivable should follow. But if ζ is truly inconceivable, how does the individual manage to "know" the set $P(\zeta)$ of those states he would regard as conceivable if the inconceivable state ζ were to occur? If the individual is assumed only to be able to explore the implications of conceivable states, then only $(P3)$ can be justified. The latter only requires that anything that is conceivable at a conceivable state be conceivable.

Samet [1987] observes that, without information partitions, an individual must be "ignoring his ignorance" to some extent. To see this point, take complements on both sides of $(K4)$ to obtain $\sim KE \subseteq K(\sim KE)$. This says that, if an individual is ignorant of something, then he *knows* he is ignorant of it. If information partitions fail to exist because $(K4)$ is not satisfied, it follows that there is at least one event of which the individual is ignorant but does not know it. The argument is then that, since he could examine the contents of his own mind and enumerate those events that he knows and those that he does not know, if he does not know that he is ignorant of an event, then it is because he ignores his ignorance. This seems thoroughly reprehensible. On the other hand, who could be blamed for being unaware of being unaware of something?

Two versions of the same story have been told to make it clear that the English language has no difficulty in distinguishing between "closed universe" and "open universe" situations. By a closed universe is meant one in which all the possibilities can be exhaustively enumerated in advance, and all the implications of all possibilities explored in detail so that they can be neatly labeled and placed in their proper pigeon-holes. In discussing the foundations of Bayesian decision theory, Savage [1954] uses the term "small world" for such a state space. Statisticians' sample spaces are invariably "small worlds" in this sense, with the set $\Omega = \{1,2,3,4,5,6\}$ of possible ways a die can fall serving as the archetypal example. Within a small world, the arguments in favor of $(K0)$ through $(K4)$, and of standard Bayesian principles, are very powerful (although, as the example at the beginning of this section shows, a bad modeling judgment can generate a world which is *too* small!)

However, there are difficulties with working in a small world. Certain things can only be expressed *informally*. For example, in game theory, it is typically understood that the structure of the game tree is to be common knowledge. But there is no way of expressing this within a formalism that takes as its small world the set $\Omega = \{0,1,2,3\}$ of decision nodes in figure 10. Just as the use of information partitions has been defended by appealing to unformalized understandings about what an individual should be expected to know as a consequence of examining the workings of his own mind, so the understandings about what is informally regarded as common knowledge will appear on the scene only when the type of game-theoretic analysis employed comes under attack.

Of course, the larger the small world, the more that can be expressed up-front in formal terms. Papers that explore what can be done in this direction are Aumann [1987], Bacharach [1985, 1987], Brandenburger and Dekel [1985], Kaneko [1987], Mertens and Zamir [1985], Samet [1985], Shin [1987] and Tan and Werlang [1985]. There is also a large and relevant philosophical literature on the theory of knowledge and epistemic logic (see Halpern [1986] and the references therein).

The deep issues discussed in these papers certainly need to be thoroughly explored, but perhaps some healthy skepticism is appropriate about how useful their conclusions are likely to prove in the foundations of game theory. Any formal characterization of how we acquire knowledge is bound to be an over-simplification and hence will generate distortions if pushed beyond its limitations. In particular, one has to expect distortions if "closed universe" methodologies are applied to "open universe" problems. The risk is greatest when attempts are made to interpret a state as incorporating a specification of the universe that is totally all-embracing. To know a state then includes knowing, not only everything there is to know about the state of the physical world, but also everything there is to know about everybody's *state of mind*, including their knowledge and beliefs.

The self-reference implicit in such an interpretation brings Gödel's theorem to mind. Recall that this says that any sufficiently complex formal deductive system cannot be complete unless it is inconsistent. That is to say, in the world of theorem-proving, the "open universe" is a *necessary* fact of life with which one has to learn to live. One is therefore perhaps entitled to be suspicious of theories of knowledge in which this fact of life is somehow evaded.

7. Algorithmic knowledge. The previous paragraph draws an analogy between proving a theorem and acquiring a piece of knowledge. This section briefly explores this idea a little further. The point is that, if what one knows is known in virtue of its being the result of applying a properly defined procedure or algorithm, then a very much more critical eye needs to be cast on assumptions $(K0)$ through $(K4)$ if the "small world" approach advocated in the previous section is not to be followed.

The notion that questions are to be settled algorithmically is captured by requiring that there exist a computer programmed for this purpose. The type of computer used in the argument is called a Turing machine. One may think of this as a machine with a *finite* program but with an indefinitely large storage capacity.

In preceding sections, doubts have been expressed about $(K4)$—the "axiom of wisdom." The purpose of what follows is to direct suspicion at the very much more fundamental $(K2)$—the "axiom of knowledge." This says that one cannot know an event that has not occurred. The corresponding condition for possibility sets is $(P2)$—i.e. $\omega \in P(\omega)$. This means that any state that actually occurs cannot be regarded as impossible.

Suppose that, in each state ω, possibility questions are resolved by a Turing machine $S = S(\omega)$ that sometimes answers NO to properly coded questions of the form, "Is it possible that . . . ?" If it answers otherwise, or not at all, possibility is conceded. (Timing issues are ignored.) Consider some specific question concerning the Turing machine N. Let the coded form of this question be $[N]$. Consider next the question, "Is it possible that Turing machine M would output NO when offered the input $[M]$?" Let the coded form of this question be $\{M\}$. Let T be a machine that, when offered $[M]$ outputs $\{M\}$. Then machine $R = ST$, constructed by composing S and T, responds to $[M]$ as S responds to $\{M\}$.

Suppose that R responds to $[R]$ with NO. Then S reports that it is impossible that $[R]$ responds to $[R]$ with NO. Thus $\omega \notin P(\omega)$. If it is to be denied that this can occur for any $\omega \in \Omega$, then it must be that R does *not* respond to $[R]$ with NO in *any* state. Nevertheless, S *always* reports that it is possible that R will respond to $[R]$ with NO. But then S will fail to report a piece of available knowledge.

Such an argument (which is a version of the halting problem for Turing machines loosely adapted from Binmore [1984]) can only serve to raise questions about how meaningful it is to work in terms of all-embracing descriptions of the universe and superhumanly endowed individuals who process such data. This goes also for more formal attempts in the same direction; such as that of Shin [1987]. But perhaps it will be enough to explain why this paper persists with a "small world" view in spite of the attractions of more ambitious foundational positions.

8. Agreeing to disagree? In specifying who knows what during the course of a game, it is usual to take $(K0)$ through $(K4)$ for granted and therefore to work with information partitions. This practice will be followed throughout the paper (except where otherwise stated), since the game trees with which we shall be concerned, like that considered already in figure 10, constitute very small worlds indeed.

If everybody always knows what has happened previously in the game, as in Chess, then the game is one of *perfect information*. In such a game, each information set encloses only a single decision node. Poker, on the other hand, is a game of *imperfect information*. The players do not know what other players have been dealt when they make their bets.

A game of perfect information must be distinguished from a game of *complete information*. All games of perfect information are games of complete information, but so are all games of imperfect information. The information that is complete in a game of complete information is the information about the *structure of the game*. This can be classified under two headings: information about the *rules of the game* and information about the tastes and beliefs of the *players*. The former includes what the game tree is, where the information sets are placed and who makes what decisions. The latter subsumes the players' Von Neumann and Morgenstern utilities for the possible outcomes of the game (i.e., the terminal nodes of the game tree) and the probabilities that the players attach to the various chance events that may occur in the game which lie outside the control of the players (as when the cards are shuffled and dealt in Poker). All this information is taken to be *common knowledge* among the players in a game of complete information. (Since everybody is then necessarily in the *same* informational state before any moves are made, a game of complete information is also commonly referred to as a "game of symmetric information." However, this should not be taken to imply that games of incomplete information cannot have a symmetric structure.)

Games of *incomplete information*, in so far as they can be dealt with realistically at all, are studied by reducing them to games of complete information using an ingenious methodology introduced by Harsanyi [1967/68]. For example, I may be in doubt whether my opponent at Chess is really aiming to win or whether he actually wants to lose. Harsanyi would say that the game should then not be seen as a two-player game of perfect information, but as a game of imperfect information with at least three players in which the opening event is a *chance move* that selects my opponent from a menu of possible opponents with various different preferences.

A standard procedure is to take for granted that everybody attaches the *same* probabilities to the possible outcomes of such a chance move. This does not need much justifying when the chance move is that of shuffling and dealing at Poker. But matters are less clear-cut with a chance move like that in the story attributed above to Harsanyi.

Recall that, in a game of complete information, the probabilities the players attach to the possible outcomes of chance events are *common knowledge*. Aumann [1976] asked whether rational players can "agree to disagree" under such circumstances by maintaining different probabilities for the same event.

His answer is that this is impossible. Intuitively, each player will use his knowledge of the other players' estimates of the probabilities to refine his own estimate and this will continue until all the estimates are the same. (See Bacharach [1985] for a more general expression of the idea.)

However, Aumann requires a strong assumption that will be important throughout the rest of the paper. He likes to refer to this as the "Harsanyi doctrine." (See section 13). The idea is that the players' informational status can be modeled as follows. Before the rational players to be considered received any data at all, they were all in the *same* position and therefore assigned the *same* probabilities $\text{prob}(\omega) > 0$ to the states $\omega \in \Omega$. Moreover, these *prior* probabilities are common knowledge among the players. The Harsanyi doctrine is therefore that there is common knowledge of a common prior. Later, the players had different experiences which led them to revise their probabilities. The current probabilities they attach to the states $\omega \in \Omega$ are therefore posterior probabilities obtained by Bayesian updating from the given common prior. It is in this sense that the posterior probabilities q_A, q_B and q_C for the event F attributed to Alice, Bertha and Cora in the proof of the following proposition are to be understood.

Proposition 6. [Aumann, 1976] *If each agent's posterior probability for an event F is common knowledge when $\omega_0 \in \Omega$ occurs, then it is common knowledge that these probabilities are equal.*

To see why this is true, focus on Alice. If ω_0 occurs, her posterior probability for F is q_A. The event that Alice actually observes when ω_0 occurs is $P_A(\omega_0)$. Thus

$$q_A = \text{prob}(F|P_A(\omega_0).$$

It follows from proposition 3 that $M(\omega_0)$ is a union of a collection of Alice's information sets. Let these information sets be $Q_1, Q_2, \ldots, Q_k$. One of these sets is, of course, $P_A(\omega_0)$.

Observe that

$$q_A = \text{prob}(F|Q_j)$$

for each $j = 1, 2, \ldots k$. The reason[5] is that, if prob $(F|Q_j)$ were not equal to q_A then, as soon as q_A became common knowledge, it would become common knowledge that Q_j had *not* happened, and hence the event $\sim Q_j$ would be common knowledge. But proposition 5 then implies that $M(\omega_0) \subseteq \sim Q_j$, and so $Q_j \subseteq Q_1 \cup Q_2 \cup \ldots Q_k = M(\omega_0) \subseteq \sim Q_j$ which is a contradiction.

It then only remains to observe that

$$\begin{aligned} \text{prob}(M(\omega_0) \cap F) &= \text{prob}(Q_1 \cap F) + \cdots + \text{prob}(Q_k \cap F) \qquad (8.1) \\ &= \text{prob}(F|Q_1)\,\text{prob}(Q_1) + \cdots + \text{prob}(F|Q_k)\,\text{prob}(Q_k) \\ &= q_A(\text{prob}\,Q_1 + \cdots + \text{prob}\,Q_k) \\ &= q_A\,\text{prob}\,M(\omega_0) \end{aligned}$$

Thus, since prob $M(\omega_0) > 0$, $q_A =$ prob $(F|M(\omega_0))$. But the same is true for q_B and q_C and therefore

$$q_A = q_B = q_C$$

as required.

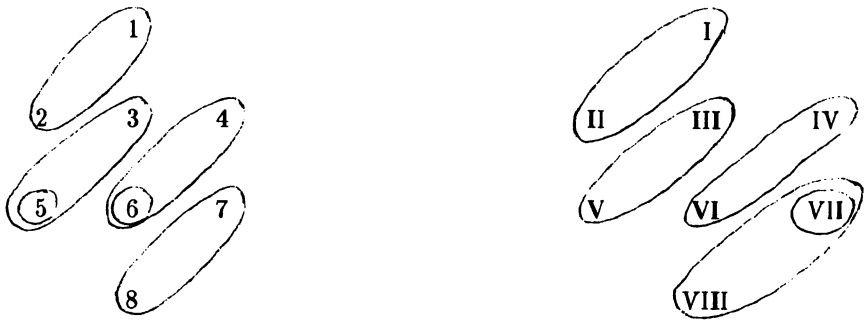
Samet [1987] has pointed out that proposition 6 does not depend on having information partitions. The result survives without $(K4)$. It is easy to see why. In the preceding proof, it is only at step (8.1) that information partitioning is required. For this step to be valid, the possibility sets $Q_1, Q_2, \ldots, Q_k$ must not overlap. Without $(K4)$, the sets may overlap. In the case $k = 2$ for example, one must then replace (8.1) by

$$\text{prob}(M(\omega_0) \cap F) = \text{prob}(Q_1 \cap F) + \text{prob}(Q_2 \cap F) - \text{prob}(Q_1 \cap Q_2 \cap F)$$

All is then well provided $Q_1 \cap Q_2$ can be expressed as the union of possibility sets. But this is guaranteed because the intersection of two truisms is again a truism. Thus, in Samet's words, we cannot "agree to disagree" even if we "ignore our ignorance."

But a pause for thought is perhaps appropriate. In a world in which we "ignore our ignorance", does it really make sense to proceed as though probabilistic assertions can sensibly be formulated and the manipulated according to the usual rules? Is it not a background presumption of probabilistic decision theories that information sets partition the universe? This is a question which returns us to the "small world" issues of section 6 and anticipates some of what is to be said in section 13.

Any doubts that may arise seem amply justified, as the following example is intended to indicate. Return to Alice's informational configuration of Figure 4, reproduced as the "Arabic world" on the left in Figure 12. Recall that the clergyman announces that a dirty-faced lady is present if and only if two or more such ladies actually are present, but Alice only knows for sure what he will do when the actual number of dirty-faced ladies differs from two. Suppose that Alice attaches equal prior probabilities to each of the states in Ω. What probability should she attach to to the event E that her own face is dirty, given that she sees that Bertha is dirty and Cora is clean?



Alice in Arabia Alice in Rome

Figure 12.

If this is posed as the question, "What is prob({3}|{3,5})?" then a conventional conditional probability calculation yields the answer 1/2. But does this make any real sense?

Consider the following alternative line of reasoning. Before anything happens, Alice does not "know" the informational structure of the world in which she lives. To be precise, she does not know whether or not the clergyman tells when there are exactly two dirty–faced ladies present. She may find this out. For example, she will learn that the clergyman *does* tell if state 5 occurs. But, if state 3 occurs, then it will remain possible for her that the clergyman does not tell. That is to say, she will have two *possible worlds* to take into account. The Arabic world on the left of Figure 12 and the Roman world on the right. We therefore ought to be looking at Figure 11 and calculating prob({3,III}|{3,5,III}). This will be 2/3 if Alice attaches equal prior probabilities to both the Arabic and the Roman worlds.

The point is that "ignoring ignorance" is not properly consistent with organizing uncertainties probabilistically. If Alice attaches a probability of 1/2 in state 5 to the event E that her face is dirty given that Bertha is dirty and Cora is clean, then she is behaving as though she knew that the roman world were impossible. But this conclusion is not one which anyone would wish to claim was a necessary consequence of her informational status. If she is not to explore the possibility that the world might be Roman, the most that one can reasonably assert about her conditional probability for the event E is that it lies between 1/2 and 1. That is to say, "ignoring ignorance" means that probabilistic judgements may be incomplete.

As Holt [1988] observes, the interdependence of assumptions regarding probability orderings, informational structures and characteristics of knowledge is not as well–known in the economics literature as in the philosophical literature (e.g. Hughes/Cresswell [1985] and Gädenfors [1975]), but it is a subject of some importance in this context.

9. Common knowledge of the game. It is clear why it should be assumed that players *know* the structure of the game they are playing. But why should it be required that the game structure is *common* knowledge?

To make this point, a version of an example of Rubinstein [1988] will be described. The game is one of "pure coordination" of the type considered by Lewis [1969] when introducing the notion of common knowledge. In such games, the players' interests coincide and their aim is to act as a "team" in getting to the best possible outcome. Their problem is that their freedom to communicate is restricted and it is therefore not necessarily easy for them to coordinate on a joint plan of action.

The payoff matrix for the game to be considered is given in figure 13. At the beginning, the players do not know how the letters A and B are used to label the strategies. This is decided by tossing a weighted coin which comes down heads with probability 2/3 and tails with probability 1/3. If it were common knowledge that a head had appeared, then both players would, of course, choose strategy A. If it were common knowledge that a tail had appeared, both would choose strategy B. However, only player I observes the fall of the coin.

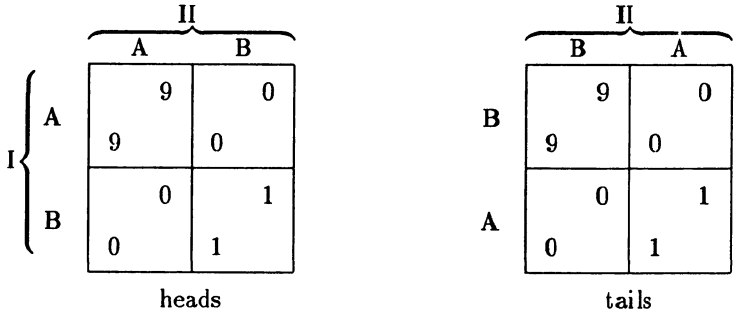


Figure 13

Player I would like to inform player II of his information but can only communicate through the electronic mail system as follows. If he sees a head he sends no message at all. If he sees a tail, he sends a message. Those familiar with the electronic mail system will know that messages fail to reach their destinations with some small probability $\varepsilon > 0$. If player II gets the message, she automatically sends an acknowledgment. If player I gets the acknowledgment, he automatically acknowledges the acknowledgment; and so on.

To study this informational set-up, the state space $\Omega = \{0, 1, 2, 3, \ldots,\}$ is introduced. A state ω in this space represents the number of messages sent. The event H that heads occurs is $H = \{0\}$. The event that tails occurs is $T = \{1, 2, 3, \ldots,\}$. Observe that

$$\omega \in (\text{everybody knows})^{\omega-1} T \qquad (\omega = 1, 2, \ldots,)$$

For example, if $\omega = 3$, then tails occurred and player I knows it because he gets to see the coin. Moreover, player II knows and knows that I knows because she got a message saying so. Further, player I knows that II knows because he got an acknowledgment.

The question is whether players I and II can ever be sufficiently well-informed to make it possible for them to choose strategy B when tails occurs. The answer is that they are *never* able to do so, no matter how large ω may be, if it is given that player I will always do the natural thing in those cases when heads occurs and choose A. If tails occurs, it may be that everybody knows that tails has occurred and everybody knows that everybody knows for 117 iterations. Nevertheless, they will still be unable to coordinate on strategy B.

To demonstrate this, some more apparatus is required. The players' information sets are shown in figure 14. For example, player II encloses 0 and 1 in the information set i, because either state is consistent with her receiving no message. It is necessary to say what each player would do at each of their information sets. It has already been noted that player I is to choose strategy A at i_0. To say what will happen elsewhere, probabilities need to be calculated.

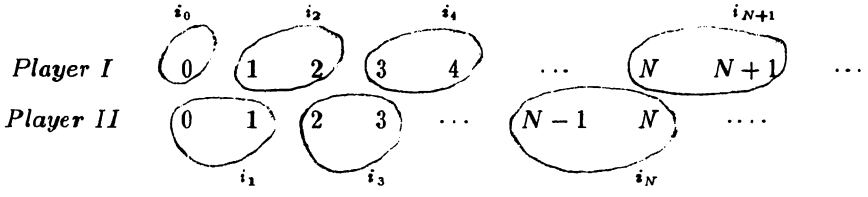


Figure 14

Let $p(\omega)$ be the probability of state ω. Then $p(0) = 2/3, p(1) = \epsilon/3$ and $p(\omega + 1) = (1 - \epsilon)p(\omega)$ $(\omega = 1, 2, 3, \ldots)$. If $i = \{\omega, \omega + 1\}$, then the conditional probabilities for ω and $\omega + 1$, given that information set i has been reached, satisfy $p(\omega + 1 \mid i)/p(\omega \mid i) = p(\omega+1)/p(\omega)$. Thus

$$p(1 \mid i_1)/p(0 \mid i_1) = \epsilon/2$$

$$p(n+1 \mid i_{n+1})/p(n \mid i_{n+1}) = 1 - \epsilon \qquad (n = 1, 2, 3, \ldots).$$

Since player I chooses A at i_0, the least that player II can get from choosing A at i_1 is $x = 9p(0 \mid i_1) + 0p(1 \mid i_1)$. The most that can be got from choosing B at i_1 is $y = 0p(0 \mid i_1) + 9p(1 \mid i_1)$. Since $y/x = \epsilon/2 < 1$, player II selects A at i_1.

The rest of the argument proceeds by induction. The induction hypothesis is that strategy A is chosen at information set i_n for a given $n \geq 1$. It is to be shown that A is also chosen at i_{n+1}. The

least that can be got by choosing A is $x = 9p(\,n \mid i_{n+1}) + 0p(\,n+1 \mid i_{n+1})$. The most that can be got by choosing B is $y = 0p(\,n \mid i_{n+1}) + 9p(\,n+1 \mid i_{n+1})$. Since $y/x = (1-\epsilon) < 1$, A is selected at i_{n+1}.

It is easy to see that $M(\omega) = \Omega$ in every state ω, and hence the outcome of the coin toss never becomes common knowledge and therefore neither does the structure of the game. This is the explanation proposed to account for the players' failure to coordinate successfully.

However, more than one loose end has been left. The first of these is taken up to some extent in the coming sections. Why should it be supposed that there will be successful coordination even if the game structure *were* common knowledge? For player I to know that a strategy is optimal, he needs to know something about what player II is going to do. To know what player II is going to do, he certainly needs to know something about what player II knows—which includes what player II knows about what player I knows, and so on. But he needs to know *more*. He needs to know how player II *uses* her knowledge. Game theorists typically seek to plug this gap with the assertion that it is "common knowledge that the players are rational." However, what this statement means in precise terms remains a vexed issue.

The second loose end concerns the question of *approximating* common knowledge. When tails occurs, the probability that $\omega \in$ (everybody knows) $^N T$ in Rubinstein's example is nearly 1 for each value of N when ϵ is sufficiently small. Does this not make T "nearly" common knowledge? If so, why is the players' behavior so far from what it would be if T were fully common knowledge?

More or less the same answers have been simultaneously proposed to these questions by Monderer/Samet [1988] and Stinchcombe [1988]. They defuse the second question by denying its premise. They suggest that only *approximately* optimizing play should be expected of the players. They can then play A when they perceive the probability of heads to be high and B when they perceive the probability of tails to be high. If ϵ is small, A will be chosen at i_0 and i_1 and B will be chosen elsewhere. This calls for suboptimal play only at i_1. But i_1 occurs only with the very small probability[6] of $\epsilon(1-\epsilon)/3$.

As to the first question, they supplement the knowledge operator K with a p-belief operator B_p defined so that $B_p E$ may be interpreted as the set of states in which E is believed with probability at least p. One can then define common p-belief in much the same way as common knowledge and use the former as an approximation to the latter when p is close to one. If there is approximate common knowledge in this sense, then there are approximate equilibria close to the equilibria that prevail when there is full common knowledge.

10. Rationalizability. As noted in the preceding section, even when the structure of a game has been assumed to be common knowledge, there remains a yawning gap between the results of game theory and the tenets of Bayesian decision theory on which it is supposedly based. This gap has been traditionally plugged with *ad hoc* additions to Bayesian rationality concerning the use of equilibria in multi-person situations. These *ad hoc* additions are then defended with vague assertions about there being "common knowledge of rationality."

Kadane and Larkey [1982] are particularly outspoken in their criticism of this lacuna in the foundations of the subject. They argue that a Bayesian decision-maker maximizes utility given his subjective probability distribution and, if the resulting actions are not in equilibrium: so what? This view depends on the naive notion that Bayesian rationality somehow confers upon its adherents the capacity to "pluck their beliefs from the air." But, what one believes must depend on what one knows. And, in game theory, the players know things about the other players.

Only recently have formal attempts been made to model the knowledge that a player has about the other players in a game *explicitly* (Aumann [1987], Bernheim [1984, 1985], Brandenburger and Dekel [1987b], Pearce [1984], Reny [1985], Tan and Werlang [1984]). These attempts all take account of the fact that a player's beliefs about the strategic choices of the other players in a game should not be entirely arbitrary.

Bernheim and Pearce's concept of "rationalizability" is the most conservative of the attempts. The only restriction imposed on beliefs is that everybody knows that everybody maximizes utility given their subjective probability distributions; and everybody knows that everybody knows; and so on. That is, it is common knowledge that the players are Bayesian rational. But the actual subjective probability distributions which the players hold are not assumed to be common knowledge.[7] In spite of the weakness of the assumptions, rationalizability can sometimes lead to a clear-cut prediction about the outcome of the game.

Consider the game illustrated in Figure 15. Suppose that player II attaches a subjective probability of p to the event that player I will choose "top" and $1 - p$ to the event that he will choose "bottom" Whatever the value of p, "right" cannot be a maximizing choice for player II because "left" is strictly better when $p > 1/3$ and "middle" is strictly better when $p < 3/4$. Since player I knows that player II is an optimizer, he therefore must assign probability 0 to the event that player II chooses "right." But then, whatever probabilities player I attaches to the other alternatives for player II, it is optimal for player I to choose "top." But player II knows that player I knows that player II is an optimizer. Thus

player II knows that player I will necessarily choose "top." Hence player II chooses "left."

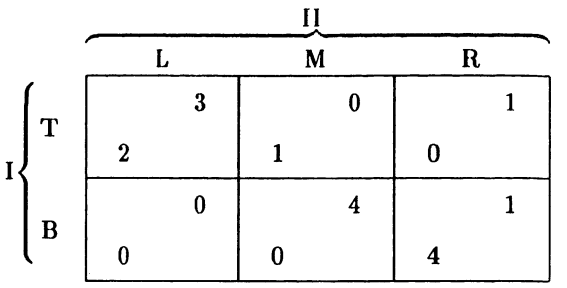


Figure 15

In two-player games, Bernheim and Pearce show that "rationalizable" strategies are those left after strictly dominated strategies are deleted from the game; and then strictly dominated strategies are deleted from the game that results; and so on. (The strategy "right" for player II in Figure 15 is strictly dominated by an equal mixture of "left" and "middle.") This process of successively deleting dominated strategies goes all the way back to Luce and Raiffa [1957]. However, it is well known that it is only in rather special circumstances that the process generates a unique prediction. For the game of "chicken," illustrated in figure 16A, the process has no bite at all. Every strategy is "rationalizable."

Bernheim [1984] regards this as a serious blow for the traditional equilibrium ideas of game theory. However, our feeling is that more is implicitly assumed by traditional game theory than Bernheim is willing to grant. One should therefore not be too surprised if, having thrown out the baby, one is left only with the bathwater.

11. Correlated equilibrium. In contrast to Bernheim, Aumann [1987] is willing to make quite strong common knowledge assumptions about players' beliefs in order to defend the notion of a correlated equilibrium.

Aumann's favorite example of a correlated equilibrium involves the game of "chicken" as illustrated in figure 16A. Suppose that a random device selects one of the cells in figure 16A according to the probabilities indicated in figure 16B. These probabilities are common knowledge, but player I is told only the row of the cell actually selected while player II is told only the column. Then player I and player II will have different but *correlated* information. Suppose that player I now uses the row reported to him as his strategy in "chicken" and player II does the same with the column reported to her. Then both players will be optimizing given the behavior of the other. For example, if player I is told "top," then he expects $6/2 + 2/2 = 4$ from playing "top" but only $7/2 + 0/2 = 3.5$ from playing "bottom."

If he is told "bottom," then he expects 7 from playing "bottom," but only 6 from playing "top." The idea is, of course, very similar to that popularized by Cass and Shell [1983] under the name of "sunspot equilibrium." Aumann argues that "Bayesian rationality in games," if properly interpreted, is nothing other than the play of correlated equilibrium strategies. His argument makes this almost into a tautology. But to explain his point requires a little apparatus.

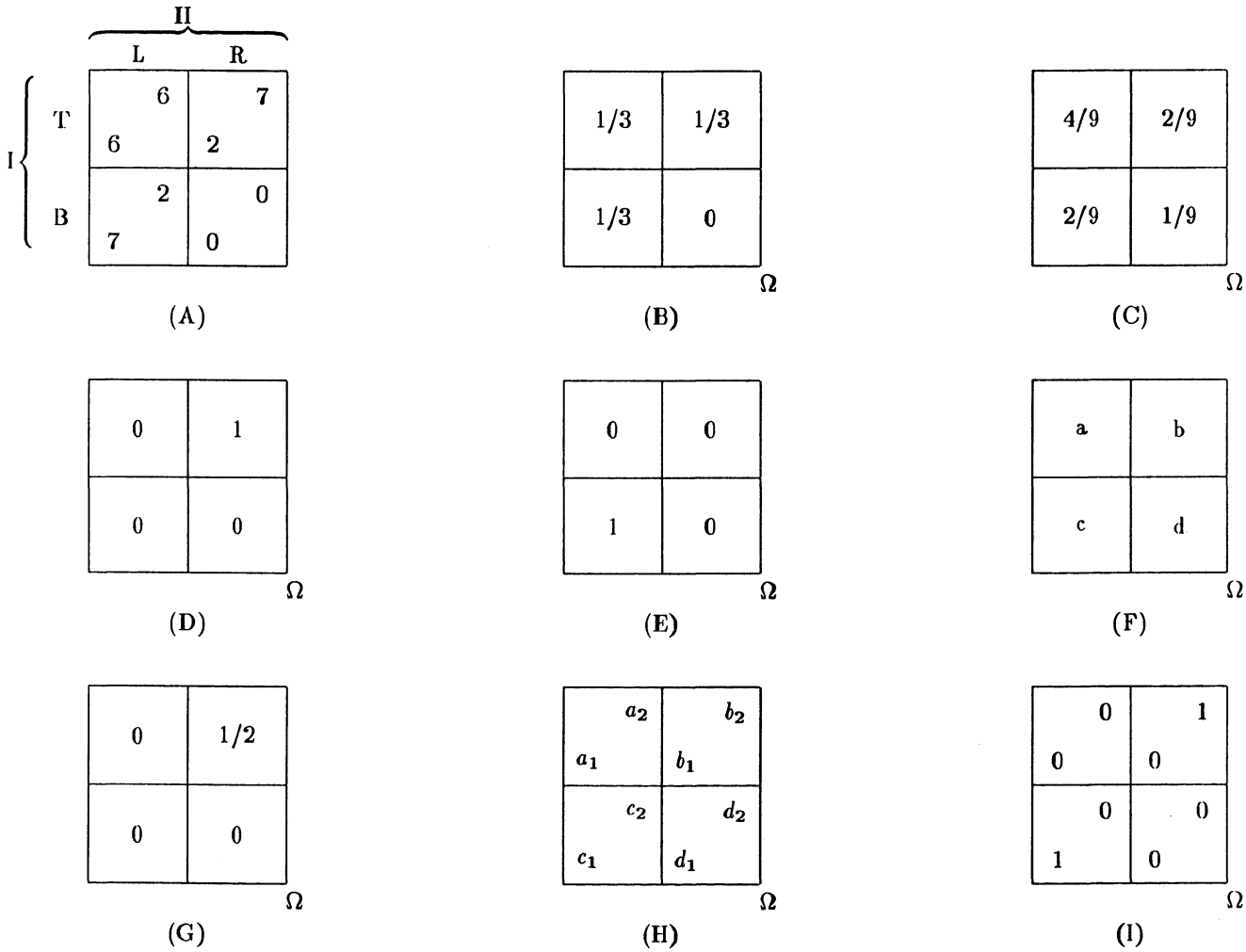


Figure 16

The specification of a two-player game in normal-form consists of the players' strategy sets, S_1 and S_2, and their Von Neumann and Morgenstern utility functions $u_1 : S_1 \times S_2 \longrightarrow R$ and $u_2 : S_1 \times S_2 \longrightarrow R$. Thus $u_1(s_1, s_2)$ is the utility that player I gets if he chooses strategy $s_1 \in S_1$ and

player II chooses $s_2 \in S_2$. In the game of "chicken," $S_1 = \{\text{top,bottom}\}, S_2 = \{\text{left,right}\}$ and, for example, $u_2(\text{top,right}) = 7$.

What does Bayesian rationality require in such a situation? Suppose that, when $\omega \in \Omega$ occurs, players I and II observe $P_1(\omega)$ and $P_2(\omega)$ respectively, and that Bayesian rationality leads then to select[8] strategies $b_1(\omega) \in S_1$ and $b_2(\omega) \in S_2$. If each player's behavior is to be optimal in each state ω given the players' beliefs, then it must be the case that

$$\begin{aligned} \mathcal{E}u_1(\, b_1, b_2 \mid P_1(\omega)) &\geq \mathcal{E}u_1(\, d_1, b_2 \mid P_1(\omega)), \\ \mathcal{E}u_2(\, b_1, b_2 \mid P_2(\omega)) &\geq \mathcal{E}u_2(\, b_1, d_2 \mid P_2(\omega)). \end{aligned} \tag{11.1}$$

where d_1 and d_2 represent any alternative strategy to that prescribed by Bayesian rationality.[9]

The essential point to be gleaned from this discussion is simply that requirement (11.1) *is* an equilibrium condition. Aumann's [1987] gloss is to observe that (11.1) implies the condition for a correlated equilibrium provided that, in each state $\omega \in \Omega$, the players *know* the Bayesian rational strategy. For example, in the game of "chicken," let TOP denote the set of those $\omega \in \Omega$ for which Bayesian rationality requires player I to choose "top". Then to know his Bayesian rational strategy, player I will need to know when the event TOP occurs and, similarly, when the event BOTTOM occurs. As far as "chicken" is concerned, he can discard any other information he might have and just retain the information partition {TOP, BOTTOM} of Ω. Similarly, player II need only retain the partition {LEFT, RIGHT}. But then we have the picture of figure 16B and hence Bayesian rationality has generated for us the correlated equilibrium discussed earlier, provided that the players' prior beliefs (i.e. the beliefs they have *before* observing anything) satisfy

$$\text{prob (TOP} \cap \text{LEFT)} = \text{prob (TOP} \cap \text{RIGHT)} = \text{prob (BOTTOM} \cap \text{LEFT)} = 1/3.$$

This is the first explicit mention of the players' prior beliefs on Ω. What do the players know about these? There is no difficulty in supposing that players know their *own* prior beliefs on Ω. They are therefore in a position to check that the behavior required by Bayesian rationality does indeed maximize their expected utility, provided the other players do not deviate. But to check that the others will not deviate, a player needs to know something about the prior beliefs of the other players. Otherwise, it would not be possible to check that the behavior required by Bayesian rationality maximizes *their* utility also. Aumann [1987] argues that we actually have no choice but to assume that all the prior subjective probability distributions are, in fact, ***common knowledge***. He reasons that, if player j were to learn that $\omega \in \Omega$ had occurred but continued to entertain two possibilities for the subjective probability attached by player i to the state ω, then ω would not be an all-inclusive description of the state of the

world. An *all-inclusive* description would include a description of the probability that player j attaches to that description. Thus, everybody must know everybody's *prior* beliefs and, by a similar argument, everybody must know that everybody knows, and so on. We share the unease that the reader will probably feel about the self-reference built into this argument, for the reasons already discussed in Section 6. For us, the assumption of common knowledge of prior probability distributions therefore remains something which needs defense.

We have seen how common knowledge assumptions provide the glue which holds together Aumann's defense of correlated equilibrium in Bayesian context. Some taxonomy is now appropriate. This depends on the answers to the following questions:

1 Are the players *independent*?
2 Do the players have *common* priors?

To say that the players are independent means that all of the players believe that their own information is statistically independent of that received by the other players. Thus, for example, in "chicken," both players' priors will satisfy $\text{prob}(TOP \cap LEFT) = \text{prob}(TOP) \times \text{prob}(LEFT)$. Think of the description of $\omega \in \Omega$ as including the fall of a weighted coin observed only by player I and the fall of an independent weighted coin observed only by player II.

To say that the players have common priors is to say that their prior beliefs on Ω are the same. This is *not*, of course, implied by the hypothesis that their priors are common knowledge. The players may "agree to disagree" about what the prior beliefs should be.

We begin with the case when it is simultaneously true that the players choose independently and that they have common priors. These are the traditional assumptions of noncooperative game theory and the conclusion is the traditional conclusion: namely, that the result will be a *Nash equilibrium*. For example, if the players' prior beliefs are as indicated in figure 16C, then the result is a Nash equilibrium for "chicken." An observer will see player I choose "top" with probability 2/3 and "bottom" with probability 1/3, while player II independently chooses "left" with probability 2/3 and "right" with probability 1/3. Each of these "mixed strategies" is an optimal reply to the other. Given player II's behavior, player I will get $6 \times 2/3 + 2 \times 1/3 = 14/3$ from the choice of "top" and $7 \times 2/3 + 0 \times 1/3 = 14/3$ from choice of "bottom." He is therefore indifferent between these two alternatives and therefore content to use any mixture of them. Similarly for player II.

"Chicken" has two other Nash equilibria. These involve only "pure strategies." In the first, player I always chooses "top" and player II always chooses "right." (The players' common prior is then

as in figure 16D.) For the second Nash equilibrium in pure strategies, player I always chooses "bottom" and player II always chooses "left." (The players' common prior is then as in figure 16E.)

Note that for Nash equilibrium, in contrast to the general case of a correlated equilibrium, the players' posterior beliefs about each other's *choice of strategy* are common knowledge. If we had not begun with a discussion of correlated equilibrium, this might have been a more natural characterization of Nash equilibrium. (See Brandenburger and Dekel [1987b] and Aumann [1988] for a different view.) In the correlated equilibrium supported by the prior of figure 16B, for example, player I's posterior belief about player II's choice of strategy depends on whether I observes TOP of BOTTOM, but II does not know which of these I observes.

Before leaving the subject of Nash equilibrium, there is a point to be made about the relevance of *mixed* equilibria. Economists sometimes reject these out of hand on the grounds that economic agents simply do do randomize when making decisions. But such a view depends on adopting a rather naive interpretation of what a mixed Nash equilibrium is. One advantage of working strictly in terms of an underlying state space Ω is that a more sophisticated interpretation lies immediately on the surface. In the mixed Nash equilibrium for "chicken" described above, neither player actively randomizes. Both simply see themselves as choosing deterministically, given their information. The random element arises from uncertainty *in the mind of the other player* about what that choice will be. It may well be that real-life economic agents do not consciously randomize when making decisions, but they will nearly always be conscious of being uncertain about what other agents will decide. No a priori case therefore exists for the wholesale rejection of mixed equilibria in the economic context.

The case when the players' choices may be correlated, but prior beliefs are common, is that of correlated equilibria. In "chicken," the correlated equilibrium described earlier is only one of many. In general, the prior subjective probability distributions which support a correlated equilibrium are described by a simple system of linear inequalities. Referring to figure 16F, it is easy to check that the appropriate inequalities for "chicken" are:

$$\left.\begin{aligned} 6a + 2b &\geq 7a \\ 7c \qquad &\geq 6c + 2d \\ 6a + 2c &\geq 7a \\ 7b \qquad &\geq 6b + 2d. \end{aligned}\right\}$$

Nash equilibria are special cases of correlated equilibria and it is not difficult to see that any convex combination of Nash equilibrium outcomes is achievable as a correlated equilibrium. (Just let the players jointly observe a random device which selects a Nash equilibrium and then require them to take

whatever independent actions are necessary to implement that Nash equilibrium.) For example, the common prior of figure 16G gives the two pure Nash equilibrium outcomes for "chicken," each with probability 1/2. But note that the resulting payoff of $7/2 + 2/2 = 4.5$ for each player is not so good as the expected payoff of $6/3 + 2/3 + 7/3 = 5$ that each player gets with the correlated equilibrium supported by the common prior of figure 16B. The expected outcome with this latter correlated equilibrium is *not* obtainable as a convex combination of Nash equilibrium outcomes. For a prettier example, see Moulin and Vial [1978].

For the case when priors are not common, consider figure 16H. The numbers a_1, b_1, c_1, and d_1 are subjective probabilities for player I and the numbers a_2, b_2, c_2, and d_2 are subjective probabilities for player II. All these numbers are common knowledge but the players "agree to disagree." Figure 16I is a special case chosen to highlight the fact that something counter-intuitive is perhaps involved in such situations. With priors as in figure 16I, both players believe it is certain that they will come away from the game with their maximum payoff of 7.

It is pleasant to be able to round off this section by pointing out that Bernheim and Pearce's rationalizability is not so distant from Aumann's correlated equilibrium after all. Brandenburger and Dekel [1978b] have shown that rationalizability can be recast in an equilibrium mold. in the two-player case, a pair of payoffs is rationalizable if and only if it is the vector of payoffs from what Aumann calls a "subjective correlated equilibrium." By this is meant an equilibrium of the type we have just been discussing: namely, one in which the players may agree to disagree and use commonly known but different priors. This result emphasizes the importance of the assumption of common priors in determining which solution concept is to be regarded as the "correct" expression of Bayesian rationality.

Finally, what of independence when the priors are not common? On this point, we want only to mention that, in discussing "rationalizability," we have restricted attention always to the case $n = 2$, because Bernheim and Pearce make independence assumptions which become relevant when $n \geq 3$ but which it would be pedantic to describe here.

12. Updating beliefs. Section 4 describes how the acquisition of new information may lead decision-makers to update their information partitions. Such refining of their information partitions will be accompanied by an updating of their prior beliefs. The study of this phenomenon is clearly very important if one hopes to get some sort of handle on markets in which speculation is an important element.

The current discussion begins with Aumann's "agreeing to disagree" result (which is proposition 6 of section 8). The agents are endowed with common knowledge of a common prior. Each then receives

some private information. Their posterior probabilities for an event F then become common knowledge. Aumann's result is that these posterior probabilities must necessarily be equal.

An example of Geanakoplos and Polemarchakis [1982] serves to show that this result leaves various stones unturned. In this example, Alice and Bertha have a common prior that attaches equal probabilities to each element of $\Omega = \{1,2,3,\ldots,9\}$. Their information partitions are $\{A_1, A_2\}$ and $\{B_1, B_2, B_3\}$ respectively as illustrated in figure 17. For example, $P_A(4) = A_2$ and $P_B(9) = B_3$. The meet M of P_A and P_B (section 4) satisfies $M(\omega) = \Omega$ for all $\omega \in \Omega$. It follows from proposition 5 that the only event that can be common knowledge when it occurs is Ω. Thus, if

$$E = \{\omega : \text{prob}(F|P_A(\omega) = q_A)\}$$

is common knowledge as required in the hypotheses of Aumann's result, then $E = \Omega$. This means that Alice *always* attaches probability q_A to the event F whatever her private information. Similarly Bertha always attaches probability q_B to F. Their private information is therefore irrelevant to the event F and hence Aumann's result is empty for this example.

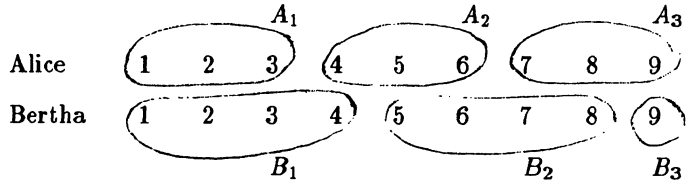


Figure 17

Consider, however, the case when $F = \{3,4\}$ and the following sequence of truthful interchanges between Alice and Bertha takes place. Suppose that $\omega = 2$. Then, Alice observes $P_A(2) = A_1 = \{1,2,3\}$. She then communicates her posterior probability of $1/3$ that the event F has taken place to Bertha. Bertha observes $P_B(2) = B_1 = \{1,2,3,4\}$ and so, before receiving Alice's message, has a posterior probability of $1/2$ for F. On getting Alice's message, she deduces that Alice has observed A_1 or A_2 but cannot deduce anything else. Since she knew this already, she leaves her posterior probability for F alone. The next step is for her to report this to Alice. Alice deduces that Bertha must have observed B_1. Since she knew this already she leaves her posterior for F alone and reports this to Bertha. But *this* tells Bertha a great deal. She knows Alice observed A_1 or A_2. If Alice had observed $A_2 = \{4,5,6\}$, then the information that Bertha observed $B_1 = \{1,2,3,4\}$ would have conveyed the fact to Alice that $\omega = 4$ and so Alice would have reported a posterior probability of 1 for F. Since she did not, Bertha deduces that Alice observed $A_1 = \{1,2,3\}$ and hence now announces a posterior probability

of 1/3 for F. After this, all future announcements of the posterior probability of F are 1/3. It has become common knowledge that the event $A_1 = \{1, 2, 3\}$ has occurred and it is common knowledge that $\text{prob}(F|A_1) = 1/3$.

Alice and Bertha therefore reach a *consensus* in respect of their beliefs about F. That is to say, the probabilities they attach to F finally become common knowledge and hence are equal (by proposition 6). But more is true in this particular case. They also end up with the *same* information; namely that A_1 has occurred. This need not always happen. Geanakoplos and Polemarchakis provide a second example in which the agents' posterior beliefs about an event are common knowledge, but *not* their full information about the event. The commonly known consensus beliefs are then *not* what they would be if the two agents pooled *all* their information.

The example is illustrated in figure 18. The state space is $\Omega = \{1, 2, 3, 4\}$ and the common prior is that all states are equally likely. When $\omega = 1$, the consensus probability for the event $F = \{1, 4\}$ is 1/2, but pooling the total information would reveal that F is certain.

Figure 18

While there are situations in which it makes practical sense to imagine agents directly and truthfully reporting their posterior probabilities for some event in the manner examined above, a more likely scenario is that only some statistic of individual beliefs becomes a public event. An example is when asymmetrically informed agents come to a market to trade. The trading process causes private information to be aggregated into a public statistic, such as a price or quantity. If agents recompute their beliefs on the basis of the value of this public statistic, and then further prices or quantities are announced, will the agents eventually reach a situation of common knowledge and hence consensus in their beliefs? This question has an obvious relevance to rational expectations equilibria but arises in numerous other contexts also.

There is a considerable literature in Statistics on the reconciliation of differing expert opinions (e.g. Dalkey [1969]). The so-called "Delphi technique" is a commonly discussed example. The experts are envisaged as offering predictions of the likelihood of some event based on their private information.

The average of their predictions is announced publicly. The experts then use this information to revise their predictions. This leads to the announcement of a new average and so on. Parimutuel betting has similar characteristics.

McKelvey and Page [1986] have demonstrated how common knowledge can arise through the publication of such aggregate statistics. (A number of people have noticed that simple proofs of the result are possible, e.g. Brandenburger and Geanakoplos [1986].) We shall look only at the "Delphic" special case of their result. Each of n agents computes his or her posterior probability $q_i(\omega)$ for an event F. Thus, originally, $q_i(\omega) = \text{prob}(F|p^i(\omega))$. then the statistic

$$\phi(\omega) = \frac{1}{n}\sum_{i=1}^{n} q_i(\omega)$$

is published. Each agent revises his or her posterior $q_i(\omega)$, using the value of the published statistic. These posteriors are then averaged and the new statistic is published; and so on. The result is that, after a finite number of iterations of this process, the agents' posteriors must become common knowledge and so consensus is achieved.

An example (taken from McKelvey and Page [1986]) may be of assistance. Suppose that Alice, Bertha and Cora are experts who face a state space $\Omega = \{1,2,3,4,5\}$. The common prior is that each state is equally likely. The initial private information is provided by the information partitions illustrated in figure 19. The experts are interested in the probability of the event $F = \{1,2,3\}$.

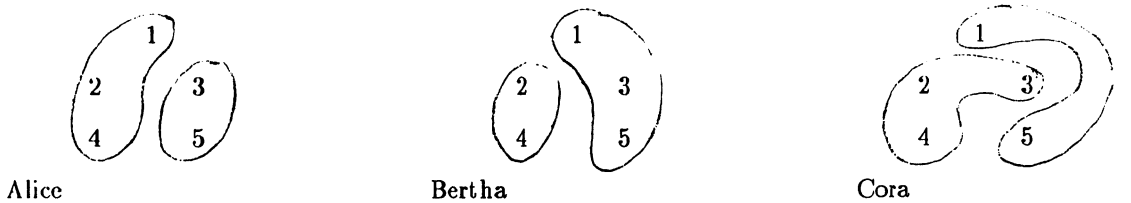


Figure 19

Suppose that the true state is $\omega = 1$ so that Alice, Bertha and Cora calculate their posterior probabilities for F to be $2/3, 2/3$ and $1/2$ respectively. The average of these assessments, $(2/3 + 2/3 + 1/2)/3 = 11/18$ is then published. This gives Alice no new information but informs Bertha and Cora that $\{5\}$ did not occur. Hence the experts' revised posteriors are $2/3, 1$ and 1. Thus a statistic of $(2/3 + 1 + 1)/3 = 8/9$ is published. At this point, Alice can conclude that $\{2,4\}$ did not occur since this would have resulted in a second-round announcement of $(2/3 + 2/3 + 2/3)/3 = 11/18$. She therefore revises her posterior probability of F to 1 and consensus is achieved.

Such simple models are reminiscent of Keynes' "beauty contest" story. Keynes drew attention to newspaper competitions of the time in which the aim was, not to identify the most beautiful young lady whose photograph was supplied, but to predict the way the voting would go on this matter. The aim, in both cases, is to draw attention to certain aspects of the problem of speculation in financial assets. What models like that of McKelvey and Page [1986] add to the Keynesian picture is a formal identification of the basic difficulty. The trading activities of rational agents with common knowledge of a common prior, who all receive different pieces of private information, must be expected to lead to an eventual consensus in beliefs, after which no rational basis for speculation will exist. One might rationalize continued speculation by supposing that the agents are in continuous receipt of new items of private information about an environment assumed to be constantly changing (although we do not know of work in which this is attempted within a formal common knowledge framework). But, is this the way things really are? Or is it simply that speculators are actually "agreeing to disagree"? (Varian [1987] contains further discussion of these issues.)

13. Are Common Knowledge Assumptions Realistic? In brief, our view is that it is precisely in those applications which lean most heavily on common knowledge assumptions, whether this is explicitly recognized or not, that the assumptions are least realistic. We have particularly in mind those models in which a great deal can happen as time passes and a number of agents, each of significant size, seek to learn what has happened and to adapt their behavior to the circumstances. One might summarize what we have to say with the observation that it seems to us painfully naive to suppose that Bayesian updating captures more than a tiny part of what is involved in genuine learning. Whatever the answer to the problem of scientific induction may be, it will surely involve more than the trivial algebraic manipulation called Bayes' rule.

Savage's [1954] theory, in which he synthesized von Neumann and Morgenstern's expected utility theory with the subjective probability ideas of Ramsey, de Finetti and others, is entirely and exclusively a *descriptive* theory of consistent behavior. It has nothing to say about how decision-makers acquired the beliefs ascribed to them: it asserts only that, if the decisions taken are consistent, then the decision-makers act *as though* they were maximizers of expected utility relative to a subjective probability distribution. (Objections to the definition of "consistent" in the theory are possible but this is not our point. In particular, the objection that "real people" are sometimes inconsistent is irrelevant to a theory of rational behavior. People often get their sums wrong. But nobody argues that we should therefore change the rules of arithmetic.)

The "Bayesian blunder" is to suppose that Savage's passive descriptive theory can be re-interpreted as an active, prescriptive theory at negligible cost. Obviously, a sensible decision-maker will be unhappy about inconsistencies. A naive Bayesian therefore assumes that it is enough to assign prior beliefs to a decision-maker and then forget the problem of where beliefs come from This was the attitude we adopted in the preceding sections in sketching the nature of the orthodox approach to common knowledge. Consistency then forces any new information that may transpire to be incorporated into the system by Bayesian updating—i.e. a posterior belief is deduced from the prior belief using Bayes' rule. The naiveté does *not* consist in using Bayes' rule, whose validity as a piece of algebra is not in question. It lies in supposing that the problem of where the priors come from can be quietly shelved. Some authors even explicitly assert that rationality somehow *endows* decision-makers with priors and hence the problem does not exist at all.

Savage [1954] had a considerably more complex view. He did argue that his descriptive theory could be of practical assistance in helping decision-makers from their beliefs. His point was that rational agents would not rest if they found inconsistencies in their belief systems. Luce and Raiffa [1957, p.302] expound Savage's view as follows:

> Once confronted with such inconsistencies, one should, so the argument goes, modify one's initial decisions [about beliefs] so as to be consistent. let us assume that this jockeying, making snap judgments, checking on inconsistency, etc.—leads ultimately to a bona fide, a priori distribution.

For what follows, we need to expand on this quotation. What is at issue is *why* a rational decision-maker should want to be consistent. After all, scientists are not consistent, on the grounds that it is not clever to be consistently wrong. When surprised by data that shows current theories to be in error, they seek new theories which are inconsistent with the old theories. This is what *genuine* learning is like. Consistency, from this point of view, is only a virtue if the possibility of being surprised can be eliminated somehow. How does Savage see this situation being achieved?

A person who makes judgments in a reasonable way will presumably prefer to make judgments when he or she has more information rather than less. A decision-maker might therefore begin to tackle the problem of constructing an appropriate belief system by asking: for every conceivable possible course of future events, what would my beliefs be after experiencing them? Such an approach automatically discounts the impact that new knowledge will have on the basic model being used to determine beliefs—i.e. it eliminates the possibility that the decision-maker will feel the need to alter this basic model after being surprised by a chain of events whose implications had not previously been considered. Next comes the question: is this system of *contingent* beliefs consistent? If not, then the

decision-maker may examine the relative *confidence* that he or she has in the "snap judgments" he or she has made and then adjust the corresponding beliefs until they *are* consistent. With Savage's definition of consistency, this is equivalent to asserting that the adjusted system of contingent beliefs can be deduced, using Bayes' rule, from a single prior. It is therefore true, in this story, that the final "massaged" posteriors can be deduced formally from the final "massaged" prior using Bayes' rule. This is guaranteed by the use of a complex adjustment process which operates until consistency is achieved. As far as the massaged beliefs are concerned, Bayes' rule therefore has the status of a *tautology*, like $2 + 2 = 4$. Together with the massaged prior, it serves essentially as an indexing system which keeps track of the library of massaged posteriors. It would perhaps be wrong to argue that there is no learning going on when an index system is used to locate a book in a library, but it will be clear that, if the above story is to be believed, then the real learning takes place *during the massaging process*. Notice that what happens during the real learning process is that a massaged *prior* is deduced from a set of primitive *posteriors*. To pursue the analogy given above, the actual learning should be thought of as assembling the right books to go in the library, the *final* step being the provision of a suitable index.

Sophisticated Bayesians, among whom we would like to count ourselves, sometimes follow Harsanyi in signaling that some analog of Luce and Raiffa's jockeying procedure is required by asserting the Bayesian theory applies only to "closed universe" or "small world" problems—i.e. to problems in which all potential surprises can be discounted in advance (section 6). Suitable examples are to be found in the small decision trees with which books on Bayesian decision theory for students of business administration are illustrated. Naive Bayesians make no such qualifications. For them, something which is not formalized in Bourbaki-type mathematics does not exist and hence no account need be taken of it at all.

The implications of taking a sophisticated view of Bayesian theory are serious enough when only one decision-maker is involved. In sections 6 and 7, and again in section 11, we considered the orthodox practice of asserting that the states of the world $\omega \in \Omega$ are to be *all-inclusive*. This allows various tricks of the trade to be practised on the unwary. However, whatever else may be clear, a problem in which *all* future histories, without exception, have to be taken into account is not one for which the massaging process described earlier makes any practical sense whatsoever. If Bayesian conclusions are not to be abandoned altogether, it is therefore necessary to be very much more circumspect about *how* inclusive a state of the world is taken to be.

However, although all this is related to what comes next, it is not the main point we wish to make. The main point concerns what Aumann [1976] calls the "Harsanyi doctrine": namely, that priors should

be taken to be common, to which Aumann adds the rider that it should be common knowledge that the priors are common. The orthodox defense need not be spelled out in detail since it is closely related to the idea of the "original position" as popularized by Rawls. Individuals are thought of entering the world with a mind which is a *tabula rasa*. Since rational individuals in this "original position" will all have the same information, how could they adopt different priors?

Obviously, such a defense of the common prior assumption will not survive the interpretation of Savage's theory offered above. However, one can recast the defense as follows. Imagine all of the players constructing their beliefs, before there is any action, using the Luce and Raiffa jockeying process. How could beliefs come to be commonly held under such circumstances? Each player would need to be able to argue with confidence that he would make the *same* subjective judgments as any other player under *all* circumstances provided that he or she were to contemplate precisely the same contingencies under precisely the *same* conditions. (Harsanyi [1977] refers to a related notion as "extended sympathy.") In this case, each player could regard the other players, in their aspect of information processors, simply as surrogates of himself. The massaging process would therefore generate a common prior.

But this new defense asks us to swallow a great deal more than the old defense. Both require us to think of individuals beginning with an identical kit of information-processing tools and no information at all. Differences between individuals then emerge entirely as a result of different information that they receive as time passes. If the kit of tools is conceived of as containing only Bayes' rule and nothing else, then the idea that we can mimic the reaction of other players to their information does not seem fraught with difficulty. But once the massaging process is taken into account, we are faced with a tool-kit of unknown composition. Not only that, people learn about how to learn as they gain experience. A person's tool-kit after the passage of some time will not therefore be the same as his or her original tool-kit. Even if one is prepared to grant the doubtful proposition that we "know" what tools we have in our current kit (i.e. we know the mechanism by means of which we currently make subjective judgments), it does not follow that we can distinguish those tools that we have acquired through experience from those with which we were equipped originally. And probably it does not even make sense to suppose that we can itemize those tools we would have had in our current kit if our experience of the world had been other than it was. But, without this sequence of increasingly unlikely requirements, the new defense is unworkable. This is not to say that the defense may not have value as a useful approximation in simple enough situations, only that it seems unwise to elevate its conclusions into holy writ.

We continue this last point with some remarks on extensive-form games. In Sections 10 and 11, we looked only at normal-form games (in which all of the action can be thought of as taking place at a single instant). But, in extensive-form games, time enters the picture. During the play of such a game, the players may receive evidence that the hypotheses they need to sustain the Harsanyi doctrine, or some other doctrine, are *false*. Figure 20 is the game-tree of an example adapted by Reny [1986] from a similar example of Rosenthal [1981]. (See also Basu [1985].) Reny calls the game "take-it-or-leave-it." A philanthropist arrives with $\$10^n$ unsure of which of two institutions to endow. Alice and Bertha represent the two institutions. The philanthropist makes them play the following game which may proceed through n stages. If the k^{th} stage is reached, the philanthropist will have placed $\$10^k$ on the table. If k is odd, Alice can take the money and so end the game. If k is even, Bertha can take the money and end the game. If the money is not taken and $k < n$, then the philanthropist multiplies the amount available by ten and the game proceeds to the k^{th} stage. If $k = n$, the philanthropist leaves in disgust taking the money with him.

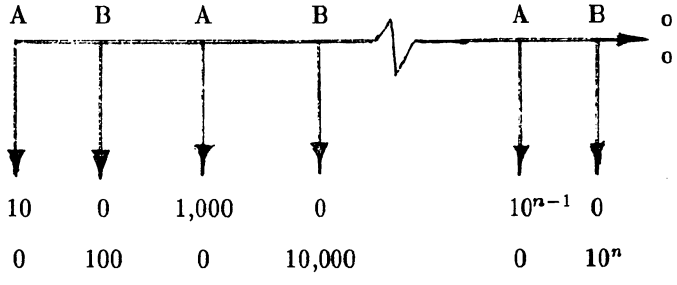


Figure 20

Reny's [1985] point is very simple. He notes that "rationalizability" (and most other game theory "solution concepts") requires Alice to take the money at step 1. But what would Bertha then deduce about Alice if Alice did *not* take the money? The natural answer is that Alice is not a rationalizer. What then is she? The "rationalizing" theory offers no clue. But suppose Bertha decides that, whatever sort of person Alice is, if she left \$10 on the table at step 1, then it is very likely that she will leave \$1,000 on the table at step 3. In this case player II will leave \$100 at step 2, expecting to be able to claim \$10,000 at step 4. However, if Alice anticipates all this, it will not be optimal for her to "rationalize" at step 1 because she will be passing up the chance of being able to claim \$1,000 at step 3.

Such discussions are absent from traditional game theory. In Selten's [1975] concept of perfect equilibrium, any deviation by players from the "Bayesian rational" course of action is attributed to small random errors. The players' hands are said to "tremble" slightly in selecting their actions, so that

there is always a small chance that they will do the wrong thing by "mistake," although they always intend to do the right thing. In Chess, for example, we are supposed to explain a long sequence of bad moves by asserting that the opponent always meant to move the correct piece, but continually picked the wrong one up by mistake. More often, however, a formalism is adopted within which it is difficult, or impossible, even to express the relevant issues so that they then do not "need" to be considered at all. One advantage of being explicit about what is being assumed about the players, as discussed in section 10, is that such evasions become very difficult to sustain.[10]

Progress on these issues must presumably await advances in the theory of bounded rationality (although Aumann [1988] has a proposal for cutting the Gordian knot). Binmore [1987/8] reviews some of the issues.

14. Conclusion. What does all this mean for the study of common knowledge? Do the criticisms and reservations of the preceding section mean that its continued investigation is fruitless? This would not be a good conclusion to draw. Firstly, our criticisms are directed at applications of the theory which hopelessly overload its currently rather tenuous underpinnings. In small-scale applications, however, such as normal-form games with a small number of strategies, these criticisms are not necessarily relevant. Even in the much more doubtful large-scale applications, there is comfort to be drawn from the fact that, without an explicit understanding of the nature of common knowledge, it would not even be possible to properly appreciate where the inadequacies lie in the theories which are current.

Footnotes

1. There are various difficulties with the raw version of Littlewood's story. One might respond to A's question by asking why B is still laughing given that she can analyze the situation just as easily as A.

2. The first criterion implies the second because, if $T = (ek)^{\infty}T$, then

$$(ek)T = (ek)(ek)^{\infty}T = (ek)^{\infty}T = T.$$

The second criterion implies the first because, if $T = (ek)T$, then $(ek)T \subseteq (ek)^2T$ by $(K1)$. But $(ek)^2T \subseteq (ek)T$ by $(K2)$. Thus $T = (ek)T = (ek)^2T$. Similarly, $T = (ek)^nT$ for all n, and so $T = (ek)^{\infty}T$.

3. In symbols

$$P(\omega) = \bigcap_{\omega \in T} T = \bigcap_{\omega \in KE} E$$

where T ranges over all truisms and E over all events.

4. The union of two truisms, S and T, is a truism because $S \cup T = KS \cup KT \subseteq K(S \cup T) \cup K(S \cup T) = K(S \cup T)$.

5. Formally, this means that the event

$$E = \{w : \text{prob}(F|P_A(\omega) = q_A)\}$$

is common knowledge. Thus $M(\omega_0) \subseteq E$ by proposition 5. Suppose $Q_j = P_A(\omega)$. Then $\omega \in P_A(\omega) \subseteq M(\omega_0) \subseteq E$. Hence $\text{prob}(F|Q_j) = \text{prob}(F|P_A(\omega)) = q_A$.

6. Of course, player II's choice of B is not approximately optimal *after* i_1 has occurred. Player II's choice is approximately optimal *before* the realization of ω.

7. A natural criticism is that Savage's [1954] theory constructs von Neumann-Morgenstern utilities and subjective probabilities *simultaneously*. At first sight it therefore seems odd that one should be the subject of common knowledge but not the other. Brandenburger and Dekel [1987b] present an expanded model with private information in which this objection does not bite. See also the discussion at the end of section 11.

8. A significant evasion is concealed in this argument. The assumption that Bayesian rationality leads to a specific recommendation for behavior rather than merely a set of constraints on behavior is left undefended.

9. Shin [1988] suggests that, with a truly all-inclusive description of a state, (11.1) is not appropriate since part of what one observes is the action one actually takes. Hence the $P_i(\omega)$ on the left of the inequalities should not be the same as that on the right.

10. Fudenberg, Kreps, and Levine [1987] and Dekel and Fudenberg [1987] address the issues raised in this paragraph in a more precise context.

References

Aumann, R. [1976]: "Agreeing to Disagree," *The Annals of Statistics*, 4, 1236–1239.

——— [1987]: "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, 55, 1–18.

——— [1988]: "Irrationality in game theory (preliminary notes)", unpublished, Hebrew University of Jerusalem.

Bacharach, M. [1985]: "Some Extensions to a Claim of Aumann in an Axiomatic Model of Knowledge," *Journal of Economic Theory*, 37, 167–190.

——— [1987]: "When Do We Have Information Partitions?," unpublished, Christ Church, Oxford.

Basu, K. [1985]: "Strategic Irrationality in Extensive Games," unpublished, Institute of Advanced Study, Princeton.

Bernheim, D. [1984]: "Rationalizable Strategic Behavior," *Econometrica*, 52, 1007–1028.

——— [1985]: "Axiomatic Characterizations of Rational Choice in Strategic Environments," unpublished, Department of Economics, Stanford University.

Binmore, K. [1984]: "Equilibria in Extensive Games," *Economic Journal*, 95, 51–59.

——— [1987/88]: "Modeling Rational Players I and II," *Economics and Philosophy*, 3 and 4, 179–214 and 9–55.

——— and A. Brandenburger [1988]: "Common Knowledge and Game Theory," ST/ICERD discussion paper 88/167, London School of Economics.

Brandenburger, A. [1986]: "The Role of Common Knowledge Assumption in Game Theory," to appear in F. Hahn (ed.) *The Economics of Information, Games, and Missing Markets.*

——— and E. Dekel [1985]: "Hierarchies of Beliefs and Common Knowledge," Research Paper No. 841, Graduate School of Business, Stanford University.

——— [1987a]: "Common Knowledge with Probability 1," *Journal of Mathematical Economics*, 16, 237–245.

——— [1987b]: "Rationalizability and Correlated Equilibria," *Econometrica*, 55, 1391–1402.

——— and J. Geanakoplos [1986]: "Common Knowledge of Summary Statistics," unpublished, Cowles Foundation, Yale University.

Cass, D. and K. Shell [1983]: "Do Sunspots Matter?" *Journal of Political Economy*, 91, 193–227.

Cave, J. [1983]: "Learning to Agree," *Economic Letters*, 12, 147–152.

Dalkey, N. [1969]: "The Delphi Method: An Experimental Study of Group Opinion," Rand Corporation, Santa Monica, CA.

Dekel, E. and D. Fudenburg [1987]: "Rational Behavior with Payoff Uncertainty," unpublished, Department of Economics, University of California at Berkeley.

Fudenburg, D., D. Kreps, and D. Levine [1987]: "On the Robustness of Equilibrium Refinements," forthcoming in *Journal of Economic Theory.*

Gädenfors, P. [1975]: "Qualitative Probability as an Intentional Logic," *Journal of Philosophical Logic*, 4, 171-185.

Geanakoplos, J. [1988]: "Common Knowledge, Bayesian Learning and Market Speculation with Bounded Rationality," unpublished, Yale University.

——— and H. Polemarchakis [1982]: "We Can't Disagree Forever," *Journal of Economic Theory*, 28, 192–200.

Gilboa, I. [1986]: "Information and Meta-information," Working Paper 3086, Foerder Institute for Economic Research, Tel Aviv University.

Halpern, J. (ed.) [1986]: *Theoretical Aspects of Reasoning about Knowledge.* Los Altos: Morgan Kaufmann Publishers, Inc.

Harsani, J. [1967–68]: "Games with Incomplete Information Played by 'Bayesian' Players," Parts I - III, *Management Science*, 14, 159–182.

——— [1977]: *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge: C.U.P.

Holt, D. [1988]: "Models of Knowledge and Information,"unpublished, University of Michigan.

Hughes, G and Cresswell, M. [1985], *An Introduction to Modal Logic*, Methuen.

Kadane, J. and P. Larkey [1982]: "Subjective Probability and the Theory of Games," *Management Science*, 28, 113–120.

Kaneko, M. [1987]: "Structural Common Knowledge and Factual Common Knowledge," RUEE Working Paper 87–27.

Lewis, D. [1969]: *Conventions: A Philosophical Study*, Cambridge: Harvard University Press.

Littlewood, J.E. [1953]: *Mathematical Miscellany*, ed. B. Bollobas, Cambridge University Press, London, 1986.

Luce, R. and H. Raiffa [1957]: *Games and Decisions.* New York: Wiley.

McKelvey, R. and T. Page [1986]: "Common Knowledge, Consensus, and Aggregate Information," *Econometrica*, 54, 109–127.

Mertens, J.F. and S. Zamir [1985]: "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1–29.

Milgrom, P. [1981]: "An Axiomatic Characterization of Common Knowledge," *Econometrica*, 49, 219–222.

——— and N. Stokey [1982]: "Information, Trade, and Common Knowledge," *Journal of Economic Theory*, 26, 177–27.

Monderer, D. and D. Samet [1988]: "Approximating Common Knowledge with Common Beliefs," typescript, M.E.D.S., Northwestern University.

Moulin, H. and J.P. Vial [1978]: "Strategically Zero-Sum Games: The Class of Games Whose Completely Mixed Equilibria Cannot be Improved Upon," *International Journal of Game Theory*, 7, 201–221.

Nielsen, L. [1984]: "Common Knowledge, Communication, and Convergence of Beliefs," *Mathematical Social Sciences*, 8, 1–14.

Pearce, D. [1984]: "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52, 1029–1050.

Reny, P. [1985]: "Rationality, Common Knowledge, and the Theory of Games," unpublished, Department of Economics, Princeton University.

Rosenthal, R. [1981]: "Games of Perfect Information, Predatory Pricing, and Chain-Store Paradox," *Journal of Economic Theory*, 25, 92–100.

Rubinstein, A. [1988]: "A Game with "Almost Common Knowledge": an Example," forthcoming in *American Economic Review.*

Samet, D. [1987]: "Ignoring Ignorance and Agreeing to Disagree," Discussion Paper No. 749, KGSM, Northwestern University.

Savage, L. [1954]: *The Foundations of Statistics.* New York: Wiley.

Schelling, T. [1960]: *The Strategy of Conflict.* Cambridge: Harvard University Press.

Selten, R. [1975]: "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25–55.

Shin, H. [1986]: "The Structure of Common Knowledge," unpublished, Nuffield College, Oxford.

——— [1987]: "Logical Structure of Common Knowledge, I and II," unpublished, Nuffield College, Oxford.

——— [1988]: "A Comment on Aumann's Definition of Bayes-Rationality," unpublished, Nuffield College, Oxford.

Stinchcombe, M. [1988]: "Approximate Common Knowledge," unpublished, University of California at San Diego.

Tan, T. and S. Werlang [1984]: "The Bayesian Foundations of Rationalizable Strategic Behavior and Nash Equilibrium Behavior," forthcoming in *Journal of Economic Theory.*

——— [1985]: "On Aumann's Notion of Common Knowledge: An Alternative Approach," WP 85–26, University of Chicago.

Varian, H. [1987]: "Differences of Opinion in Financial Markets," unpublished, Department of Economics, University of Michigan.