

Knowledge, Beliefs and Game-Theoretic Solution Concepts*

Oliver Board[†]

`oliver.board@economics.ox.ac.uk`

November 2002

Abstract

In situations of strategic interaction it is important for agents to consider not only what their rivals will do, but also what they know, what they know about what they know, and so on. Formal models of knowledge have been developed to help us keep track of these levels of knowledge. This paper provides a non-technical introduction to one of these models, and investigates its foundations. It is then shown how the model can be used to analyze game-theoretic solution concepts, in particular Nash equilibrium.

Fred, a young faculty member in the Department of Economics, is up for reappointment. Two professors, Gillian and Helen, have been asked to assess his competence at research and teaching respectively. Only if he is found to be good at both will he be granted tenure. The (uninformed) faculty gossip is that he is good at teaching but bad at research. After carrying out her part of the assessment, however, Gillian finds out that he is in fact good at research. Before filing her report, she offers Helen a bet that Fred will make tenure. This looks like a good bet to Helen (assuming even odds): even if her assessment is positive, she believes on the basis of the faculty gossip that Fred is probably bad at research. They are just about to shake on it when Helen has the following thought: “Gillian is about to accept the bet that Fred will make tenure. But he’ll only make tenure if his research is good. Since Gillian has assessed his research, she must know that it is good.” From the fact that Gillian is willing to accept the bet, Helen infers that Gillian knows Fred’s research is good, and the bet may not be such a good deal after all.

But what if Helen’s assessment reveals that Fred is actually bad at teaching? Now she is on to a sure thing if she accepts the bet, since Fred will be granted tenure only if he is good at research

*I am grateful to Meg Gleason and an anonymous referee for helpful comments.

[†]Department of Economics, University of Oxford, Manor Road, Oxford, OX1 3UQ.

and teaching. It is Gillian's turn to reconsider. Seeing the smile on Helen's face as she offers her hand, she reasons as follows: "If Helen knew Fred was good at teaching, she would not accept the bet, since she must know that I know he is good at research or I wouldn't have offered the bet. But Helen is about to accept the bet, so she must know that Fred is bad at teaching". So Gillian infers that Helen knows that Fred is bad at teaching, and the bet is called off.

The lesson to be learned from this simple story is that in situations of strategic interaction, it is important to consider what the agents know not just about the world and about each other's actions, but also what they know about what they know, what they know about what they know about what they know, and so on. This theme is taken up by other papers in this issue: see Myatt *et al.* [15] for details.

It can be hard to keep track of all these levels of knowledge with the kind of informal reasoning used above. In this paper we develop a formal model of knowledge that allows us to be precise about *every* level of knowledge, and we apply the model in a particular setting: the comparison of game-theoretic solution concepts. Section 1 develops the model of knowledge, while section 2 shows how it can be applied to games. Section 3 provides some concluding remarks. Examples have been used to illustrate general points as much as possible. Formal definitions and theorems have been relegated to boxes, and can safely be ignored by those who have no taste for such things. In addition, section 1.3 is technical and can be omitted without affecting the flow of argument.

1 Modelling knowledge and beliefs

Philosophers have long been concerned with formal models of knowledge. The logic of knowledge, or *epistemic logic*, was pioneered by Hintikka [12], who showed how structures based on *possible worlds* could be used to evaluate statements about agents' knowledge. These structures are similar in many respects to the structures discussed below (where states take on the role of possible worlds), developed independently by Aumann [1]. Aumann used his model to show how one could formalize the notion of *common knowledge* first introduced by Lewis [13]. The generalization of Aumann's model discussed in section 1.3 was brought to the attention of economists by Bacharach [4] and Samet [18].

An encyclopedic treatment of epistemic logic is provided by Fagin *et al.* [9]. Useful surveys of the Aumann structure model and its extensions include Geanakoplos [10] and Osborne & Rubinstein [16].

1.1 Information partitions and Aumann structures

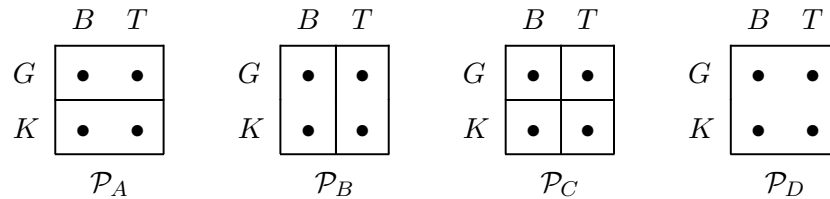
At the heart of Aumann's model of knowledge is the notion of a *state of the world*, or *state* for short. A state provides a complete description of everything of importance to a group of agents in a particular situation. It must specify the physical universe, or at least those aspects of it which may affect our group of agents; it must describe what actions are available to each agent, which actions are actually taken, and the utilities the agents obtain from each action combination; finally, each state must detail what every agent thinks about the world and about every other agent, and what every agent thinks about what every other agent thinks, and so on.

Because agents typically do not know everything about the world in which they live, we must consider a set of states, W , only one of which is the true state. An agent who is less than perfectly informed is unsure which this true state is, and considers several states in W possible on the basis of the information she possesses; her information is represented by this subset of W . A useful analogue is to think of a player's information sets in an extensive form game: if an information set contains more than one node, the player cannot tell which of these nodes has been reached in the same way that an agent who considers several states to be possible cannot tell which of these states is the true state.

More precisely, each agent's knowledge is represented by an *information partition*, which divides the states in W into a number of *cells*. The cells are a collection of mutually disjoint and exhaustive subsets of W , i.e. each state is in precisely one cell. If two states are in the same cell, then the agent cannot tell them apart: whichever happens to be the true state, the agent considers both of them to be possible.

As an example, suppose that Alan, Bruce, Chris and Dave care only about the outcome of the Football World Cup semi finals. Germany are playing Korea in match 1, and Brazil are playing Turkey in match 2. W consists of 4 states, corresponding to the four possible outcomes of these matches. We shall denote these states by GB , GT , KB , and KT , i.e. $W = \{GB, GT, KB, KT\}$. Now suppose that Alan went to Seoul to watch match 1 and Bruce went to Saitama to watch match 2; Chris stayed at home and watched both matches on television, while Dave was on holiday in the USA where there is no football coverage. Their knowledge can be described by information partitions in the following way. Alan knows the outcome of Germany vs. Korea. He can distinguish states GB and GT from states KB and KT , but cannot distinguish between GB and GT or between KB and KT . His information partition contains two cells: $\mathcal{P}_A = \{\{GB, GT\}, \{KB, KT\}\}$.

Similarly, Bruce’s information partition contains two cells, but the cells are different. He cannot distinguish between GB and KB or between GT and KT : $\mathcal{P}_B = \{\{GB, KB\}, \{GT, KT\}\}$. Chris knows the outcome of both matches. He is perfectly informed and can tell the difference between every state: $\mathcal{P}_C = \{\{GB\}, \{GT\}, \{KB\}, \{KT\}\}$. Finally, Dave does not know the outcome of either match. His information partition consists of just one cell containing all four states: $\mathcal{P}_D = \{\{GB, GT, KB, KT\}\}$. A diagrammatic representation of these information partitions is given below.



To see how the partitions work, consider the proposition “Germany wins”. Notice that there is more than one state in which this proposition is true: it is true in state GB and in state GT . We call the set of all these states $G = \{GB, GT\}$ the *event* that Germany wins. G is true¹ at state w just if $w \in G$ (i.e. w is a member of G). When (if ever) do our football fans know that G is true? In general, an agent knows that an event is true if that event is true at every state the agent considers possible. If the true state is GB , then Alan considers GB and GT possible. G is true at both of these states, so Alan does know that G . Bruce, on the other hand, considers GB and KB possible. G is true at the first of these states, but not at the second, so he cannot be sure that G is true: he does not know that G . Similar reasoning tells us that Chris knows that G , while Dave does not.

Let $K_A(G)$ denote the event that Alan knows that Germany wins. $K_A(G)$ is the set of states at which Alan knows that G . We have just shown that $GB \in K_A(G)$. If the true state is GT , Alan again considers GB and GT to be possible, so he still knows that G . At KB and at KT , on the other hand, he considers precisely KB and KT possible; at neither of these states is G true, so he does not know that G . Thus $K_A(G) = \{GB, GT\}$. The reader is invited to check that

$$K_B(G) = \emptyset; \quad K_C(G) = \{GB, GT\}; \quad K_D(G) = \emptyset,$$

¹The observant reader will have noticed the double usage of the word “true”, to refer to propositions (i.e. sentences of the English language) and to refer to events (i.e. sets of states). This slight abuse of terminology should not cause any confusion here.

where \emptyset denotes the empty set. So Alan and Chris knows that Germany wins precisely when Germany really does win; but whatever the true state, Bruce and Dave do not know that Germany wins. This is what we would expect given our original description of the situation.

The model we have just been working with is an example of an *Aumann structure*. Box 1 provides the formal definitions.

Box 1: Aumann structures

An *Aumann structure* for n agents consists of a set of states W and a partition over W for each agent:

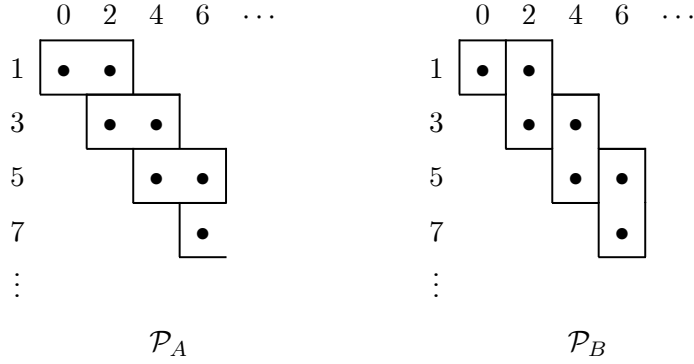
$$\langle W, \mathcal{P}_1, \dots, \mathcal{P}_n \rangle.$$

An agent knows an event $E \subseteq W$ precisely if that event holds at every world the agent considers possible. So letting $\mathcal{P}_i(w)$ denote cell of \mathcal{P}_i containing w , i knows E at w if $\mathcal{P}_i(w) \subseteq E$. The event that i knows E is the set

$$K_i(E) = \{w : \mathcal{P}_i(w) \subseteq E\}.$$

Since $K_A(G)$, $K_B(G)$, etc. are themselves events, we can use the model to analyze when they are known in just the same way as we used it to analyze when G was known, and build up hierarchies of knowledge. $K_D(K_A(G))$, for example, is the event that Dave knows that Alan knows that Germany won (this event is never true: $K_D(K_A(G)) = \emptyset$).

To illustrate this point further, consider a new example (borrowed from Morris [14]). Suppose a father wants to give some money to his two children, Alice and Bob. He wants to share it as evenly as possible, but he has an odd number of dollars so one of the children will end up with one more dollar than the other. Alice ends up with an odd number of dollars, Bob with an even number. The children know all this, but they do not know how exactly how much money their father has. The (infinite) state space and the information partitions are shown below, with rows giving the number of dollars in Alice's envelope and columns the number of dollars in Bob's envelope:



Label the states 10, 12, 32, 34, 54, 56, 76, etc. and consider the event that neither child has more than $\mathcal{L}3$: $E = \{10, 12, 32\}$. Suppose that 10 is the true state. Then Alice considers states 10 and 12 possible. E is true at both of these states, so Alice knows that E . Furthermore, it is easy to check that Bob knows that E at both of these states, so Alice knows that Bob knows that E^2 . But Alice does not know that Bob knows that Alice knows that E : Alice considers state 12 possible, and at this state Bob considers both 12 and 32 possible; Alice does not know that E at the second of these, since at 32 she considers 34 possible as well, where E is false. Thus Bob does not know that Alice knows that E at state 12, and so Alice does not know that Bob knows that Alice knows that E at state 10. It is easier to figure out what is going on if we build up the hierarchy of knowledge events:

$$\begin{aligned}
 E &= \{10, 12, 32\} \\
 K_A E &= \{10, 12\} \\
 K_B K_A E &= \{10\} \\
 K_A K_B K_A E &= \emptyset
 \end{aligned}$$

(we have omitted the parentheses since there is no risk of confusion).

One of the great strengths of the Aumann structure model is that it can be used to analyze hierarchies of beliefs in this way. Economists often assume that certain facts are *common knowledge*, which means that everyone knows them, everyone knows that everyone knows them, and so on. In the example above, event E is never common knowledge among the agents. We use the symbol

²We are here making the implicit assumption that Alice knows Bob's partition; thus she knows, for each state, whether or not Bob knows that E at that state (though she may not know which the actual state is). In fact, it is usual to assume that the partitions are *common knowledge*. Aumann [2] provides a more detailed discussion of this issue.

C to represent common knowledge: if $w \in C(E)$, it is common knowledge that E at state w . A formal definition is given in Box 2.

Box 2: Common knowledge

As a preliminary, notice that for any two events E and F , the event that E and F is the intersection of the two sets, $E \cap F$. For E and F are both true at a state w only if E is true at w and F is true at w . And $w \in E$ and $w \in F$ precisely if $w \in E \cap F$.

The definition of common knowledge is built up in three stages.

(i) Define a new operator E , “everyone knows that...”. Everyone knows an event if 1 knows it and 2 knows it and ... and n knows it, so

$$E(E) = \bigcap_{i=1}^n K_i(E).$$

(ii) Iterate the E operator to define a sequence of events, everyone knows that E , everyone knows that everyone knows that E , etc.

$$E^1(E) = E(E),$$

$$E^{k+1}(E) = E(E^k(E)) \text{ for } k \geq 1.$$

(iii) Common knowledge of E , written $C(E)$, is defined as the intersection of all the E^k events.

$$C(E) = \bigcap_{k=1}^{\infty} E^k(E).$$

Having introduced the notion of common knowledge, it is natural to ask when it might exist. Using the hierarchical procedure above, it could take a long time to check whether a particular event is common knowledge among a group of agents. A remarkable result first mentioned by Aumann [1] enables us to short-cut the whole procedure. The result uses the notion of a *self-evident* event. An event is self evident to an agent if that she knows that it is true whenever it actually is true. Formally, event E is self-evident to i if $E = K_i(E)$. It turns out that if E is self evident to every agent, then E is common knowledge whenever it is true. Furthermore, even if E is *not* self-evident to every agent, but we can find an event F which *is*, and F is a subset of E (i.e. every state in F

is also in E), then E is common knowledge at every state in F . Box 3 gives a formal statement of this result, and an alternative characterization of common knowledge.

Box 3: Characterizing common knowledge

1. An event E is common knowledge at state w if and only if there is an event F which is self-evident to every agent, such that $w \in F \subseteq E$.

The task of checking for common knowledge can be further simplified if we are willing to accept another round of definitions. Given any two partitions, \mathcal{P}_i is said to be *finer* than \mathcal{P}_j (and \mathcal{P}_j *coarser* than \mathcal{P}_i) if $\mathcal{P}_i(w) \subseteq \mathcal{P}_j(w)$ for all $w \in W$ (this can be thought of as a way of formalizing the idea that agent i has better information than agent j). The *meet* of a collection of partitions $\mathcal{P}_1, \dots, \mathcal{P}_n$ is the finest partition which is at least as coarse as each \mathcal{P}_i . The symbol \sqcap is often used to denote the meet of two partitions; let $\mathcal{M} = \mathcal{P}_1 \sqcap \dots \sqcap \mathcal{P}_n$.

2. An event E is common knowledge at w if and only if $\mathcal{M}(w) \subseteq E$, i.e. $C(E) = \{w : \mathcal{M}(w) \subseteq E\}$.

Self-evident events are easy to find, so it is easy to check for common knowledge. Consider again the previous example. Since there is no subset of $E = \{10, 12, 32\}$ that is self-evident to both agents (try to find one!), there can be no state w at which E is common knowledge. Our calculations above confirm that this is the case.

1.2 Introducing probabilities

The Aumann structure model is an excellent tool for analyzing agents' knowledge about the world and about each other. But there will typically be many events which agents do not know, and in economic applications and many other contexts it is reasonable to assume that they make their decisions based on probabilistic beliefs. Aumann's model can be extended to deal with probabilistic beliefs simply by including a probability distribution p_i over the set W of states for each agent i . The idea is that each p_i function encodes the agent's *prior* beliefs about the states. From this we can calculate beliefs about an event simply by adding up the probability of each state at which that event is true. When the agent receives information according to her information partition (i.e. agent i learns that the true state is one of those in $\mathcal{P}_i(w)$, if the true state is w), she updates

these beliefs using the formula for conditional probabilities to obtain *posterior* beliefs. The formal version is given in Box 4 below, but intuitively this formula says that agent i 's posterior belief at state w that E is true (denoted $p_i^w(E)$) is equal to her prior belief that $\mathcal{P}_i(w)$ and E divided by her prior belief that $\mathcal{P}_i(w)$.

An example will make it easier to see what is going on. Returning to the case of Alice and Bob and their father, suppose that Alice is pessimistic about her father's wealth. She believes that he is most likely to have only £1; he is half as likely to have £3, half as likely again to have £5, and so on. Thus state 10 is twice as likely as state 12, which is twice as likely as state 23, and so on. Thus Alice's prior beliefs are as follows:

	0	2	4	6	...
1	$\frac{1}{2}$	$\frac{1}{4}$			
3		$\frac{1}{8}$	$\frac{1}{16}$		
5			$\frac{1}{32}$	$\frac{1}{64}$	
7				$\frac{1}{128}$	
⋮					

p_A

How likely does she consider event E , that she nor her brother has more than £3? Her prior belief that E is true is simply the sum of the probabilities of each state in which E is true: $p_A(E) = p_A(10) + p_A(12) + p_A(32) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$. Now suppose that the true state is 32. When Ann looks into her own envelope she will find £3; she knows that the state is either 32 or 34, but cannot tell which. Her prior belief that she will receive this information and that E is true is $\frac{1}{8}$ ($= p_A(32)$), and her prior belief that she will receive this information is $\frac{3}{16}$ ($= p_A(32) + p_A(34)$). Dividing the first by the second, we obtain $p_A^{32}(E) = \frac{2}{3}$. This makes sense: 32 and 34 are the only states which Ann considers possible, and the first of these (at which E is true) she considers to be twice as likely as the second (at which E is false). Thus E must be twice as likely as not.

Box 4: Probabilistic beliefs

An Aumann structure with probabilities can be written as

$$\langle W, \mathcal{P}_1, \dots, \mathcal{P}_n, p_1, \dots, p_n \rangle.$$

Agent i 's prior beliefs about an event E are calculated simply by summing up over all the states in that event:

$$p_i(E) = \sum_{w \in E} p_i(w).$$

Posterior beliefs at state w are obtained by conditionalizing on the information she has at that state:

$$p_i^w(E) = \frac{p_i(\mathcal{P}_i(w) \cap E)}{p_i(\mathcal{P}_i(w))}.$$

1.3 Beyond information partitions

In this (technical) section we examine the foundations of the Aumann structure model, and show that it makes some very strong assumptions about agents' rationality.

The information partitions in an Aumann structure enable us to analyze exactly what information each agent has at each state of the world. But there is a more direct way of doing this: an *information function* for agent i is a function which associates with every state in W a (nonempty) subset of W . Letting \mathcal{I}_i denote this function, $\mathcal{I}_i(w)$ is to be interpreted as the set of states which i considers possible when the true state is w . An information function works in exactly the same way as an information partition, but it is more general: all information partitions can be represented by information functions but the converse is not true.

To see how information functions and information partitions differ, consider a very simple model with only three states, $W = \{1, 2, 3\}$, an one agent, A . Suppose that $\mathcal{I}_A(1) = \{1, 2\}$; $\mathcal{I}_A(2) = \{1\}$; and $\mathcal{I}_A(3) = \{2, 3\}$. Is there any information partition \mathcal{P}_A of W which represents this information function, in the sense that $\mathcal{I}_A(w) = \mathcal{P}_A(w)$ for all w ? It is easy to see that there is not. An information partition divides W into separate cells, so that every state is in exactly one cell. If $\mathcal{P}_A(w)$ is the cell which contains w , clearly w must always be a member of $\mathcal{P}(w)$. But $2 \notin \mathcal{I}_A(2)$. Furthermore, if x is contained in $\mathcal{P}_A(w)$, this means w and x are in the same cell of \mathcal{P}_A , and so $\mathcal{P}_A(w) = \mathcal{P}_A(x)$. Here, however, 2 is member of $\mathcal{I}_A(3)$, and yet $\mathcal{I}_A(2) \neq \mathcal{I}_A(3)$. Thus \mathcal{I}_A cannot be represented by any partition.

It turns out that these two conditions are not only necessary for an information function to be representable by an information partition, they are also sufficient. Stated more formally:

For any information function \mathcal{I}_i over W , there exists an information partition \mathcal{P}_i of W such that $\mathcal{I}_A(w) = \mathcal{P}_A(w)$ for all w if and only if

- I1** $w \in \mathcal{I}_i(w)$ for all $w \in W$; and
- I2** if $x \in \mathcal{I}_i(w)$, then $\mathcal{I}_i(w) = \mathcal{I}_i(x)$.

This result tells us that it is correct to think of an information partition as a special kind of information function, which satisfies **I1** and **I2**.

An *information structure* consists of a set of states and an information function for each agent:

$$\langle W, \mathcal{I}_1, \dots, \mathcal{I}_n \rangle.$$

Just as an information partition is a special kind of information function, an Aumann structure is a special kind of information structure. The knowledge operators of each agent are defined in exactly the same way as before. Agent i knows that E at state w precisely if E is true at every state she considers possible:

$$\mathsf{K}_i(E) = \{w : \mathcal{I}_i(w) \subseteq E\}.$$

For any information function, the knowledge operator must satisfy the following three properties:

- K1** $\mathsf{K}_i(W) = W$
- K2** If $E \subseteq F$ then $\mathsf{K}_i(E) \subseteq \mathsf{K}_i(F)$
- K3** $\mathsf{K}_i(E) \cap \mathsf{K}_i(F) = \mathsf{K}_i(E \cap F)$

K1 says that the agent knows all tautologies. Tautologies are propositions that are always true, and so the event representing any tautology is W , the set of all states. If $\mathsf{K}_i(W) = W$, then at every state the agent knows that W . **K1** is a stronger assumption than it may at first seem. Some tautologies are very complex (Fermat's Last Theorem is a good example), so an agent who knows all of them is a very powerful reasoner indeed.

K2 says that if E implies F , then knowledge of E implies knowledge of F . If one event is a subset of another, then at any state at which the first is true the second is also true; so the first

implies the second. **K2** is also a strong assumption: if it holds, then arguments are a waste of time: people can disagree about basic facts, but not about the implications of those facts.

K3 says that if an agent knows that E and knows that F , then she knows that E and F . It is reasonable to suppose that rational agents satisfy **K3**.

If we impose additional restrictions on \mathcal{I}_i , we can obtain further properties of knowledge. In particular, if \mathcal{I}_i satisfies **I1** we have:

$$\mathbf{K4} \quad K_i(E) \subseteq E$$

This says that everything the agent knows is actually true. **K4** is sometimes referred to as the Knowledge Axiom, and is thought of as the distinguishing feature of knowledge as opposed to belief. It would be strange to claim that an agent knew something that was in fact false, while it is not uncommon for beliefs to be mistaken.

If \mathcal{I}_i satisfies **I2** as well, we have:

$$\mathbf{K5} \quad K_i(E) \subseteq K_i(K_i(E))$$

$$\mathbf{K6} \quad W \setminus K_i(E) \subseteq K_i(W \setminus K_i(E))$$

K5, the Positive Introspection Axiom, says that if an agent knows something, she knows that she knows it; **K6**, the Negative Introspection Axiom, says that if an agent doesn't know something, she knows that she doesn't know it ($W \setminus E$ denotes the *complement* of the set E , i.e. the set of all states not in E ; thus $W \setminus E$ is the event that E is not true). Philosophers have cast doubt on these axioms; in particular, it has been argued that they may be appropriate for belief but not for knowledge. This is not the place to discuss these issues further. Rather, we shall satisfy ourselves with an example in which negative introspection fails. Imagine that Gary is unlucky enough to have a tutorial during the Football World Cup final (in 2006) when England are due to play Germany. He rushes off at the end of the tutorial to find out the the result: Germany won 1–0. Had England won, on the other hand, there would have been no need to ask the result; the cheers from the JCR would have been loud enough for him to hear on the other side of college. Here we have a failure of negative introspection: although Gary would have known that England had won had he heard cheers from the JCR, he fails to infer from the silence that England lost. To model this situation, let $W = \{E, G\}$, the two states corresponding to an English victory and a German victory respectively. If England win, Gary can tell that they have won from the cheers: $\mathcal{I}_i(E) = E$; but if Germany win, he doesn't hear any cheers and is unable to figure out who has won: $\mathcal{I}_i(G) = \{E, G\}$. Thus

$K_i(E) = E$, and $W \setminus K_i(E) = G$; but $K_i(W \setminus K_i(E)) = \emptyset$. So $W \setminus K_i(E) \not\subseteq K_i(W \setminus K_i(E))$, and **K6** is violated: if Germany win, Gary doesn't know that England won, but he doesn't know that he doesn't know it.

Where does all of this leave Aumann structures? We have seen that Aumann structures are information structures in which **I1** and **I2** are satisfied. This means that all agents satisfy **K1–K6**. These assumptions may or may not be reasonable in a particular context, but it is important that they are made explicit.

2 Solution concepts in game theory

A game is a formal model of a situation of strategic interaction. The key feature of a game as opposed to a single-agent decision problem is that the players' actions affect each other's payoffs. And it is well established both theoretically and empirically that strategic reasoning in games requires agents to form beliefs not just about each other's actions, but also about each other's knowledge and beliefs, which can then be used to infer what actions they might take.

On the other hand, game-theoretic solution concepts such as Nash equilibrium make no explicit reference to players' rationality or beliefs. Whether or not rational players must always play Nash equilibrium strategies is a question that has troubled game theorists for many years. Furthermore, Nash equilibrium is only one of an array of solutions concepts available to the game theorist: as well as a whole range of refinements, such as sequential equilibrium, there are also weaker concepts such as correlated equilibrium and iterated deletion of dominated strategies. Even the strongest of these does not guarantee a unique solution to every game, and real-life game players often play strategies that are not included even in the weakest.

So how can we adjudicate between these various solution concepts? An approach that has gained popularity in recent years is to abandon the search for a single solution concept that yields a unique prediction in all games (the Holy Grail of game theory), and to attempt instead a systematic evaluation of the existing concepts. In particular, it has been proposed that we link various assumptions about the knowledge and behavior of the players with each concept. The intention is to describe precise conditions under which rational players might be expected to act in the way described by a particular solution concept; these conditions can then be said to *characterize* the concept.

There is a very large and often complex literature on the epistemic foundations of solution con-

cepts. Seminal papers include Bernheim [6] and Pearce [17], who examine the implications of common knowledge of rationality; Aumann [2], who investigates correlated equilibrium; and Aumann and Brandenburger [3], who provide characterization theorems for Nash equilibrium. Comprehensive surveys of work in this area are provided by Brandenburger [7], Dekel & Gul [8], and Battigalli and Bonanno [5]. Here we present just a few of the more important results. The following analysis is most closely related to Stalnaker [19].

2.1 Models for games

A normal form game is defined by a set of players, a strategy set for each player, and a payoff function for each player:

$$G = \langle n, S_1, \dots, S_n, u_1, \dots, u_n \rangle.$$

Letting $S = \times_i S_i$ denote the set of strategy profiles³, each u_i is a function from S into the real numbers: $u_i(s)$ is i 's payoff if strategy profile s is played. It will be convenient to use $S_{-i} = \times_{j \neq i} S_j$ to denote the set of strategy profiles of i 's opponents. If $s_i \in S_i$ and $s_{-i} \in S_{-i}$, then $(s_i, s_{-i}) \in S$ is the resulting strategy profile when i plays s_i and everyone else plays s_{-i} . We shall also need to consider mixed strategies for each player: $\Sigma_i = \Delta S_i$ is the set of player i 's mixed strategies, formed by taking all the possible probability distributions over S_i . Σ is the set of all mixed strategy profiles, and Σ_{-i} is the set of mixed strategy profiles of i 's opponents. Just as we used s_i , s_{-i} and s to refer to typical members of S_i , S_{-i} and S , we shall use σ_i , σ_{-i} and σ to refer to typical members of Σ_i , Σ_{-i} and Σ . The range of u_i must also be extended. To calculate the payoff to player i resulting from a mixed strategy profile, we multiply the probability of each pure strategy profile by the payoff if that profile is played, and sum up over all the pure strategy profiles.

We can use the methods developed above to analyze rational play in games. A *model* of a game gives a complete description of each player's knowledge and (probabilistic) beliefs, in much the same way as an Aumann structure with probabilities. It consists of a set of states, W ; and an information partition \mathcal{P}_i , probability distribution p_i , and *strategy function* f_i for each agent. The only new element here is the strategy function: it tells us which strategy the agent plays at every state. In order to capture the assumption that each player knows her own strategy choice, we must assume that she plays the same strategy in every world she considers possible: if $x \in \mathcal{P}_i(w)$, then

³ $\times_i S_i$ is the *Cartesian product* of all the S_i 's, i.e. all the ways of taking precisely one object from each set. For example, if $S_1 = \{U, D\}$ and $S_2 = \{L, R\}$, then $S_1 \times S_2 = \{(U, L), (U, R), (D, L), (D, R)\}$.

$$f_i(w) = f_i(x).$$

We are now in a position to define what it is for a player to be rational. As usual, we shall identify rationality with expected utility maximization. A player's expected utility from choosing a particular strategy will vary across the states, since her information and beliefs vary across states. Let $EU_i^w(s_i)$ denote her expected utility at state w if she chooses s_i . The formal definition can be found in Box 5 below, but it is sufficient to know that $EU_i^w(s_i)$ is calculated by multiplying the probability of each state (e.g. $p_i^w(x)$ for state x) and the utility obtained in that state if player i plays s_i ($u_i(s_i, f_{-i}(x))$ for state x), and then summing over all the states. Two points are worthy of note. First, $p_i^w(x)$ will typically equal zero for many states: precisely those states which i does *not* consider possible given the information she has at state w . Second, $u_i(s_i, f_{-i}(x))$ is not the utility player i actually gets at state x , unless $s_i = f_i(x)$. It is the utility she *would* get if she played s_i , given that all of her opponents stick with $f_{-i}(x)$. If she changes her strategy she changes the state, and so possibly her beliefs about the other players' strategy choices; but it is her beliefs about s_{-i} in the actual state that are important for evaluating expected utility. We are keeping these beliefs fixed while varying s_i .

A rational agent is one who maximizes expected utility. A player is rational at a particular state, then, if her strategy choice at that state yields at least as much expected utility as any other strategy choice. We denote this event RAT_i (i.e. RAT_i is the set of all states at which player i is rational). RAT is the event that everyone is rational, and finally CKR is the event that there is common knowledge of rationality.

Box 5: Models of games

A model of a game consists of the following elements:

$$\langle W, \mathcal{P}_1, \dots, \mathcal{P}_n, p_1, \dots, p_n, f_1, \dots, f_n \rangle,$$

where W is a set of states, and \mathcal{P}_i, p_i , and $f_i : W \rightarrow S_i$ are player i 's information partition, prior probability distribution, and strategy function respectively. It is assumed that $f_i(w) = f_i(x)$ for all $x \in \mathcal{P}_i(w)$. For convenience, we define $f = (f_1, \dots, f_n)$ and $f_{-i} = (f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n)$, so that $f(w) \in S$ denotes the strategy profile played at state w , and $f_{-i}(w)$ denotes the strategy profile of all of i 's opponents.

The definition of $EU_i^w(s_i)$ was discussed in the main text. Formally,

$$EU_i^w(s_i) = \sum_{x \in W} p_i^w(x) \cdot u_i(s_i, f_{-i}(x)).$$

A player is rational at a particular state if her chosen strategy (as given by f_i) yields at least as much expected utility as any alternative strategy. The events RAT_i , RAT , and CKR are defined as expected:

$$\begin{aligned} RAT_i &= \{w : EU_i^w(f_i(w)) \geq EU_i^w(s_i), \text{ for all } s_i \in S_i\} \\ RAT &= \bigcap_{i=1}^n RAT_i \\ CKR &= C(RAT) \end{aligned}$$

The next sections discuss some characterization results. In general, a set of assumptions characterizes a solution concept if:

- (i) at every state in every model in which the assumptions are satisfied, the players' strategies are consistent with the solution concept;
- (ii) if the players' strategies are consistent with the solution concept, then we can find a model and a state in that model in which the assumptions are satisfied.

Intuitively, a typical characterization result will tell us that players with certain beliefs will behave in a particular way; and if they behave in a particular, we can explain that behavior by supposing they have certain beliefs.

2.1.1 Iterated deletion of dominated strategies

A strategy is (strictly) dominated for a player if she has another available strategy which beats it whatever her opponents do. The archetypal example of a dominated strategy is cooperation in the Prisoner's Dilemma game: whether one's opponent chooses to cooperate or defect, defection yields a higher payoff than cooperation. The formal definition is as follows:

The strategy s_i of player i is *dominated* if there is a mixed strategy $\sigma_i \in \Sigma_i$ such that $u_i(\sigma_i, s_{-i}) > u_i(s_i, s_{-i})$ for all $s_{-i} \in S_{-i}$.

(Note that a pure strategy may be beaten by a mixed strategy even if it is not beaten by any pure strategies. We shall see an example of this shortly.) It is easy to see that a rational player will never play a dominated strategy: whatever she believes her opponents are going to do, she will obtain higher expected utility if she plays the strategy which dominates it, and so the original strategy cannot be expected utility maximizing. So if we are trying to predict the behavior of rational players, we can rule out all strategies which are dominated. Iterated deletion of dominated strategies repeats this process again and again. Consider the example below.

	L	C	R
U	0,2	3,1	2,3
M	1,4	2,1	4,1
D	2,1	4,4	3,2

Careful examination of this game reveals that U is dominated by D for player 1 (who chooses rows): whatever player 2 does, U yields a lower payoff than D . We delete U and look at the 2×3 game that remains. R is now dominated by a $\frac{1}{2}, \frac{1}{2}$ mix between L and C for player 2. If player 1 plays M , player 2 obtains an expected payoff of $2\frac{1}{2}$ from the mixed strategy compared with a payoff of 1 from R ; and if player 1 plays D , player 2 obtains an expected payoff of $2\frac{1}{2}$ from the mixed strategy compared with a payoff of 2 from R ; we do not need to consider what happens if player 1 plays U , since that strategy has already been deleted. We are now left with a 2×2 in which player 1 chooses between M and D and player 2 chooses between L and C . M is dominated by D in this game; once M has been deleted, L is dominated by C and we are left with (D, C) as the unique strategy profile which survives iterated deletion of dominated strategies. Each stage of this process is shown below.

	<i>L</i>	<i>C</i>	<i>R</i>		<i>L</i>	<i>C</i>	<i>R</i>		<i>L</i>	<i>C</i>		
<i>U</i>	0, 2	3, 1	2, 3	→	<i>M</i>	1, 4	2, 1	4, 1	→	<i>M</i>	1, 4	2, 1
<i>M</i>	1, 4	2, 1	4, 1		<i>D</i>	2, 1	4, 4	3, 2		<i>D</i>	2, 1	4, 4
<i>D</i>	2, 1	4, 4	3, 2		<i>D</i>	2, 1	4, 4	3, 2		<i>D</i>	2, 1	4, 4

	<i>L</i>	<i>C</i>		<i>C</i>
<i>D</i>	2, 1	4, 4	→	4, 4

Box 6 gives the formal definition of D , the set of strategy profiles which survive iterated deletion of dominated strategies.

Box 6: Iterated deletion of dominated strategies

Define a sequence of strategy sets for each player D_i^0, D_i^1, \dots where

$D_i^0 = S_i$; and

$D_i^{k+1} = \{s_i : \text{there is no } \sigma_i \in \Sigma_i \text{ such that } u_i(\sigma_i, s_{-i}) > u_i(s_i, s_{-i}) \text{ for all } s_{-i} \in D_{-i}^k\}$

Starting with the full strategy set of each player, at each stage we delete strategies which are dominated given that everyone can play only strategies which have survived until the previous round. These strategy sets give us a corresponding sequence of sets of strategy profiles D^0, D^1, \dots . The set of strategy profiles which survive iterated deletion of dominated strategies is the limit of this sequence:

$$D = \lim_{k \rightarrow \infty} D^k.$$

As long as the original game is finite (i.e. n is finite and each S_i is finite), it can be shown that this limit is well defined and nonempty.

Iterated deletion of dominated strategies does not usually have as much predictive power as in the game above. In most games, several strategies for each player will survive the process, and in many games none of the players have any dominated strategies so it cannot even get started. But the advantage of using such a weak solution concept is that it is widely applicable. It turns out that whenever there is common knowledge of rationality among the players, they will play strategies which survive iterated deletion of dominated strategies.

The intuition behind this result is straightforward. We have already said that rational players will not play dominated strategies. Then if all the players know they are all rational, they know none will play a dominated strategy. And being rational, they will play strategies that are not dominated in the game remaining after one round of deletion. This gives us two rounds of deletion. If they all know that they all know they are all rational, we have three rounds of deletion, and so on. Thus common knowledge of rationality gives us iterated deletion of dominated strategies.

Does this result exhaust the implications of common knowledge of rationality? Are there any strategies which we can rule out in addition to those eliminated by the iterated deletion procedure? In fact the answer is no: every strategy profile which survives the procedure is consistent with common knowledge of rationality given appropriate beliefs of the players.

The proof relies on the following (non trivial) result: if a player's strategy is not dominated, then we can find beliefs for the player which make that strategy a rational choice. (This is the converse of the result which says that rational players do not play dominated strategies.) If the strategy survives two rounds of deletion, then, we can find beliefs for the player which make that strategy a rational choice given that she places positive probability only on strategies of her opponents which survived the first round of deletion, i.e. such that she is rational and knows that everyone else is rational. Repeating this argument completes the proof.

We now have our first characterization result: common knowledge of rationality characterizes iterated deletion of dominated strategies (the formal version is given in Box 7). What is the implication of this result? If all we know or wish to assume about a certain strategic situation is that the players possess common knowledge of rationality, then (i) we can be sure that they will play a strategy profile which survives iterated deletion of dominated strategies, but (ii) we cannot say anything more than this. In game theory as elsewhere in economic theory, weak assumptions yield weak results.

Box 7: Characterization of iterated deletion of dominated strategies

- (i) Suppose w is a state in some model of game G . If $w \in CKR$ then $f(w) \in D$.
- (ii) Suppose $s \in D$ for some game G . Then there is a state w in some model of G such that $w \in CKR$ and $f(w) = s$.

2.1.2 Nash equilibrium

Nash equilibrium is the most widely used solution concept in game theory. We have a Nash equilibrium if each player's strategy choice maximizes her expected utility taking the strategy choices of her opponents as given. More formally,

A strategy profile $s \in S$ is a *Nash equilibrium* if, for every player i , $u_i(s) \geq u_i(s'_i, s_{-i})$, for all $s'_i \in S_i$.

Let N denote the set of Nash equilibrium strategy profiles. Nash equilibrium is a stronger solution concept than iterated deletion of dominated strategies: every Nash equilibrium strategy profile survives iterated deletion of dominated strategies, but there are some profiles which survive iterated deletion and are not Nash equilibria ($N \subseteq D$). This might lead us to expect that a stronger set of assumptions characterizes Nash equilibrium. In fact this is not quite true: the assumptions are stronger in one sense but weaker in another.

The formal details can be found in Box 8, but in words the characterization result for Nash equilibrium says that (i) if the players are all rational and know each other's strategies, then they will play a Nash equilibrium strategy profile; and (ii) every Nash equilibrium could be played by players who are rational and know each other's strategies. More neatly: mutual⁴ rationality and knowledge of the strategy profile characterize Nash equilibrium. So common knowledge of rationality is no longer required; it is enough that everyone is rational, with the additional assumption that the strategy profile is known. The reason is simple. Rational players maximize expected utility given their beliefs. If they actually know which strategies everyone else is choosing, they will play the utility maximizing response. This is precisely what is required by the definition of a Nash equilibrium.

Box 8: Characterization of Nash equilibrium

(i) Suppose w is a state in some model of a game G . If $w \in RAT \cap K_1(f(w)) \cap \dots \cap K_n(f(w))$ then $f(w) \in N$.

(ii) Suppose $s \in D$ for some game G . Then there is some state w in some model of G such that $w \in RAT \cap K_1(f(w)) \cap \dots \cap K_n(f(w))$ and $f(w) \in N$.

⁴A *mutual* property is one that everyone possesses; so mutual rationality means that everyone is rational, and mutual knowledge is something that everyone knows.

This characterization result sums up much that is good and much that is bad about Nash equilibrium. All we need to assume is that we have rational players who know each other's strategies. But we are begging the question of how they come to have this knowledge. One answer is that they might figure it out from assumptions about each other's rationality. But even common knowledge of rationality, as we have seen, does not usually predict a unique strategy profile; all it gives us is iterated deletion of dominated strategies. The unfortunate truth is that there is no reason in general why the players should know each other's strategies. The Holy Grail of game theory was never found, and never will be. We must look at the particular game that is being played, and the context in which it is being played, before we can assess the reasonableness of this assumption.

Before moving on, we issue an important *caveat*. The definition of a Nash equilibrium presented here applies only to a Nash equilibrium in *pure strategies*. But in many games there is no pure strategy Nash equilibrium. Nash's existence theorem stating that every finite game has a Nash equilibrium assumes that the players are allowed to play mixed strategies which randomize over their pure strategies. But in the models we have considered, each state represents a particular way the game might be played, and we have assumed that each player knows her own strategy choice. We could relax this assumption, and allow players to randomize, obtaining a characterization result analogous to that above. But the idea that players actually play mixed strategies is troubling on several counts. First, there is never any incentive for randomization. In a mixed strategy Nash equilibrium players must obtain the same expected utility from each of the pure strategies they are randomizing over. And second, most people have the feeling that they simply do not randomize when making decisions (though they may have a hard time making up their minds). Indeed, psychologists tell us that people *cannot* randomize, even when they want to. An alternative interpretation of a mixed strategy is to think of it as representing uncertainty in the mind of the player's opponents about what she will do⁵. In two-player games this works well: even though both players actually play pure strategies, we can have an equilibrium in *conjectures* (i.e. the probabilities each places on the other's pure strategy choices).

Thinking of mixed strategies as conjectures, it can be shown that, in two-player games, mutual knowledge of rationality and of conjectures characterize mixed-strategy Nash equilibrium. To see why knowledge of rationality is required here when it wasn't for the pure strategy case, suppose that (σ_1, σ_2) is a mixed strategy Nash equilibrium; σ_1 is interpreted as player 2's conjecture about what player 1 will do, and σ_2 as player 1's conjecture about what 2 will do. If the players were

⁵This idea is originally due to Harsanyi [11], and is discussed in detail by Aumann [2].

merely rational but did not know each other to be rational, there would be no restrictions on these conjectures, and all we could say is that the players would play strategies which are not dominated. But now suppose that player 1 knows player 2 is rational, and knows player 2's conjecture, σ_1 . Then every strategy of player 2 assigned positive weight by player 1's conjecture, σ_2 , must be expected utility maximizing given σ_1 . And by the same reasoning, each of player 1's strategies assigned positive weight by player 2's conjecture, σ_1 , must be expected utility maximizing given σ_2 . This is precisely what is required for a mixed strategy Nash equilibrium.

In the case of games with more than two players, the result becomes more complicated. To interpret mixed strategies as conjectures, all of a given player's opponents must have the *same* beliefs about what she will do. Additional conditions are required to guarantee that this will be the case: that the players have a common prior probability distribution over the set of states, and that their conjectures are common knowledge⁶. For more details, the interested reader is referred to Brandenburger [7] or Aumann and Brandenburger [3].

3 Conclusion

In this paper we have presented a formal model of knowledge which allows us to analyze what agents know about the world and about each other, and we have shown how this model can be used to provide a systematic evaluation of game theoretic solution concepts by describing the circumstances under which it might be appropriate to use a particular concept. Conspicuous by its absence is any discussion of the many refinements of Nash equilibrium that have been developed over the years. Although some of these refinements, such as trembling-hand perfection and proper equilibrium, relate to the normal form of the game, the majority use the extra structure provided by the extensive form to reduce the set of Nash equilibria. For example, subgame perfect equilibrium and perfect Bayesian equilibrium refer to information sets and decision nodes, both features of extensive form games, in their definitions.

Much progress has been made in this area (details can be found in the more recent surveys mentioned in section 2), but dealing with extensive form games raises additional issues beyond the scope of this paper. In particular, if agents take turns to move, they can respond to each other's actions. Observing moves that are made as the game progresses, they will gain more information. It is not reasonable to assume that their knowledge and beliefs remain unchanged. But the Aumann

⁶This result is related to Aumann's [1].famous "agreeing to disagree" theorem, discussed in Morris [14].

structure model and its extensions discussed in section 1 are essentially static. The dynamic models that have been developed and used to analyze extensive form solution concepts are necessarily much more complex than Aumann structures; and the mechanism by which information is received and updated by the players is still a matter for open debate.

What of the extensions to Aumann structures discussed in section 1.3? How are the characterization results affected if we replace information partitions with the more general information structures? The answer is, surprisingly, very little. Most of the results survive unchanged, although some differences emerge when we consider dynamic models and extensive form games. Details can be found in Dekel & Gul [8].

We hope we have managed to convince the reader that formal models of knowledge and beliefs are a useful tool of analysis for adjudicating between the plethora of solution concepts on offer in game theory. The specification of a game is very sparse, and in most cases does not provide enough information for us to predict how the players will behave. A single game might be used to represent two very different situations of strategic interaction, and until we are given more details about the context in which the game is being played, it may be difficult to say much about what will happen. The more details we have, the more predictive power we have. Characterization results make the link between contextual details and particular solution concepts precise: for example, if we can expect that the players are rational and know each others strategies, we can predict they will play a pure strategy Nash equilibrium, but nothing more than this. If the game has two Nash equilibria, these assumptions and the description of the game itself do not allow us to choose between them. But often we may have more information about the context than this. Formal models of knowledge can show us how to use it.

References

- [1] Aumann, R. J. (1976), “Agreeing to Disagree”, *Annals of Statistics* **4**, 1236–1239.
- [2] Aumann, R. J. (1987), “Correlated Equilibrium as an Expression of Bayesian Rationality”, *Econometrica* **55**, 1–18.
- [3] Aumann, R. J. and A. Brandenburger (1995), “Epistemic Conditions for Nash Equilibrium”, *Econometrica* **63**, 1161–1180.

- [4] Bacharach, M. (1985), “Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge”, *Journal of Economic Theory* **37**, 167–190.
- [5] Battigalli, P, and G. Bonanno (1999), “Recent results on belief, knowledge and the epistemic foundations of game theory”, *Research in Economics* **53**, 149–225.
- [6] Bernheim, B. D. (1984), “Rationalizable Strategic Behavior”, *Econometrica* **52**, 1007–1028.
- [7] Brandenburger, A. (1992), “Knowledge and Equilibrium in Games”, *Journal of Economic Perspectives* **6**, 83–101.
- [8] Dekel, E. and F. Gul (1997), “Rationality and Knowledge in Game Theory”, pp. 87–172 in *Advances in Economics and Econometrics* vol. I, ed. by D. M. Kreps and K. F. Wallis. pp. 87–172. Cambridge University Press, Cambridge, UK.
- [9] Fagin, R., J. Y. Halpern, Y. Moses and M. Y. Vardi (1995), *Reasoning about Knowledge*. The MIT Press, Cambridge, MA.
- [10] Geanakoplos, J. (1992), “Common Knowledge”, *Journal of Economic Perspectives* **6**, 53–82.
- [11] Harsanyi, J. C. (1973), “Games with Randomly Disturbed Payoffs: A New Rationale for Mixed-Strategy Equilibrium Points”, *International Journal of Game Theory* **2**, 1–23.
- [12] Hintikka, J. (1962), *Knowledge and Belief*. Cornell University Press, Ithaca, NY.
- [13] Lewis, D. (1969), *Conventions: A Philosophical Study*. Harvard University Press, Cambridge, MA.
- [14] Morris, S. (2002), “Coordination, Communication and Common Knowledge: A Retrospective on the Electronic Mail Game”, *Oxford Review of Economic Policy* **??**, ???–???
- [15] Myatt, D., H. S. Shin and C. Wallace (2002), “The Assessment: Games and Coordination”, *Oxford Review of Economic Policy* **??**, ???–???
- [16] Osborne, M. J. and A. Rubinstein (1984), *A Course in Game Theory*. The MIT Press, Cambridge, MA.
- [17] Pearce, D. G. (1984), “Rationalizable Strategic Behavior and the Problem of Perfection”, *Econometrica* **52**, 1029–1050.

- [18] Samet, D. (1990), “Ignoring Ignorance and Agreeing to Disagree”, *Journal of Economic Theory* **52**, 190–207.
- [19] Stalnaker, R. (1994), “On the Evaluation of Solution Concepts”, *Theory and Decision* **37**, 49–73.