COMMON KNOWLEDGE

by

John Geanakoplos

Note: This paper is a draft of a survey of common knowledge prepared for the *Journal of Economic Perspectives*. Another, more technical version, has been prepared for the *Handbook of Game Theory*.

I wish to acknowledge many inspiring conversations, over the course of many years, I have had with Bob Aumann on the subject of common knowledge.

I. INTRODUCTION

People, no matter how rational they are, usually act on the basis of incomplete information. If they are rational they recognize their own ignorance and reflect carefully on what they know and what they do not know, before choosing how to act. Furthermore, when rational agents interact, they also think about what the others know, and what the others know about what they know, before choosing how to act. Failing to do so can be disastrous. When the notorious evil genius Professor Moriarty confronts Sherlock Holmes for the first time he shows his ability to think interactively by remarking that "all I have to say has already crossed your mind"; Holmes, even more adept at that kind of thinking responds "then possibly my answer has crossed yours." On account of Moriarty's limited mastery of interactive epistemology, he let Holmes and Watson escape from the train at Canterbury, a mistake which ultimately led to his death, because he himself went on to Paris after calculating that Holmes would normally go on to Paris, failing to deduce that Holmes had deduced that he would deduce what Holmes would normally do and in this circumstance get off earlier.

Knowledge and interactive knowledge are central elements in economic theory. Nobody can buy stock unless somebody else is selling it. If the buyer has information suggesting the price will go up, he must consider that the seller might have information indicating that the price will go down. If the buyer further considers that the seller is willing to sell the stock having also taken into account that the buyer is willing to purchase the stock, he must ask himself, should he still buy? Does the answer depend on how rational the agents are? For example, suppose one of them always ignores unpleasant news. Does that affect the chances for a sale?

Can rational agents agree to disagree? Is there a connection between this question and whether rational agents will speculate in the stock market? What relevance to this question is the degree of rationality of the agents? Or the length of time they talk before agreeing to disagree?

A crucial role in the analysis of these questions is played by the notion of common knowledge. We say that an event E is common knowledge among a group of agents if each one knows it, and if each one knows that the others know it, and if each one knows that each one

knows that the others know it, and so on. Common knowledge is thus the limit of a potentially infinite chain of reasoning about knowledge.

In different situations, different kinds of events are common knowledge, and with different consequences. Public events are the most obvious candidates for common knowledge, even when their occurrence is due to causes entirely independent of the agents in question. When the agents bring about the events themselves, as for example in collectively designing the rules of some game or agreeing to some contract, the plausibility of common knowledge is strengthened. Certain facts about human nature might also be taken to be common knowledge. We are especially interested, for example, in the consequences of the hypothesis that it is common knowledge that all agents are optimizers, i.e. maximize their utilities. Finally, it often comes about after lengthy periods of observing behavior that what people are going to do is common knowledge, though the reasons for their actions may be difficult to disentangle.

The purpose of this chapter is to survey some of the implications for economic behavior of the hypotheses that events are common knowledge, that actions are common knowledge, that optimization is common knowledge, and that rationality is common knowledge.

In general we shall discover that a talent for interactive thinking is advantageous, but if everyone can think interactively, and infinitely deeply all the way to common knowledge, then sometimes there are puzzling and counterintuitive consequences. We are therefore led to investigate whether bounded rationality can help explain some of the common sense events that we take for granted, but which are ruled out when they become common knowledge between agents who are perfectly rational.

In the rest of Section I we describe several well-known puzzles illustrating the strength of the common knowledge hypothesis. In Section II we formalize the definition of knowledge. In Section III we formalize the definition of common knowledge, and in Section IV we formalize learning and show how common knowledge can be achieved. In Section V we describe Bayesian equilibrium. In Section VI we describe speculation. In these latter two sections we derive two surprising consequences of common knowledge: common knowledge of actions negates asymmetric information about events, and common knowledge of rationality eliminates financial speculation, even between people who know different things. In Section VII we explore ways in which deviations from perfect rationality, or bounded rationality, can restore

some of the commonsense explanations of phenomena like speculation and cooperation that are ruled out by common knowledge of perfect rationality.

1. Two Old Puzzles Depending on Reasoning about the Reasoning of Others

The most famous example illustrating the ideas of reasoning about knowledge can be told in many equivalent ways.[1] The earliest telling of the story that I could find appears in Littlewood's Miscellania, published in 1953. He said it was well known and had caused a sensation in Europe some years before when it had become popular. In the colonial version of the story there are many cannibals married to unfaithful wives, and of course a missionary. I shall be content to describe a more prosaic version of the story involving a group of logical children wearing hats.

Imagine 3 children sitting in a circle, each wearing either a red hat or a white hat. Suppose that in fact all the hats are red. No matter how many times the teacher asks if there is any student who can identify the color of his own hat, there is always a negative response, since nobody can see his own hat. Now if the teacher happens to remark that there is at least one red hat in the room, a fact which is well-known to every child (who knows that there are at least 2 red hats in the room) then there is a surprising change in the answers to his question. The first student who is asked cannot answer, nor can the second. But the third will be able to answer with confidence that he is indeed wearing a red hat.

The story is surprising because aside from the apparently innocuous remark of the teacher, the students appear to learn from nothing except their own ignorance. Indeed this is precisely the case. The third student could answer because he reasoned that if his hat were white then the second student would have given a positive response since he would have figured that the first student must have been looking at that second student's red hat when he could not answer about his hat color.

There are several crucial elements to this story. First, everybody knows that everybody else can see 2 hats. In fact this is common knowledge. Second, the pronouncements of ignorance are public; each time a student maintains his ignorance, he knows that everyone else

---

[1]These examples are so well-known that it is difficult to find out who told them first. They appeared for example in Martin Gardner's column in *Scientific American* as early as 19xx. Our first example was first discussed in the economics literature by Geanakoplos-Polemarchakis (1982). It appeared in the computer science literature in Halpern-Vardi (1984).

knows he said he didn't know, etc. Third, everybody knows the reasoning used by everyone else. The apparently innocuous fact related by the teacher that there is at least one red hat in the room was known to all the students, but it was not common knowledge between them. Before the remark, the first student (who must consider that his own hat might be white) could not rule out that the second student might have thought that both his and his own hat were white, and that therefore the second student might have thought that the third student was looking at two white hats, in which case he would not be sure that there was any red hat in the room.

Consider a second example, also described by Littlewood, involving betting. A generous but mischievous father tells his two sons that he has placed 10_n dollars in one envelope, and 10^{n+1} dollars in the other envelope, where n is chosen with equal probability among the integers between 1 and 6. Since the father's wealth is well known to be 11 million dollars, the sons completely believe their father. He randomly hands each son an envelope. The first son looks inside his envelope and finds $10,000. Disappointed at the meager amount, he calculates that the odds are fifty fifty that he has the smaller amount in his envelope. Since the other envelope contains either $1,000 or $100,000 with equal probability, the first son realizes that the expected amount in the other envelope is $50,500. Unbeknownst to him the second son has seen that there are only $1,000 in his envelope. Based on his information, he expects to find either $100 or $10,000 in the first son's envelope, which at equal odds comes to an expectation of $5,050. The father privately asks each son whether he would be willing to pay $1 to switch envelopes, in effect betting that the other envelope has more money. Both sons say yes. The father then calls both his sons in together and tells them that they have each offered to pay $1 to switch envelopes, and asks them to shake hands on the deal, it being understood that if either son refuses the deal is off. The sons take a hard look at each other. What should they do? Suppose instead that the sons were not permitted to look at each other, but instead they had to write their confirmation of the deal on separate pieces of paper and hand them to their father. What should they write?

In Sections II and III we shall discuss a model of knowledge based on possible worlds formulated in the 1920's by the philosopher C. I. Lewis, and in the 1960's by Hintikka and Kripke, which will allow us to formalize these popular puzzles, as well as some phenomena of greater economic significance.

2. A Newer Puzzle

Consider two detectives trained at the same police academy. Their instruction consists of a well-defined rule specifying who to arrest given the clues that have been discovered. Suppose now that an important murder occurs, and the two detectives are hired and ordered to conduct independent investigations. They promise not to share any data gathered from their research, and indeed they begin their sleuthing in different corners of the town.

Suddenly the detectives are asked to appear and announce who they plan to arrest. Neither has had the time to complete a full investigation, so they each have gathered different clues. On the way to the station they happen to meet. Recalling their pledges, they do not tell each other a single discovery, or even a single reason why they were led to their respective conclusions. But they do tell each other who they plan to arrest, and who each plans to arrest becomes common knowledge between them.

Then they must both announce the same suspect at the station! This is so even though if asked to explain their choices, they may each produce entirely different motives, weapons, scenarios, etc.

The conclusion depends on only two assumptions. First, both detectives use the same decision rule: given the same clues they would arrest the same suspect or suspects. Second, the decision rule specifying who to arrest must satisfy the sure-thing-principle. This means that if Moriarty should be arrested if the victim's blood type is O, given the other clues, and similarly if Moriarty should be arrested if the blood type is X, given the same other clues, then given those other clues and the clue that the blood type is either X or O (but not which), the detective should also arrest Moriarty.[2]

This third story appears a bit surprising. The detectives, who have different clues, and quite likely have in mind different scenarios for the murder, nevertheless must have the same suspect in mind once their choices are common knowledge. The riddle of course must lie in the (perhaps) unsuspected strength of the common knowledge hypothesis.

The importance of this last puzzle to economic behavior derives from the fact that much of economic theory rests on the foundation of Bayesian decision making, and Bayesian optimal decisions satisfy the sure-thing-principle. Indeed when Savage (1954) formulated the

[2]This story so far is originally due to Bacharach, perhaps embellished somewhat embellished by Aumann, from whom I learned it.

Bayesian framework that economists use, he named the sure-thing-principle as one of the main characteristics of Bayesian behavior. Examples of Bayesian decisions that we will discuss later include betting, "opining," trading, and more generally, optimizing in Bayesian games.

The moral of the detective story for economics becomes more relevant if the story is extended. Suppose now that the detectives went to different academies, but that the rules each teach are well-known. Each detective follows those of his own academy. Now after their conversation the detectives may announce different suspects. But these choices are not arbitrary. There must be a single set of clues which would lead the first detective, following his academy's prescription, to choose his suspect, and at the same time lead the second detective, following different instructions, to choose his suspect. This must be the case even though in fact the two detectives use quite different sets of clues.

It is commonplace in economics nowadays to say that many actions of optimizing, interacting agents can be naturally explained only on the basis of asymmetric information. But the riddle shows that if those actions are common knowledge, and if they can be explained by asymmetric information, then they can also be explained on the basis of symmetric information. In short, common knowledge of actions negates asymmetric information about events.

II. INTERACTIVE EPISTEMOLOGY

1. Possible Worlds

The fundamental conceptual tool we shall use is the notion of state of the world. First introduced by Leibnitz, the idea has been refined as it has been applied to game theory and other formal models of interactive epistemology by Kripke, Savage, Harsanyi, and Aumann, among others. We develop the features of a "state of the world" slowly, but ultimately a state shall specify the entire physical universe, past, present, and future; it shall describe what every agent knows, and what every agent knows about what every agent knows etc; it shall specify what every agent does, and what every agent thinks about what every agent does, and what every agent thinks about what every agent thinks about what every agent does etc; it shall specify the utility to every agent of every action, not only of those that are taken in that state of nature, but also those that hypothetically might have been taken, and it specifies what everybody thinks about the utility to everybody else of every possible action etc.; it specifies

not only what agents know, but what probability they assign to every event, and what probability they assign to every other agent assigning some probability to each event etc.

2. Possible Worlds and Partitions

Let Ω be a set which we shall interpret as the set of all possible worlds. How all-embracing Ω turns out to be will gradually become clear, so we shall defer our discussion of this issue until much later.

Consider an agent i. His knowledge will be described (throughout most of this survey) by a collection of mutually disjoint and exhaustive equivalence classes of states called cells, that is by a partition $\mathbf{P}_i$ of Ω. If two states of nature are in the same cell, then we say that agent i cannot distinguish them: if one is the true state of the world, then i cannot rule out the other.

For example, suppose that Ω is the set of integers from 1 to 8, and that agent i is told whether the true number is even or odd. Agent i's partition consists of two cells, $\{1, 3, 5, 7\}$ and $\{2, 4, 6, 8\}$. If the true state were 4, then i would think that any of the even states is possible, while none of the odd states is possible. If he were asked whether the true state is prime, he would not be sure, since as far as he knows, the true state might be 2, which is prime, or one of 4, 6, 8, which are not prime. If on the other hand the true state were 3 and he were asked whether the true state is prime he would be able to answer yes since all of the numbers 1, 3, 5, 7 are prime. The event E that the true state is prime is denoted by $\{1, 2, 3, 5, 7\}$. Whenever ω is in E, the true state is prime. In some states where the event E occurs, i knows it, and in some states where E occurs he does not know it.

We call any subset E contained in Ω an event. If the true state of the world is ω, then we say that E occurs or is true provided that $\omega \in E$. We say that i knows that E occurs when ω is the true state of the world, if for the cell $P_i(\omega)$ containing ω, $P_i(\omega) \subset E$. Thus we can think of $P_i(\omega)$ as the set of all states that i regards as possible when ω is the true state. When $P_i(\omega) \subset E$, every state that i thinks is possible (given that ω is the true state) entails E, so i must think $\sim E$ is impossible, that is he (thinks he) knows E. Following this interpretation of i's knowledge we shall usually refer to P_i as the possibility correspondence of agent i. Note that i may well know E at some $\omega \in E$, if $P_i(\omega) \subset E$, but not at other $\omega' \in E$, if $P_i(\omega') \not\subset E$.

If $P_i(\omega) \subset E$ for all $\omega \in E$, then we say that E is self-evident to i. Such an event E cannot happen unless i knows it.

In the example from the last paragraph, there are only three self-evident events to agent i: the event that the real state is even, the event that the real state is odd, and the all inclusive event that the real state is in Ω. In general, the self-evident events are the unions of partition cells.

So far we have described the knowledge of agent i by what he would think is possible in each state of nature. There is an equivalent way of representing the knowledge of agent i at some state ω, simply by enumerating all the events which the information he has at ω guarantees must occur. The crispest notation to capture this idea is a "knowledge operator" K_i taking any event E into the set of all states at which i is sure that E has occurred: $K_i(E) = \{\omega \in \Omega : P_i(\omega) \subset E\}$. At ω, agent i has enough information to guarantee that event E has occurred iff $\omega \in K_i(E)$. Thus in the example above, $K_i(\{1, 3, 5, 7\}) = \{1, 3, 5, 7\} = K_i(\{\omega : \omega \text{ is prime}\})$, while $K_i(\{1,4\}) = \phi$. A self-evident event can now be described as any subset E of Ω satisfying $K_i(E) = E$, i.e. the self evident events are the fixed points of the K_i operator.

As long as the possibility correspondence P_i is a partition, the knowledge operator applied to any event E is the union of all the partition cells that are completely contained in E. Thus it can easily be checked that the knowledge operator K_i derived from the partition possibility correspondence P_i satisfies the following five axioms: for all events A and B contained in Ω,

(1) $K_i(\Omega) = \Omega$. It is self evident to agent $_i$ that there are no states of the world outside of Ω.

(2) $K_i(A) \cap K_i(B) = K_i(A \cap B)$. Knowing A and knowing B is the same thing as knowing A and B.

(3) $K_i(A)$ contained in A. If i knows A, then A is true.

(4) $K_iK_i(A) = K_i(A)$. If i knows A, then he knows that he knows A.

(5) $-K_i(A) = K_i(-K_i(A))$. If i does not know A, then he knows that he does not know A.

Kripke (1963) called any system of knowledge satisfying the above five axioms S5. We shall later encounter descriptions of knowledge which permit less rationality. In particular, the last axiom, which requires agents to be just as alert about things that do not happen as about things that do, is the most demanding. Dropping it has interesting consequences for economic theory, as we shall see in Section VII. The clever reader can convince himself that axiom 5 implies axiom 4.

The partition approach to knowledge is completely equivalent to the knowledge operator approach satisfying S5. Given a set Ω of states of the world and a knowledge operator K_i satisfying S5, we can define a unique partition of Ω that would generate K_i. Simply define P_i as the collection of minimal fixed point events of the operator K_i. That is, look for all self evident events $E = K_i(E)$ such that there is no self evident $A = K_i(A)$ strictly contained in E. By axiom 1 there is at least one fixed point, namely $E = \Omega$, and since Ω is finite there is at least one minimal fixed point. By axiom 2, minimal fixed point events are disjoint, for otherwise their intersection would be a fixed point contradicting the minimality. It also follows from axiom 2 that if $A \subset B$, then $K_i(A) \subset K_i(B)$. For $K_i(A) = K_i(A \cap B) = Ki(A) \cap K_i(B)$. Hence if A is the union $\cup E$ of fixed point events E, then $E = K_i(E) \subset K_i(A)$, so $A = \cup E \subset K_i(A)$. But from axiom 3, $K_i(A)$ is contained in A. Thus we conclude that the union of fixed point events is a fixed point event. From axiom 5, if E is a fixed point of K_i, then so is $-E$. Hence the collection P_i of minimal fixed point events of K_i is indeed a partition of Ω. We must now check that the partition P_i generates the knowledge operator K_i, that is we must check that for any event A, $K_i(A)$ is the union E of all the minimal fixed point events that are contained in A. But we have seen that $A \supset K_i(A) \supset K_i(E) = E$. By axiom 4, $K_iK_i(A) = K_i(A)$. If $K_i(A) \supsetneq E$ is a fixed point, and so must contain a minimal fixed point, contradicting the definition of E.

Let us reconsider the puzzle of the red and white hats that we discussed in Section 2. Let there be three children, so $N = 3$. A state of nature ω corresponds to a description of the color of each child's hat. We list the 8 states in the table below:

		STATES OF THE WORLD							
	1	R	R	R	R	W	W	W	W
PLAYER	2	R	R	W	W	R	R	W	W
	3	R	W	R	W	R	W	R	W

Note that $\Omega = \{a, b, c, d, e, f, g, h\}$. The partitions of the agents are given by:

$$P_1 = \{\{a, e\}, \{b, f\}, \{c, g\}, \{d, h\}\}$$

$$P_2 = \{\{a, c\}, \{b, d\}, \{e, g\}, \{f, h\}\}$$

$$P_3 = \{\{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}\}.$$

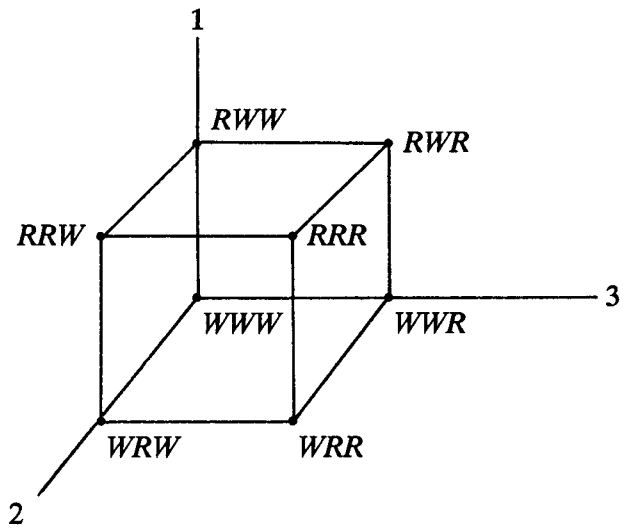
Observe that these partitions give a faithful representation of what the agents would know at the outset, no matter what the true colors of the hats. For example, if the true state of the world is $\omega = a = RRR$, i.e. all red hats, then agent 1 is informed of $P_1(a) = \{a, e\}$, agent 2 knows $P_2(a) = \{a, c\}$, and agent 3 knows $P_3(a) = \{a, b\}$. Agent 1 thus knows that the true state is either $a = RRR$, or $e = WRR$. Thus he cannot imagine that the other two children don't have red hats on, i.e. he knows their hats are red. Agent 1 however cannot be sure of his own hat color, since under one possible state of the world it is red, and under another it is white. Agent i knows his hat color only if at all ω which he regards as possible, his hat color is the same.

Were the actual state something else, say $\omega = b$, then $P_1(b) = \{b, f\}$, and again agent 1 would know the other children's colors, but not his own. The reader can check that the same is true for the other agents as well, no matter what the state.

We are in a position to use our model of knowledge to explain the puzzle of the hats. We shall see that after the teacher's announcement, every time a child announces that he does or does not know his own hat color, he is revealing information, namely that he was *not* informed of certain of the cells in his information partition, and hence that the states of nature contained in those cells cannot be the true state of nature. Before the teacher's announcement, the children could not reveal any information by admitting they did not know their own hat color. To help motivate the formal definition of common knowledge that will be presented later, we use the expression "common knowledge" in an informal way in the following discussion.

We can represent the state space more clearly as the vertices of a cube.[3] Think of R as 1 and W as 0. Then every corner of the 3 dimensional cube has three coordinates which are either 1 or 0. The ith coordinate denotes the hat color of the ith agent, $i = 1, 2, 3$. For each agent i we connect two vertices with an edge if they lie in the same information cell in agent i's partition. These edges should be denoted by different colors to distinguish the agents, but no confusion should result even if all the edges are given by the same color. The edges corresponding to agent i are all parallel to the ith axis.

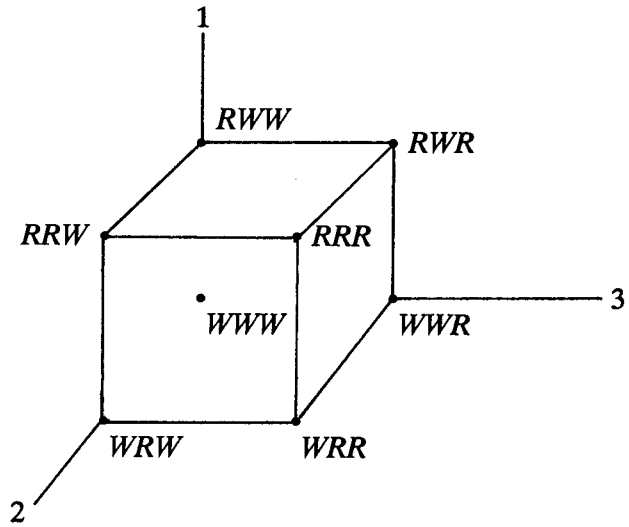
[3]This has been pointed out in public talk by Halpern.



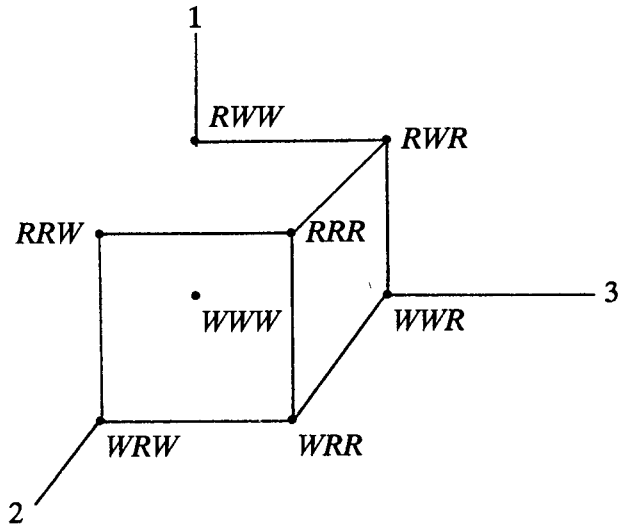
Note that every edge of the cube is present in the diagram. In particular every two vertices are connected by at least one path. Consider for example the state *RRR* and the state *WWW*. At state *RRR*, Agent 1 thinks *WRR* is possible. But at *WRR*, agent 2 thinks *WWR* is possible. And at *WRR* agent 3 thinks *WWW* is possible. In short, at *RRR* agent 1 thinks that agent 2 thinks that agent 3 thinks that WWW is possible. This chain of thinking is indicated in the diagram by the directed path marked by arrows.

An agent i knows his hat color at a state if and only if the state is not connected by one of i's edges to another state in which i has a different hat color. In the original situation sketched above, there is no state at which any agent knows his hat color.

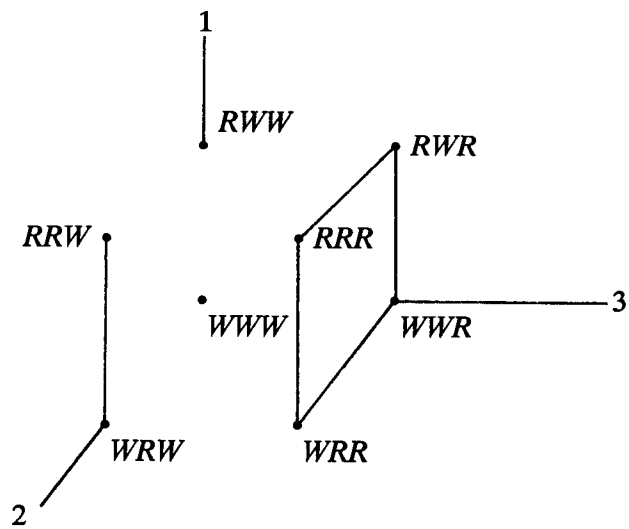
We now describe the evolution of knowledge resulting from the teacher's announcement and the responses of the children. The analysis proceeds independent of the actual state, since it describes what the children would know at every time period for each state of the world. Consider the situation after the teacher has announced that there is at least one red hat in the room. That announcement is tantamount to declaring that the actual state is not *WWW*. The clever students realize that if the teacher says nothing, then indeed the state is *WWW*. Every agent's partition has been augmented by the knowledge whether or not *WWW* is the true state of nature. This can be captured pictorially simply by dropping all the edges leading out of the state *WWW*, as seen below. There are now two components to the graph: the one consisting of the state *WWW* on its own, and the rest of the states.



If, after hearing the teacher's announcement, the first student announces he does not know his hat color, he confirms that the state could not be *WWW* (for then he would know the state precisely, and therefore his hat color), but he also reveals that the state could not be *RWW*, since if it were he would also be able to deduce the state from his own information and the teacher's announcement and therefore would have known his hat color. On the other hand, had he given the only other possible answer and said that he did know his hat color, then he would have revealed that the state was indeed *WWW* or *RWW*, since those are the only states at which he would have known his hat color. We can capture the effect of the first student's announcement on every agent's information by severing all the connections between the set {*WWW, RWW*} and its complement. There are now three different components to the graph.



The announcement by agent 2 that he still does not know his hat color reveals that the state cannot be any of {*WWW, RWW, RRW, WRW*}, since these are the states in which the above diagram indicates agent 2 would have the information (acquired in deductions from the teacher's announcement and the first student's announcement) to unambiguously know his hat color. Conversely, if 2 knows his hat color, then he reveals that the state must be among those in {*WWW, RWW, RRW, WRW*}. We represent the consequences of student 2's announcement on the other agents' information partitions by severing all connections between the set {*WWW, RWW, RRW, WRW*} and its complement, producing the final diagram. Notice now that the diagram has four separate components.



In this final situation agent 3 knows hit hat color at all the states. After hearing the teacher's announcement, and each of student 1 and student 2's announcements, it would have been impossible for him to have said no. Thus no more information is revealed.

If after student 3 says yes, student 1 is asked again what is the color of his hat, he will give the same answer he did the first time. So will student 2. The answers will repeat indefinitely as the question for students 1 and 2 and 3 is repeated over and over. Eventually every student will know what every other student is going to say, and each student will know that each other student knows what each student is going to say etc. Their responses will be common knowledge. By logic alone the students come to a common understanding of what must happen in the future. Note that after the teacher made the first announcement, and similarly after any student speaks, what was said is common knowledge since everybody heard it, and everybody knows that everybody heard it etc. We shall make precise a bit later the sense in which an event that has already occurred is common knowledge, in contrast with the sense in which an event which is yet to occur can also be regarded as common knowledge.

Assume now that the true state is *RRR*, as in Littlewood's story. The answers are no, no, yes. The formal treatment of Littlewood's puzzle has confirmed his heuristic analysis. But it has also led to some further results which were not immediately obvious. The analysis shows for example that for any initial hat colors (such as RWR) that involve a red hat for student 3, the same no, no, yes sequence will repeat indefinitely. For initial hat colors *RRW* or *WRW*,

the responses will be no,yes,yes repeated indefinitely. Finally, if the state is either *WWW* or *RWW*, then after the teacher speaks every child will be able to identify the color of his hat.

We have seen that no matter what the true state of nature, after the teacher identifies whether or not *WWW* is true, eventually one student will be able to deduce his hat color if they answer consecutively. We shall see in Section IV after we have introduced some more general principles of reasoning about knowledge that one student must eventually realize his hat color no matter which state the teacher begins by confirming or denying, and no matter how many students there are, and no matter what order they answer in, including possibly answering simultaneously or in groups etc. To prove this claim will not require any of the nitty gritty calculations that we have just made.

We can model the second puzzle about the envelopes along similar lines. In that story we can take Ω to be the set of ordered pairs (m, n) with m and n integers between 1 and 7, and $|m-n| = 1$. At state (m, n), agent 1 has 10^m dollars in his envelope, and agent 2 has 10^n dollars in his envelope. The knowledge of agent 1 can be described by the partition

$$P_1 = \{\{(n, n-1), (n, n+1)\} \text{ for all integers } 1 \le n \le 7\}.$$

Similarly

$$P_2 = \{\{(m-1, m), (m+1, m)\} \text{ for all integers } 1 \le m \le 7\}.$$

Agent i can always tell his own number, but never the other player's.

We graph the state space and partitions for this example below. The dots correspond to states with coordinates giving the numbers of agent 1 and 2, respectively. Agent 1 cannot distinguish states lying in the same row, and agent 2 cannot distinguish states lying in the same column.

	1	2	3	4	5	6	7
1		•					
2	•		•				
3		•		•			
4			•		•		
5				•		•	
6					•		•
7						•	

There are two components to the state space. In one component agent 1 has an odd number and 2 has an even number, and this is common knowledge, i.e. 1 knows it and 2 knows it and 1 knows that 2 knows it etc. In the other component, the parity is reversed, and again that is common knowledge. At states (1, 2) and (7, 6) agent 1 knows the state, and in states (2, 1) and (6, 7) 2 knows the state. In every state in which an agent i does not know the state for sure, he can narrow down the possibilities to two states. Both players have the same prior probability in mind for every state -- all states are equally likely. Thus each son quite rightly calculates that it is preferable to switch envelopes when first approached by his father. The sons began from a symmetric position, but they each have an incentive to take opposite sides of a bet because they have different information.

When their father asks them to shake hands, however, the situation changes completely. Now they must each take into account that the other one is also accepting the bet. Neither son would bet if he had the maximum $10 million in his envelope. Hence the instant after the two sons look at each other and neither jumps to renege on the bet, it becomes common knowledge that neither number is 7. The state space is now broken into four pieces, with the end states (6, 7) and (7, 6) each on their own. But a moment later neither son would allow the bet to stand if he had $1 million in his envelope, since either his bet would be taken, in which case he would get only $100,000, or else his bet would be rejected if the other son had

$10 million in his envelope. Hence if the bet still stands after the second instant, both sons conclude that the state did not involve a 6, and the state space is broken into two more pieces; now (5, 6) and (6, 5) also stand on their own. If after one more instant the bet is still not rejected by one of the sons, they both conclude that neither has $100,000 in his envelope. But at this moment the son with $10,000 in his envelope recognizes that he must lose, and he voids the bet.

If the sons could not look at each other, then *both* of them would write down that they did not want the bet. Each would realize that if there was a reason to bet, then the other son would also deduce that, and then exploit it. We shall explain this later.

Let us give one more example, that is reminiscent of the detective story. Suppose, following Aumann (1976) and Geanakoplos-Polemarchakis (1982), that two agents are discussing their opinions about the probability of some event, or more generally, of the expectation of a random variable. (If the random variable is the indicator function of the event, then the expectation of the random variable is the probability of the event). Suppose furthermore that the agents do not tell each other why they came to their conclusions, but only what their opinions are.

For example, let $\Omega = \{1, 2, ..., 9\}$, let both agents have identical priors which put uniform weight 1/9 on each state, and let $P_1 = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}$ and $P_2 = \{\{1, 2, 3, 4\}, \{5, 6, 7, 8\}, \{9\}\}$. Suppose that x takes on the values

1	2	3	4	5	6	7	8	9
17	-7	-7	-7	17	-7	-7	-7	17

We can represent the information of both agents in the following graph, where heavy lines connect states that agent 1 cannot distinguish, and dotted lines connect states that agent 2 cannot distinguish.

1 —.... 2 —.... 3 .... 4 — 5 —.... 6 .... 7 —.... 8 — 9

When agent 1 declares that he thinks that the expectation of x is 1, he reveals nothing, since no matter what the real state of the world, his information would have led him to the same conclusion. But when agent 2 responds with his opinion, he is indeed revealing information. If he says -1, then he reveals that the state must be between 1 and 8, whereas if he says

17 then he is revealing that the state of the world is 9. After his announcement, the partitions take the following form:

$$1 \overline{\ldots} 2 \overline{\ldots} 3 \ldots 4 \overline{\quad} 5 \overline{\ldots} 6 \ldots 7 \overline{\ldots} 8 \qquad 9$$

If agent 1 now gives his opinion again, he will reveal new information, even if he repeats the same number 1 he gave the last time. For 1 is the appropriate answer if the state is 1 through 6, but if the state were 7 or 8 he would say -7, and if the state were 9 he would say 17. Thus after 1's second announcement, the partitions take the following form:

$$1 \overline{\ldots} 2 \overline{\ldots} 3 \ldots 4 \overline{\quad} 5 \overline{\ldots} 6 \qquad 7 \overline{\ldots} 8 \qquad 9$$

If agent 2 now gives his opinion again he will also reveal more information, even if he repeats the same opinion of -1 that he gave the first time. Similarly if 1 responds a third time, he will yet again reveal more information, even if opinion is the same as it was the first two times he spoke. The evolution of the partitions after 2 speaks a second time, and 1 speaks a third time are given below:

$$1 \overline{\ldots} 2 \overline{\ldots} 3 \qquad 4 \qquad 5 \overline{\ldots} 6 \qquad 7 \overline{\ldots} 8 \qquad 9$$

Finally there is no more information to be revealed. But notice that 2 must now have the same opinion as 1! If the actual state of nature is $\omega = 1$, then the responses of agents 1 and 2 would have been (1, -1), (1, -1), (1,1).

As pointed out in Geanakoplos-Sebenius (1983), if instead of giving their opinions of the expectation of x, the agents instead were called upon to agree to bet or not (where $x(\omega)$ represents the amount of money that 2 must transfer to 1), or more precisely, they were asked only if the expectation of x is positive or negative, exactly the same information would have been revealed, and at the same speed. In the end the agents would have agreed on whether the expectation of x is positive or negative, just as in the envelopes problem. This convergence is a general phenomenon. In general, however, the announcements of the precise value of the expectation of a random variable conveys much more information than the announcement of its sign, and so the two processes are quite different, though they both result in a kind of agreement. When three or more agents are involved, the betting process and the opinion exchange look superficially still more different, as we shall see later.

So far in our formal treatment of knowledge we have dealt with the individuals separately, leaving the interaction of their knowledge to informal discussion. We now become more precise about the interactive epistemology, gradually building up to the notion of common knowledge.

3. Reasoning about the Reasoning of Others

Although we have used the concept in our three examples, so far our *formal* models lack a crucial ingredient: they do not specify what agent i thinks about what j thinks. We now describe how this is defined at each state $\omega \in \Omega$.

Let there be I agents, $i = 1, ..., I$; now I is finite. Each agent's knowledge is given by a possibility correspondence P_i, all acting on the same underlying state space Ω. For any $A \subset \Omega$, we can write $P_i(A) \equiv \bigcup_{\omega \in A} P_i(\omega)$. $P_i(A)$ is the set of all states i would think possible, if the true state varied over A. Thus E is a self-evident event for i if and only if $P_i(E) \subset E$.

Suppose j knows i's partition. At ω, j thinks any state in $P_j(\omega)$ is possible. Hence j cannot rule out any state in $P_i(P_j(\omega))$ as being one that i thinks is possible. We have already introduced the notion that i knows an event A at ω if and only if $P_i(\omega) \subset A$. If $P_i(P_j(\omega)) \subset A$, then we shall say that j knows that i knows A, if ω is the actual state of nature. This is sensible, since $P_i(P_j(\omega)) \subset A$ means that at every state that j can imagine that i can imagine, A occurs.

We can express these relations in an equivalent way. Let us write i knows A at ω by $\omega \in K_iA$. Thus we have defined a "knowledge operator on events" K_i by $K_iA \equiv \{\omega \in \Omega \,|P_i(\omega) \subset A\}$. Given any event $A \subset \Omega$, we shall write that i knows that j knows that A occurs at ω by $\omega \in K_iK_jA \equiv K_i(K_j(A))$. Notice that since K_j maps events $A \subset \Omega$ into events $K_j(A) \subset \Omega$, and therefore in the domain of K_i, the operator K_iK_j is well-defined. Moreover, $\omega \in K_iK_jA \Longleftrightarrow P_jP_i(\omega) \subset A$. Similarly we can write $K_iK_jK_mK_jA$ to mean that i knows that j knows that m knows that j knows that A.

In the hats example, before the teacher's announcement, does agent 3 know that agent 2 knows that agent 1 knows that there is at least one red hat at $\omega = a = RRR$, at least according to our model of knowledge? The set A where there is at least one red hat is $A = \{a, b, c, d, e, f, g\}$. The set K_1A is the set of all ω' with $P_1(\omega') \subset A$, and the reader can check that $K_1A = \{a, b, c, e, f, g\}$. The set K_2K_1A is the set of all ω' with $P_2(\omega') \subset K_1A$. This

is $\{a, c, e, g\}$. The set $K_3K_2K_1A$ is the collection of all ω' with $P_3(\omega') \subset K_2K_1A$. The reader can check that $K_3K_2K_1A = \phi$, so $a \notin K_3K_2K_1A$. In summary, though everybody knew what the teacher said, and they all knew that they all knew, they did not know that they knew that they knew!

III. COMMON KNOWLEDGE

1. Common Knowledge of Events

Having given an interpretation to the operators $K_{i_1} \cdots K_{i_n}$ we are now in position to introduce our most important concept.

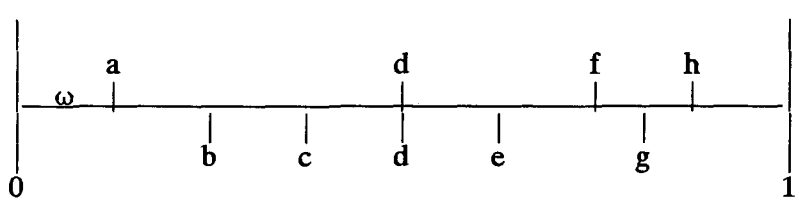
DEFINITION: The event $A \subset \Omega$ is *common knowledge* among agents $i = 1, ..., I$ at ω iff for any n and any sequence $(i_1, ..., i_n)$, $\omega \in K_{i_1}K_{i_2} \cdots K_{i_n}A$.

This definition of common knowledge was first introduced by the philosopher D. Lewis (1969), and first applied to economics by R. Aumann (1976). Note that there are an infinite number of conditions that must be checked to verify that A is common knowledge at ω. Yet when Ω is finite, Aumann (1976) has shown that there is an equivalent definition of common knowledge that is easy to verify in a finite number of steps.

2. A Characterization of Common Knowledge of an Event

THEOREM 1: *Let P_i, $i \in I$, be possibility correspondences representing the (partition) knowledge of individuals $i = 1, ..., I$ defined over a common state space Ω. Then the event A is common knowledge at ω if and only if $M(\omega) \subset A$, where $M(\omega)$ is the smallest event containing ω that is self-evident to every agent $i \in I$.*

We illustrate the theorem with a diagram



The whole interval [0, 1] represent Ω. The upper subintervals with endpoints $\{0, a, d, f, h, 1\}$ represent agent 1's partition. The lower subintervals with endpoints $\{0, b, c, d, e, g, 1\}$ repre-

sent agent 2's partition. At ω, 1 thinks $[0, a]$ is possible; 1 thinks 2 thinks $[0, b]$ is possible; 1 thinks 2 thinks 1 thinks $[0, d]$ is possible. But nobody need think outside $[0, d]$; note that $[0, d]$ is the smallest event containing ω that is self-evident to 1 and 2.

PROOF: Let $Q(\omega) \equiv \bigcup_n \bigcup_{i_1,\ldots,i_n} P_{i_1} \cdots P_{i_n}(\omega)$, where the union is taken over all strings $i_1, \ldots, i_n \in I$, of arbitrary length. Clearly A is common knowledge at ω if and only if $Q(\omega) \subset A$. But notice that for all $i \in I$, $P_i Q(\omega) \subset Q(\omega)$, so $Q(\omega)$ is self-evident for each i. QED

3. Common Knowledge of Actions

Let $f : \Omega \to A$ be a function of the state of nature. If the function is known, it is sensible to ask whether it is common knowledge at ω that f takes on the value a. We interpret this to mean that the event $E_a \equiv \{\omega' \in \Omega \mid f(\omega') = a\}$ is common knowledge at ω. From the last theorem we know that it is common knowledge at ω that f takes on the value a if and only if $f(\omega') = a$ for all $\omega' \in M(\omega)$.

A state of nature so far has described the prevailing physical situation, what everybody knows, what everybody knows about what everybody knows etc. We now allow each state to describe what everybody does. Each ω thus specifies an action $a_i = f_i(\omega)$ in A_i for each agent i in I. Since action results from individual choice we limit the function f_i by requiring that it not change if the information of agent i does not change:

$$[P_i(\omega) = P_i(\omega')] \text{ implies } [f_i(\omega) = f_i(\omega')]$$

Now let φ map nonempty subsets of Ω into some space A. We will interpret such a φ to be an action rule, because it specifies what element $a \in A$ to choose given any set of possibilities. Given a possibility correspondence P_i, φ induces an action function $f_i : \Omega \to A$ by $f_i(\omega) = \varphi(P_i(\omega))$. Thus given an action rule φ_1 for agent 1, among a group of agents $i = 1, \ldots, I$, with possibility correspondences P_i, $i = 1, \ldots, I$, we say that it is common knowledge at ω that φ_1 takes on the value a if and only if $\varphi_1(P_1(\omega')) = a$ for all $\omega' \in M(\omega)$.

4. Common Knowledge of Actions Negates Asymmetric Information about Events

In this section we state and prove our main theorem. The proof is trivial, but the conclusion is surprisingly strong.

We say that a decision rule $\varphi : 2^{\Omega}/\phi \to A$ satisfies the sure-thing-principle iff $\varphi(A) = \varphi(B) = a, A \cap B = \phi$ implies $\varphi(A \cup B) = a$.

The decision rules in our three examples all satisfy the sure thing principle. An agent who cannot tell his hat color if he is told only that the true state of nature is in A, and similarly if he is told it is in B, will definitely not know if he is told only that the true state is in $A \cup B$. Similarly if he could deduce from the fact that the state lies in A that his hat color is red, and if he could deduce the same thing from the knowledge that the state is in B, then he could also deduce this fact from the knowledge that the state is in $A \cup B$. (Note that we did not use the fact that A intersection B is empty). If the expectation of a random variable is equal to a (alternatively greater than zero) conditional on the state of nature lying in A, and similarly if the expectation of the same random variable is also a (alternatively greater than zero) conditional on the state lying in B, and if A and B are disjoint, then the expectation of the random variable conditional on $A \cup B$ is also a (alternatively greater than zero).

Now we can prove the following theorem showing that common knowledge of actions negates asymmetric information about events.[4] It shows that if agents follow action rules satisfying the sure-thing principle and if with asymmetric information the agents i are taking actions a_i, *then* if those actions are common knowledge, there is symmetric information that would lead to the same actions.

THEOREM 2: *Let $\varphi_i : 2^{\Omega}/\phi \to A_i$ be a decision rule for each i is satisfying the sure-thing-principle. If for each i it is common knowledge at ω that φ_i takes on the value a_i, then there is some event E such that $\varphi_i(E) = a_i$, for every i.*

COROLLARY: Under the conditions of the theorem, if $\varphi_i = \varphi$ for all i, then $a_i = a$ for all i.

---

[4]A special case of the theorem was proved by Aumann (1976), for the case where $\varphi_i = \varphi$ = the posterior probability of a fixed event A. The logic of Aumann's proof was extended by Cave [1983] to all "union consistent" decision rules. Bacharach (1985) identified union consistency with the sure-thing-principle. Both authors emphasized the agreement reached when $\varphi_i = \varphi$, and it is sometimes called the Agreement Theorem. However the aspect which I emphasize here (following Geanakoplos (1987)) is that even when the φ_i are different, and the actions are different, they can all be explained by the same information E.

PROOF: Let $E = M(\omega)$. Since it is common knowledge that φ_i takes on the value a_i at ω, $\varphi_i(P_i(\omega')) = a_i$ for all $\omega' \in E$. Since E is self-evident to each i, it is the disjoint union of cells in P_i on which i takes the same action a_i. Hence by the sure-thing-principle, $\varphi_i(E) = a_i$. QED

In the last diagram, if it is common knowledge that agent 1 is taking action a_1 at ω, then he must also be taking action a_1 at $[a, d]$. Hence by the sure-thing principle he would take action a_1 on $[0, d]$. Similarly if it is common knowledge at ω that agent 2 is taking action a_2 at ω, then not only does he do a_2 on $[0, b]$, but also on $[b, c]$ and $[c, d]$. Hence by the sure-thing-principle, he would have taken action a_2 had he been informed of $[0, d]$. Furthermore, if the action rules of the two agents are the same, then we must have $a_1 = a_2$.

Theorem 2 shows that it cannot be common knowledge that two players with common priors want to bet with each other, even though they have different information. Choosing to bet (which amounts to deciding that a random variable has positive expectation) satisfies the sure thing principle, as we saw previously. If players with common priors had the same information, they would not bet against each other. Hence by theorem 2 they will not bet if they have different information. This application is due to Milgrom-Stokey (1982). See also Sebenius-Geanakoplos (1983).

Similarly agents who have the same priors will not agree to disagree, as long as their opinions satisfy the sure thing principle. For if they had the same information, they would have the same opinion. Hence by Theorem 2 they must have the same opinion, even with different information. The agreeing to disagree theorem is due to Aumann (1976).

IV. LEARNING

So far we have taken a static approach to knowledge and action in our formal analysis. By the simple device of extending the notion of state of nature to take into account time, the same formal theory can also be used to understand learning.

1. Finer and Coarser Information

Consider two different possibility correspondences P_1 and P_2, corresponding to partitions $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively. We say that P_1 gives *finer* information at ω than P_2 if $P_1(\omega) \subset P_2(\omega)$.

P_1 leaves fewer possibilities to consider, and since $\omega \in P_1(\omega) \cap P_2(\omega)$, it brings the knower closer to the truth. We say that P_1 is *finer* than P_2 if P_1 is finer at all ω than P_2.

When P_2 is a partition, then P_1 is strictly finer than P_2 only if the cardinality of P_1 is strictly greater than the cardinality of P_2. The finest possible information is given by P_1 with $P_1(\omega) = \{\omega\}$ for all $\omega \in \Omega$. The coarsest possible information is given by P_2 with $P_2(\omega) = \Omega$ for all $\omega \in \Omega$. For arbitrary P_1 and P_2 it is often the case that neither P_1 nor P_2 refines the other. In none of our examples is one agent's knowledge finer than the other's: there are some things one agent knows better than the other, and other events the latter knows better.

2. The Join of Partitions

Consider an agent whose knowledge is given by the possibility correspondence P. He is presented with some new information. How can he incorporate that into his view of the world? Recall that an agent who is perfectly rational will infer as much from what is not said as he does from what is said. If another agent is due to say what color his hat is if he knows it, and the moment passes without him speaking, then he will infer that the agent did not know the color of his hat. Thus new information, if properly processed, always takes the form of a partition S of the state space. Recall that for any function $s : \Omega \to A$, s generates a partition S with cells equal to $s^{-1}(a)$ for all a in A.

We define the *join* of partitions P and S to be the partition J defined by

$$J(\omega) = P(\omega) \cap S(\omega), \text{ for all } \omega \in \Omega.$$

Clearly J is a partition if P and S are, and J is finer than either P or S. The reader can check that there is no other partition that is less fine than J and also finer than P and S.

3. A Dynamic State Space

Let T be a discrete set of consecutive integers, possibly infinite, denoting calendar dates. We shall now consider an expanded state space $\underline{\Omega} = \Omega \times T$. A state of nature ω in Ω prescribes what has happened, what is happening, and what will happen at every date t in T. An event $\underline{E}$ contained in $\underline{\Omega}$ now specifies what happens at various dates. The simplest events are called dated events and they take the form $\underline{E} = E \times \{t\}$ for some calendar time t, where E is contained in Ω.

Knowledge of agent i can be represented in the dynamic state space precisely as it was in the static state space as a partition $\underline{P}_i$ of $\underline{\Omega}$. We shall always suppose that agent i is aware of the time, ie we suppose that if (Ω', t') is in $\underline{P}_i(\omega, t)$, then $t' = t$. It follows that at each date t we can define a partition P_{it} of Ω corresponding to what agent i knows at date t about Ω, i.e. $P_{it}(\omega) = \{\omega' \text{ in } \Omega : (\omega', t) \text{ in } \underline{P}_i(\omega, t)\}$. The snapshot at time t is exactly analogous to the static model described earlier. Over time the agent's partition of Ω evolves.

In the dynamic state space we can formalize the idea that agent i knows at time t' about what will happen later at time t, perhaps by applying the laws of physics to the rotation of the planets for example. We say that at some (ω, t'), agent i knows that a (dated) event $\underline{E} = E \times \{t\}$ will occur at time $t > t'$ if $P_{it}(\omega) \subset E$. We say that it is common knowledge among a group of agents i in I at time t' that the event $\underline{E}$ occurs at time t iff $E = \{\omega : (\omega, t) \text{ in } \underline{E}\}$ is common knowledge with respect to the information partitions P_{it}'s, i in I.

4. A Dynamic Model of Knowledge and Action

We now describe how actions and knowledge co-evolve over time. Let $f_{it}(\omega)$ denote the action agent i takes at time t if the state of nature ω prevails. Let $s_{it}(f_{1t}, ..., f_{It}, \omega)$ be the signal that agent i receives as a result of the actions of all the agents, and the state of nature. Since each f_{it} is a function of ω, s_{it} is also a function of ω, and so it generates a partition S_{it} of Ω. At one extreme s_{it} might be a constant and the corresponding S_{it} consist of a single cell, Ω. This is the case if agent i does not observe any actions. At the other extreme s_{it} might be the whole vector $((f_{1t}, ..., f_{It}), \omega)$, which means that i is perfectly informed of all the actions of the players immediately after they occur, and of the state. (Incidentally, it may well be that s_{it} depends on the last term ω directly, without depending at all on the actions f_{it}. In that case agent i does not observe the actions of the other agents, but he does learn something about the state of the world.) An interesting intermediate case occurs if every agent whispers something to the person sitting on his right. Then $s_{it}(f_{1t}, ..., f_{It}, \omega) = f_{(i+1)t}$ (take $I+1 = 1$).

A consistent model of knowledge and action with perfect recall is described in the notation $(\Omega, T, I, P_i, f_i, s_i)$ satisfying for all t in T and i in I:

(1) $[P_{it}(\omega) = P_{it}(\omega')]$ implies $[f_{it}(\omega) = f_{it}(\omega')]$.

(2) $P_{it+1} = P_{it}$ join S_{it}, where S_{it} is the partition of Ω generated by $s_{it}(f_{1t}, ..., f_{It}, \omega)$.

Condition 1 says that agents can only take actions on the basis of what they know, so that if at time t they cannot distinguish states ω and ω', then they must take the same actions there. Condition 2 says that at each date an agent puts together everything he already knew at the previous date plus what he observed at the previous date.

5. Getting to Common Knowledge

THEOREM 3: *Let* $(\Omega, T, I, P_i, f_i, s_i)$ *be a consistent model of knowledge and action. Suppose* Ω *and* I *are finite and* T *is infinite. Then there is some* t^* *such that for all* $t, t' > t^*$, *it is common knowledge at* t' *that each agent knows what signals he will see at* t.

PROOF: As t increases, P_{it} becomes increasingly fine, hence the number of cells in each partition is a monotonic function of t. Since this number is bounded above by the cardinality of Ω, there must be some date t^* after which the number of cells in every partition is constant. From this moment on it is common knowledge that no agent will learn anything more. But that implies that each agent can predict with certainty every signal he will receive thereafter.

QED

Consider again the puzzle of the children and the hats. Suppose that from the beginning it is common knowledge that each agent can see the color of the other children's hats, and that it is common knowledge that at least one combination of hat colors is ruled out. (In the example we discussed earlier, *WWW* was ruled out. But any assignment will do. Let us call the assignment of hat colors that is known to be impossible *abc*.) Suppose that the agents each answer the question do you know the color of your own hat in some preassigned order. It may be that they alternate 1, 2, 3, 1, 2, 3, etc., or that they all speak at once, (123), (123), etc., or some combination of the two. We only need to assume that over the course of the infinite T, every agent i hears every agent i' an infinite number of times. Then it must be that at least one child eventually identifies the color of his hat, no matter what the true state *def*.

The proof of this claim is easy. From the convergence to common knowledge theorem, after some date t^* it is common knowledge what every child will say. Suppose that it is common knowledge that no child can identify his hat color. Then throughout the common knowledge component E defined by the partitions P_{it}^* containing *def*, no child can identify his hat color. Since agent 1 cannot identify his hat color at *def*, both *Ref* and *Wef* must be possible

for him. Hence *aef* is in E. Since agent 2 cannot identify his hat color throughout E, *aRf* and *aWf* are possible for him at *aef*. Hence they are in E. In particular *abf* is in E. But then agent 3 cannot identify his own hat color at *abf*, hence he must think both *abW* and *abR* are possible there, hence they are both in E. In particular *abc* is in E. But this is a contradiction, since it is common knowledge that *abc* is not in E. Thus it could not be that it was common knowledge that every child could not identify his hat color.

The same argument applies to the betting scenario. Suppose that at every date t each agent declares, on the basis of the information that he has then, whether he would like to bet, assuming that if he says yes the bet will take place (no matter what the other agents say). Then eventually one agent will say no. From the convergence to common knowledge theorem, at some date t^* it becomes common knowledge what all the agents are going to say. From the common knowledge of actions negates asymmetric information theorem, at that point they would do the same thing with symmetric information, provided it were chosen properly. But no choice of symmetric information can get agents to bet against each other, if they have the same priors. (This is the argument given in Sebenius-Geanakoplos (1983).)

An analogous argument given in Geanakoplos-Polemarchakis (1982), covers the case where agents with the same priors announce their beliefs about the probability of some event, or more generally, of the conditional expectation of a random variable. Eventually it is common knowledge what they will say, and, then there must be some common information which would get each of them to say what he had been saying. With the common prior assumption, that means all of them are saying the same thing. It is worth noting that although agents with common priors cannot agree to disagree, the number they ultimately agree upon may not be the same one they would have chosen if they had been allowed to share the reasons for their opinions, as opposed to simply their opinions. To see this, consider a state space $\Omega = \{1, 2, 3, 4\}$, and suppose agent 1 can see odd or even, while agent 2 has a partition $\{\{1, 2\}, \{3, 4\}\}$. Both agents have prior belief of 1/4 on every state. If asked to share their opinion about the probability of the event $\{1, 4\}$, both agents would say 1/2 (no matter what the real state of nature). Their opinions are therefore common knowledge, and indeed they are the same. However, if the agents shared their reasons, ie revealed which cells of their partitions the true state was in, then they would also agree, but on either 1 (if the true state was either 1 or 4), or else 0 (if the true state was 2 or 3).

6. Generalizations of Agreeing to Disagree

The conclusion that agents with common priors cannot agree to disagree can be reached by a somewhat different route. After agent i announces his opinion f_{it} at time t giving his conditional expectation of x, each agent j who hears that opinion will have a new partition P_{jt+1} which is finer than the partition F_{it} generated by the announcement f_{it}. In particular, each cell C in F_{it} is not only the disjoint union of cells in P_{it}, but also the disjoint union of cells in P_{jt+1}. Recall that each cell C in F_{it} consists of all states of the world at which i has the same opinion c about the conditional expectation of x. The key observation is that the average of the opinions that j announces at $t+1$ on those cells D whose disjoint union is some cell C in F_{it} (weighted by the *a priori* probability of each D) is simply the expectation of x conditional on C, which in turn must equal the opinion c that i announced at every ω in C. The expectation of j's opinion, given that i's opinion is c and assuming that j has heard i's opinion, must also be c.

In mathematical language, we have just shown that the sequential announcement of agents' conditional expectations of x is a martingale, provided that the agent making the announcement at time $t+1$ has heard the announcement made at time t. (Recall that a martingale is a sequence of random variables y_t, ie a sequence of functions $y_t : \Omega \to R$, defined on a state space Ω with some probability p on Ω, such that the expectation of y_{t+1} given the realization of y_t is the realization of y_t). A famous theorem of probability theory shows that martingales converge, i.e. for p-almost every ω, the sequence $y_1(\omega), y_2(\omega), \ldots$ converges. But since the sequence of opinions includes all the different agents' opinions, the convergence of this sequence is exactly what we wanted to show.

The martingale proof that agents cannot agree to disagree is actually much stronger than our previous proof, since it uses fewer assumptions. Note that it is only necessary that one agent hear every announcement (and that every agent speaks and is spoken to an infinite number of times if there are an infinite number of time periods). For example, following Parikh-Krasucki (1989), consider N agents sitting in a circle. Let each agent whisper his opinion in turn to the agent on his right. By our getting to common knowledge theorem, after going around the circle enough times, it will become common knowledge that each agent knows the opinion of the agent to his immediate left, assuming as usual that the state space is finite. It seems quite possible, however, that an agent might not know the opinion of some-

body two places to his left, or indeed of the agent on his right to whom he does all his speaking but from whom he hears absolutely nothing. Yet we have just shown that all the opinions must eventually be the same, and hence that every agent does in fact know everybody else's opinion.

Following McKelvey-Page (1987), suppose that instead of whispering his opinion to the agent on his right, each agent whispers his opinion to a poll-taker who waits to hear from everybody and then publicly reveals the average opinion of the N agents. (Assume as before that all the agents have the same prior over Ω). After hearing this pollster's announcement, the agents think some more and once again whisper their opinions to the pollster who again announces the average opinion etc. From the convergence to common knowledge theorem, if the state space is finite, then eventually it will be common knowledge what the average opinion is even before the pollster announces it. But it is not obvious that any agent i will know what the opinion of any other agent j is, much less that they should be equal. But in fact it can be shown that everyone must eventually agree with the pollster, and so the opinions are eventually common knowledge and equal.

We can see why by reviewing the proof given in Nielson-Brandenburger-Geanakoplos-McKelvey-Page (1990). Observe first that the opinion of any agent is itself a function on Ω and if it is not a constant it must be positively correlated with x. If it is constant, then of course it has 0 correlation with x. Moreover, exactly the same claim of positive or zero correlation between an agent's opinion and x itself can be made if attention is restricted to any set C which is the union of information cells of the agent's partition. Suppose now that at a time t after the average opinion is common knowledge, it takes on the value c on some set C contained in Ω. Restricted to C, the average opinion is not correlated with x since it is a constant there. But because each agent knows the average opinion and uses that information to refine his own partition, each agent's partition after date t contains cells whose disjoint union is C. Hence if any one of those agent's opinions is not constant on all of C, it must be positively correlated with x on C. But in that case the average opinion, which after all is just a scalar multiplied by the sum of opinions, each of which has either 0 correlation or positive correlation with x (and which includes at least one opinion with positive correlation with x) must have positive correlation with x when restricted to C. But this contradicts the fact that the average opinion is a constant on C. Hence each agent's opinion must be a constant on all of

C. But with common information, all agents must have the same opinion, so they must all have the same opinion *c*.

7. Infinite State Space

The hypothesis that the state space is finite, even though time is possibly infinite, is very strong, and often not justified. But without that hypothesis, the getting to common knowledge theorem is clearly false. Consider for example the famous coordinated attack problem, often mentioned in the distributed computing literature. Two generals plan to attack a town, either at dawn or at dusk. The circumstances may be such that a dawn attack is better, or it may be slightly better to attack at dusk. What is essential, however is that the generals coordinate their attack, for an attack by one army alone would be crushed. Unfortunately the generals are on opposite sides of the city, and can only communicate by messenger, and only the first general has the information telling whether dawn or dusk is the more propitious in which to attack. Suppose the first general sends a messenger at noon the day before to the second general saying when to attack. If the messenger is sure to arrive within 1 hour, then it is common knowledge at noon that it will be common knowledge at 1 PM when to attack the next day.

But now suppose that the messenger dies with probability p before arriving at the second general's camp. Then even after the messenger successfully completes his trip and conveys his information, it is not common knowledge when the right time to attack is. The first general knows when to attack, because he sent the message. The second general knows when the right time to attack is, because he got the message, but the first general does not know that the second general knows, because he cannot be sure that the messenger actually arrived. Suppose now that the second general sends the messenger back with an acknowledgement that he got the message. If the return trip is completely safe, and bound to be completed in under an hour, then by 2 PM it is common knowledge when to attack. But suppose that the return trip is just as dangerous as the original trip. Then even after the messenger arrives back in the first general's camp with the acknowledgement from the second general, it is still not common knowledge when to attack. The first general knows, the second general knows, the first general knows that the second general knows, but the second general does not know that the first general knows that he knows. It is easy to see that even if the state space were

allowed to be infinite, and the generals could send the messenger back and forth indefinitely, with probability one he would die in finite time and it would still not be common knowledge when to attack. More to the point, even if time is finite, to model the situation with a faulty messenger, the number of states must increase at least as fast as the number of time periods, for with each extra time period we must add one more state corresponding to the number of successful trips the messenger makes before he gets shot. One wonders whether common knowledge, as opposed to approximate common knowledge, makes any difference to how the generals would actually behave. This is an interesting question to which we shall return later. We shall also reexamine the question whether agents can bet against each other, and whether they can agree to disagree, if the state space is infinite.

For now let us consider the agreement result that if agents with the same priors announce their conditional expectations of some random variable back and forth, then eventually they will agree. With an infinite state space this can no longer be true; indeed, for any t and any probabilities p and q it is easy to construct an example in which the two agents repeat p and q back and forth to each other t times before finally agreeing on some probability r. However, as we just saw, the iterated announcement of expectations of x is a martingale, and so long as x is bounded above and below the martingale convergence theorem assures us that the opinions will converge to each other with probability one, even if the state space is infinite. In particular, if the opinions are common knowledge, and hence do not change over time, then convergence requires that they must be equal. Thus an infinite state space does not much affect the results on agreeing to disagree.

In the betting example with the two envelopes, we could imagine an infinite state space in which the father first picks any positive integer n with probability $1/2^n$, puts 10^n dollars in that envelope, and $10^{(n+1)}$ dollars in the other envelope and randomly gives them to his two sons. If one of the sons looks inside his envelope and finds 10^n dollars, $n > 1$, then he should calculate that the chances are two thirds that the other envelope has $10^{(n-1)}$ dollars, and one third that the other envelope contains $10^{(n+1)}$ dollars. No matter what the value of n, it would appear that the son would reckon it profitable to switch, and hence that he would always agree to it. This would seem to suggest that with an infinite number of states it could be common knowledge that agents want to bet against each other. However, as we shall see in a few sections down the road, the infinity of states has nothing to do with this problem. It is

the unboundedness of the payoffs that is the cause of the paradox. An infinite state space does not affect the no betting results.

V. BAYESIAN GAMES

1. Introduction

In this section we extend the description of a state of nature to include two more features. First, each state prescribes to each agent i a probability distribution over all possible states. Agent i not only has a view as to what is possible, but also a view of how likely each possibility is. In principle these beliefs could be arbitrary, but the Bayesian hypothesis is that each agent has a prior probability over all the states of nature, and that his beliefs at any one state ω are obtained from this prior and Bayesian updating from the additional information that the true state is in $P_i(\omega)$.

The Bayesian updating hypothesis is central to game theory and decision theory, and this is not the place to go into the philosophical arguments about its plausibility. (The interested reader should consult Savage (1954)). Note however that the agent who acts in a state ω has the beliefs prescribed by ω. It is not necessarily obvious that he knows his prior, ie that he can imagine what he would believe if he did not know what he does in fact know at ω. One philosophical explanation of the prior is that the agent did indeed begin with those beliefs, and that his partition is the result of information he has acquired since then. In that case he would naturally update his beliefs using Bayes Law. The importance of the Bayesian hypothesis, as will become clear shortly, is that it ties together decisions made at different information cells.

Second, we suppose that each state specifies the utility to every agent of choosing every possible action (not just the actions they do indeed choose at that state). If an agent knew the state ω, he could calculate his utility from choosing arbitrary action a by calculating the payoff in that state to him if he plays a and the others play the moves specified by ω. If he does not know the state, then he must calculate the payoffs to him of choosing a in state ω' while the others choose the actions specified by ω', multiplied by the probability of ω' believed at ω, summed over all such ω' that the agent believes have positive probability at ω.

If the action prescribed to agent i by every state ω that agent i assigns positive prior probability gives agent i higher expected utility than any other action (where the expectation is

computed using the beliefs specified by state ω and the utilities specified by ω and the other states ω' which i considers possible at ω), and if this holds for all agents i, then we say that we are in a Bayesian Nash equilibrium of a static Bayesian game. The notion of Bayesian Nash equilibrium is usually attributed to Harsanyi (1968), although most of the elements of its construction can be found in von Neumann and Morgenstern (1944).

If the agents must act over time, (so that in the language of game theory we might say that we are in an extensive form game), then each state (ω, t) must specify a complete strategy of what to do not only at t, but at all future dates as well, depending on what signals have been observed.

2. Examples

The concept of a Bayesian equilibrium is a bit subtle and calls for a formal definition, but before giving one let us describe some examples.

Consider the game called matching pennies with payoff matrix listed below.

	L	R
T	1, -1	-1, 1
B	-1, 1	1, -1

We know that there is a unique mixed strategy Nash equilibrium in which each player randomizes with equal probability over both of his strategies. This Nash equilibrium, like all others, is a special kind of Bayesian Nash equilibrium. Consider as the state space Ω the four entries in the above payoff matrix. Corresponding to each state there is indicated a pair of moves, one for each player. The first player has a partition of the state space consisting of the two rows of Ω. Notice that the moves he must make are (trivially) consistent with his partition. Similarly the second player has a partition of Ω given by the two columns of Ω. Once again the moves of player two are consistent with his partition. Both players have prior 1/4 on each state. Thus at any state, say the one in the top left hand corner, player one thinks the chances are 1/2 that the state is either one in the top row, and player two thinks the chances are even between either state in the left column. The payoffs to each player for any state are given by the same payoff matrix.

To check that we have indeed described a Bayesian Nash equilibrium, note that at any state, such as top left, each player is indeed satisfied to choose the move indicated by that state since his expected payoff is the same (namely 0) no matter which action he chooses. To player 1 for example, the expected payoff to going T is $(1/2)1 + (1/2)(-1) = 0$, where the sum indicates that player 1 does not know which state he is in, and hence given the uniform prior, must assign equal probability to the two states he thinks are possible, and the consequences arising in each from his choice of T. Similarly he can calculate his expected payoff to choosing B as $(1/2)(-1) + (1/2)1 = 0$.

The Bayesian Nash equilibrium gives a slightly different interpretation to behavior from the usual mixed strategy Nash equilibrium. In a mixed strategy Nash equilibrium each player is flipping a coin to decide how to play. In Bayesian Nash equilibrium, there is one actual state. Thus each player is making a unique choice of (pure) move, namely the one assigned by that state. But the other player does not know which move that is, so to him the choice seems random. This reinterpretation of mixed strategy Nash equilibrium in terms of Bayesian Nash equilibrium is due to Ambruster-Boge (1980).

In the above example the set of moves available to each player, and the payoffs to the players of those choices, does not depend on the state of nature. Moreover, for every player i, the conditional distribution of moves by i's opponent is the same at each state. When the action spaces and payoffs are independent of the state (thus defining a unique game G) but the conditional distribution of opponent's actions is allowed to depend on the state, then Bayesian Nash equilibrium reduces to what has been called a correlated equilibrium of G. The notion of correlated equilibrium was invented by Aumann in 1974. An elementary but important example of a correlated equilibrium is a traffic light. In every state, the choices are the same: stop or go. If both cars go, it is disastrous. If both stop it is OK for both. If one stops and the other goes, it is good for the one who goes, and OK for the one who stops. Each player has a prescribed action given the state, and the state is random. But if one player sees a green light, his distribution over the moves of the other player is very different from what it would be if he saw a red light.

In general the action spaces and the payoffs will depend on the state of nature. Recall the problem of the father and the two sons who were each left money in envelopes. The state space and the partitions have already been defined. Moreover we have already specified a

prior probability for each son, namely that every state is equally likely (since there are 12 states, that means each has probability 1/12). In the static game in which the sons must write down once and for all what they want to do, the action spaces of the two sons are the same at every state and consist of two elements, namely to bet (call it B) or not to bet (call it N). The payoffs must now be described at each state $\omega = (m, n)$, as a function of the 4 possible combinations of action choices. Under the rules given, if both choose N, then each keeps the amount of money in his own envelope (so that the payoff to 1 is 10^m, and the payoff to 2 is 10^n). If both choose B, then the envelopes are switched, but each must pay \$1, so that the payoff to 1 is 10^n-1, and the payoff to 2 is 10^m-1. If 1 chooses B, and 2 chooses N, then the envelopes are not switched, but 1 loses a dollar, and if 1 chooses N and 2 chooses B, then again the envelopes are not switched but 2 loses 1 dollar. The following matrix summarizes these payoffs at state (m, n).

	B	N
B	$10^n-1,\ 10^m-1$	$10^m-1,\ 10^n$
N	$10^m-1,\ 10^n-1$	$10^m,\ 10^n$

To complete the description of the Bayesian equilibrium we must specify the actions that the agents take at each state. Let us assign each agent the action N at each state. Then it is clear that we have a Bayesian equilibrium. There is no state at which either agent can profitably switch to B, since doing so costs \$1 and does not result in the transfer of the envelopes. Notice that in this Bayesian equilibrium it is common knowledge at every ω that both sons are choosing not to bet. We have already seen that if it is common knowledge what the agents are doing, then they must not be betting. Thus it cannot be a Bayesian Nash equilibrium for each son to bet no matter what he sees in his envelope. (One can verify this directly. If son 1 adopts this decision rule, then son 2 can take advantage of him by betting no matter what he sees, unless it is the maximum amount, in which case he refuses to bet). But perhaps there is another Bayesian equilibrium in which the agents choose to bet some of the time but not all the time, so it is not common knowledge what they are doing. In the next section we show that there can be no such equilibrium either.

The example where I agents with fixed partitions write down their best guess of the conditional expectation of some random variable x can also be described as a Bayesian game. Let

the action space for each agent be the entire set of real numbers. Let the payoff to agent i if the state of nature is ω and he chooses the real number r be $-(x(\omega) - r)^2$. It is easy to see that the optimizing assignment of actions to states is to require that at each ω, agent i choose the action $r(\omega)$ which maximizes

$$\underset{r}{Max} \sum_{\omega' \in P(\omega)} - (x(\omega') - r)^2 \pi_i(\omega')$$

and this clearly works out to choosing $r(\omega) = \mathrm{Exp}[x|P_i(\omega)]$.

Let us reconsider the problem of the two generals trying to decide whether to attack at dawn or at dusk, and follow the analysis of Rubinstein (1989). How "close" to common knowledge do we need to get in order to coordinate the generals? The state space (when the messenger travels infinitely fast) is made up of all possible pairs (dawn, n) or (dusk, n), where the first component represents the best time to attack, and the second component n is any nonnegative integer representing the number of successful trips the messenger makes before being shot. The partitions of the two generals is given in the picture below, where each dotted line connects two states that general 1 cannot distinguish, and the dotted lines demark states that general two cannot distinguish.

```
(dawn, 0) —— (dawn, 1) --- (dawn, 2) —— (dawn, 3) --- (dawn, 4) —— (dawn, 5) ---
    |
    |
    |
(dusk, 0) —— (dusk, 1) --- (dusk, 2) —— (dusk, 3) --- (dusk, 4) —— (dusk, 5) ---
```

Let us suppose that the messenger makes exactly n successful trips, where n is even. General 1 cannot tell whether the messenger made n or $n+1$ successful trips, since all he knows at say (dawn, n) is that the messenger arrived back to him on his nth trip, but failed to show up after what would have been his $n+2^{nd}$ trip. General 2 cannot tell whether the messenger did not return on his $n+2^{nd}$ trip because he died on the way out (on his $n+1^{st}$ trip) or on the way back (on his $n+2^{nd}$ trip, after completing $n+1$ trips). General 1 always knows whether it is more propitious to attack at dawn or dusk. General 2 on the other hand does not know whether it is better to attack at dawn or at dusk without benefit of the messenger. But he does know once the messenger has made his first trip, i.e. if $n \geqq 1$. If $n = 0$, then the second general knows n exactly, since he does not get any message. But if n is even, and $n > 1$, then

the second general receives the messenger on his $n-1^{st}$ trip but he does not return on the $n+1^{st}$ trip, so general 2 cannot tell whether the messenger died after $n-1$ or n successful trips.

Recalling that the probability of the messenger getting shot on any trip is p, independent of how many trips he has made or what his message says, the prior on Ω for each general is the same, namely prob(dawn, n) = prob(dusk, n) = $(1/2)[(1-p)^{(n)}]p$. For concreteness, let p = .1. We can easily calculate the probability beliefs each general has at any state of the world. Let us choose for example (dawn, 4). The first general knows that there are only two possible states, and using Bayes law we calculate that the conditional probability of (dawn, 4) is $1/(1+(1-p))$ = 10/19, and the conditional probability of (dawn, 5) is $(1-p)/((1+(1-p))$ = 9/19. In short, if the first general does not see the messenger arrive on his 6th trip, he assumes he was more likely to have died on the way out after successful trip 4 than on the way back after successful trip 5 (since he would not even get the chance to come back if he had not made it out). Similarly if general 2 receives the messenger after his third successful trip, but does not see the messenger after trip 5, he assumes it is more likely that the messenger died on trip 4 than on trip 5, so general 2 assigns probability 10/19 to 3 successful trips, and probability 9/19 to 4 successful trips.

The action spaces of each of the two generals is to attack either at dawn (call it M for morning) or at dusk (call it N for night). The payoffs to states (dawn,n) and (dusk, n) depend on the time but not on when the messenger is shot.

	dawn	
	M	N
M	4,4	2,0
N	0,2	3,3

	dusk	
	M	N
M	3,3	2,0
N	0,2	4,4

Let us now try to assign actions for each general to the states to create a Bayesian Nash equilibrium. One Bayesian Nash equilibrium is for both generals always to choose M, no matter what their information. This results in a coordinated attack, and an ex ante expected payoff of 3.5, but the information that general 1 receives about when to attack is completely wasted, even at states where n is very large. There is also an equilibrium where each general always attacks at night, which similarly gives an expected payoff of 3.5.

Surprisingly, there is no Bayesian equilibrium in which the first general's information is ever used, even when he knows it, 2 knows it, 1 knows that 2 knows it, etc. To see this, suppose that general 1 always plays M when it is dawn. We shall show that he must always play M when it is dusk as well, in any Bayesian equilibrium. For example, suppose we assign him to move M for (dawn, n) for all n, and M for (dusk, n) if $n < K$, where K is even, and N for (dusk, n) where $n \geqq K$. Note that this is a feasible strategy for him, since he knows when the state is less than K for any even K. In words, the first general sticks with attacking in the morning unless the messenger announcing that dusk is a good time to attack makes at least K trips, in which case he switches to attacking at night. But if this is what 1 is doing in each state, then general 2 will choose to play M at (dawn, n) for all n, and will play M at (dusk, n) for $n < K+1$, and N at (dusk, n) for $n \geqq (K+1)$. The reason is that at $n = K$, general 2 cannot be sure that $n = K$ or $n = (K-1)$. Indeed he assigns probability 10/19 to $n = (K-1)$, and only probability 9/19 to $n = K$. By playing M he gets expected utility (10/19)3 + (9/19)2, while if he plays N he gets expected utility (10/19)0 + (9/19)4, which is smaller. But if general 2 is going to play M when $n = K$, then general 1 is also better off playing M when $n = K$, by a similar argument. Thus the assignment given above for 1's moves is not optimal, and we have reached a contradiction.

It is interesting to see how the game changes if we truncate the states at some finite K. For simplicity, take K even. Suppose that after K trips the messenger stops traveling. This is indeed quite realistic, for if he does not move infinitely fast, the day of the attack will eventually arrive and there will be no more time for messages. In that case general 1 knows it when $n = K$. Furthermore, when $n = K$ or $K-1$, general 2 assigns probability .9 to $n = K$, and probability .1 to $n = K-1$, for once he sees that the messenger has made $K-1$ successful trips he knows he will never see the messenger again, and chances are .9 that the messenger will make it back to general 1's headquarters. For $n < K-1$, everything is as before. The reader can check that there is now another Bayesian equilibrium in which general 1 plays M at every state of the form (dawn, n), and he plays M at every state of the form (dusk, n) for $n < K$, while at (dusk, K) he plays N. General 2 plays M at every state of the form (dawn, n) and he plays M for every state of the form (dusk, n) with $n < (K-1)$, but for states (dusk, $K-1$) and (dusk, K) he plays N. The upshot is that if the messenger succeeds in making the maximum number of trips, then there is coordination at the right time of day.

Finally, let us consider the famous repeated Prisoner's Dilemma. The two players have two possible moves at every state, and in each time period, called cooperate (C) and defect (D). The one-shot payoffs to these choices are given by

	C	D
C	5,5	0,6
D	6,0	1,1

Let us suppose that the game is repeated T times. A strategy for an agent consists of a designation at each t between 1 and T of which move to take, as a function of all the moves that were played in the past. One example of a strategy, called the grim strategy, is to defect at all times, no matter what. The tit for tat strategy is to play C at every t until the other player has defected, and then to play D for ever after. Other strategies typically involve more complicated history dependence in the choices. Since at any t there are $4^{(t-1)}$ histories, there are in total $s = 2(4^T - 1)/3$ possible (pure) strategies for each player. The number of strategy pairs is thus s^2. For each strategy pair, there is unique path of play that is recursively defined. For example, if tit for tat meets grim, then in the first period tit for tat plays C and grim plays D, and thereafter both play D. To each choice of strategy pair we can define a payoff to each player, namely the sum of the T one-shot payoffs specified by the T moves defined by the path of play corresponding to that strategy pair.

Let us take for Ω a set of s^2 states, corresponding to all strategy pairs. The set of states of the world is thus $\Omega \times T$. At each (ω, t) we can define the payoff to each pair of possible strategies by the sum of the T-t one shot payoffs that they specify in the time remaining. The actions are of course specified by ω in the obvious way. The set $P_i(\omega, t)$ consists of all (ω', t) such that player i's strategy is the same in ω' as in ω, and the history of the first $t-1$ realized outcomes is the same in ω' as it is in ω.

To complete the description of a Bayesian Nash equilibrium, we need only specify the prior probabilities of the agents on Ω. One simple equilibrium is the one in which each agent puts probability 1 on the state ω in which both players use the grim strategy. The path of play resulting from this state is (D, D) at each date, a dismal result. We now show that every Bayesian Nash equilibrium yields the same outcome.

Let us restrict attention to those Bayesian Nash equilibria in which the priors of the two players have the same support, that is each state ω has positive probability under both priors or neither prior. In that case both players' priors must put probability 1 on states in which the path of play consists entirely of (D, D). To see why, note that if ω has positive probability under 1's prior, then on the path of play defined by ω, 1 must be choosing D in the last period T, otherwise he could improve his payoff by switching to D for that last period. From the common support hypothesis, we deduce that ω also has positive probability under player 2's prior, hence by the same reasoning 2 must be playing D in period T on the path of play generated by ω. Take any time period t such that at all time periods s later than t, the path of play under all ω that have positive probability consists entirely of (D, D). (We have just shown that $t = T-1$ is such a t). Then each player must also be playing D on the path of play in period t at any ω with positive probability. For otherwise, by switching to D in period t the player improves his one shot payoff in period t and cannot make any subsequent payoff worse than it was before (if he plays D in every subsequent period), and also leaves the first $t-1$ periods unaffected. Hence by backward induction we conclude that (D, D) is played in every period t at any ω that has positive probability.

The upshot is that in Bayesian Nash equilibrium, repeated interaction does not work to improve cooperation. We shall come back to this point later.

Before leaving this example, let us note that sometimes it is possible to represent essentially the same situation with two different state spaces. For example, let us delete from the above Bayesian Nash equilibrium all states which have probability zero under both players' priors. Restrict the partitions in the obvious way, simply by leaving out the probability zero states. In the grim equilibrium, there is now only one state of nature. The effect of this compression is that agents now know much more than they did before, since there are so many possibilities they need not consider any more.

3. Bayesian Games: A Formal Definition

A (Bayesian) game is a vector $\Gamma = (I, \Omega, (P_i, \pi_i, A_i, u_i)_{i \in I})$ where $I = \{1, ..., I\}$ is the set of players, Ω is the set of states of the world, P_i is a partition of Ω, π_i is a prior probability on Ω, A_i is the set of possible actions for player i, and $u_i : A_i \times \Omega \times \mathbb{R}$, where $A = A_1 \times \cdots \times A_I$,

is the payoff to player i. For any product $Y = Y_1 \times \cdots \times Y_I$, the notation Y_{-i} means $Y_1 \times \cdots \times Y_{i-1} \times Y_{i+1} \times \cdots \times Y_I$.

A (Bayesian) Nash equilibrium for the game Γ is a vector $f = (f_1, ..., f_I)$ where $\forall i$, $f_i : \Omega \to A_i$ and

(1) $[P_i(\omega) = P_i(\omega')] \Rightarrow [f_i(\omega) = f_i(\omega')]$, $i = 1, ..., I$ and

(2) $\forall i, \forall a \in A_i, \forall \omega \in \Omega$,

$$\sum_{\omega' \in P_i(\omega)} u_i(f_i(\omega), f_{-i}(\omega'), \omega')\pi_i(\omega') \geq \sum_{\omega' \in P_i(\omega)} u_i(a, f_{-i}(\omega'), \omega')\pi_i(\omega') .$$

Condition (1) means that if player i cannot distinguish ω' from ω, then he must choose the same action $f_i(\omega) = f_i(\omega')$ in both states. Condition (2) means that at each ω agent i prefers the action $f_i(\omega)$ to any other action $a \in A_i$, given his information $P_i(\omega)$ and given the action rule f_{-i} of the other agents. Implicit in the definition is the idea that each player i knows the decision functions f_{-i} of all the other players. Without f_{-i} it would be impossible to define the payoff to agent i, since u_i depends on the choices of the other players, as well as the state and agent i's choice.

4. Ex Ante and Ex Post Optimality in Bayesian Games

In any state of nature ω each agent has a probability distribution over the set of states he considers possible. To distinguish this probability from the prior on all states he might have begun with (under the Bayesian hypothesis) we call this the ex post beliefs at ω. If that agent is optimizing at ω, then he is choosing the action which gives him the highest expected utility calculated with respect to the ex post probability, in short the highest ex post utility. The force of the Bayesian hypothesis is that given any decision function mapping states into actions, we can also calculate the ex ante expected utility to agent i by using his prior probability.

Furthermore, it is easy to see that if an action rule is ex post optimal at every state of the world, then there can be no other decision function which respects the knowledge of the agent (i.e., is constant inside each of his partition cells) and gives higher ex ante utility. To put it another way, if the agent had full understanding of the whole model, and was interested in

maximizing his ex ante utility, then faced with the informational constraints given by his partition, he could not do any better than to maximize his ex post utility.

It follows, assuming that the other agents do not change their behavior, that if agent i is given more information (and hence a finer partition), then optimizing his ex post utility will make his ex ante expected utility at least as great. At each cell of the finer partition, the agent could always choose the same action he took when he had the coarser partition. If he changes his action, it will improve his payoff on that cell without affecting it anywhere else. *Ex ante, more knowledge always helps a Bayesian decision-maker.*

Bayesian (ex post optimal) decisions also satisfy the sure-thing principle. The argument showing that more knowledge is ex ante helpful also confirms the sure-thing principle. From Theorem 1 we deduce the result that given any Bayesian Nash equilibrium in which the actions of the agents are common knowledge at some state ω, we can find another Bayesian Nash equilibrium with the same agents, state space, utilities, priors, and actions, but in which the partitions of the players have been altered in such a way that at ω all agents have the same information. We deduce that asymmetric information about events can never explain joint actions that are common knowledge in Bayesian Nash equilibrium.

In closing this section, let us note that although it may help the players in getting to Bayesian Nash equilibrium to understand the whole model, including the priors and the information partitions they have and the partitions and utilities the other players have, in fact the definition only requires that each agent optimize his own utility ex post. Therefore it is only necessary for him to understand how the moves of the other players depend on the states of nature. He need not know anything about what they know, or what they want.

5. Common Knowledge of Rationality and Optimization

In Bayesian Nash equilibrium it is common knowledge at every ω that all the players have partitions of the state space, that their posterior beliefs are derived from a prior, that they each know their own move, and that they are each optimizing. Indeed, together with the meaning of state of nature (which one might view as tautological), these properties characterize Bayesian Nash equilibrium. Aumann (1987) has summarized all these conditions under the rubric of rationality. Thus he proves that if it is also common knowledge what the payoffs of the players are to each move, then common knowledge of rationality is equivalent to being

in a correlated equilibrium (with common priors or with subjective priors). In section VI we shall weaken some of these "rationality" conditions in our definition of equilibrium, hence there it will be convenient to confine the term rationality to the representation of knowledge through partitions, so as to keep separate the question of whether agents always optimize, and so on.

VI. SPECULATION

The cause for financial speculation and also gambling has for a long time been put at differences of opinion. "It takes a difference of opinion to make a horse race" is an old refrain. Since the simplest explanation for differences of opinion is differences in information, it was only natural to conclude that gambling and speculation could be explained by differences in information. Take for example the question of whether a Republican will win the next Presidential election. Two agents who think exactly alike, that is, in the language of Bayesian analysis, have the same priors, very likely will disagree about the exact probability of a Republican victory because they have different information. Does that not imply that they can find odds at which to bet on the outcome?

We distinguish two kinds of speculation. One is simple betting, in which two agents face each other and agree on some contingent transfer of money. At the moment of the agreement they may shake hands or sign a contract, or give some other sign that the arrangement is common knowledge between them. The other kind of speculation occurs between many agents, say on the stock market or at a horse race or a gambling casino, where an agent may commit himself to risk his money before knowing what the odds may be (as at a horse race or roulette table) or whether anyone will take him up on the bet (as in submitting a buy order to a stockbroker). In the second kind of speculation what the agents are doing is not common knowledge.

When two people bet with each other, one wins what the other loses. Assuming that the outcome of the bet is independent of the players' original incomes, and that they are both risk averse, and have the same priors, the only explanation for betting would appear to be different information. Note that if they had different priors, then that could explain their betting. Of course we must remember that it is not just the expectation of winning, but the expectation of the marginal utility of winning that determines whether a bet is accepted. An agent may

put low probability on some event but still be reluctant to risk losing any money in that event if his marginal utility for income is higher when he is poor and his income is particularly low in that event.

When one agent buys a stock from another agent he puts himself in a position where he will gain if the stock price goes up and the seller will lose. This appears to be a bet. But another reason for trading the stock could be that the seller's marginal utility for money at the moment of the transaction is relatively high (because his children are starting college), whereas the buyer's marginal utility for money is relatively higher in the future when the stock is scheduled to pay dividends. In such a situation we would not say that the trade is for speculative reasons.

Speculation shall be understood to mean any sort of actions which are taken purely on account of differences of information. To capture the idea that actions are taken solely on account of differences of information, we suppose that there is a status quo action for each agent, which does not take any knowledge to implement, such that if every agent pursued that action in every state the resulting (ex ante) utilities would be Pareto optimal. As we have said, except in special cases ex ante Pareto optimality is neither implied by nor does it imply identical priors. Pareto optimality ex ante means that no other combination of state-contingent actions, no matter how much knowledge their implementation requires, can make every agent better off (ex ante). In an ex ante Pareto optimal situation with a status quo action, the only reason for any agent to choose something other than the status quo action is that he thinks he has a chance to gain at somebody else's expense. A typical kind of ex ante Pareto optimal situation might arise as follows. Risk averse agents, with possibly different priors, trade Arrow-Debreu state contingent claims for money, one agent promising to deliver in some states and receive money in others etc. Agents will of course arrange to get deliveries in those states where they were relatively poor to begin with, and which they consider relatively more likely. After the signing of all the contracts for delivery, but before the state has been revealed, a well-known theorem guarantees that if there is a sufficiently rich collection of assets, the agents are at an ex ante Pareto optimum. Now suppose that each agent receives additional information revealing something about which state will occur. If different agents get different information, so that they now have different beliefs about which is the likeliest state, should they trade again? The following theorem shows that the answer is no.

THEOREM 4 *(Non-speculation Theorem for Bayesian Nash Equilibrium): Let* $\Gamma = (I, \Omega, (P_i, \pi_i, A_i, u_i)_{i \in I})$ *be a Bayesian game. Suppose each player i has an action* z_i *such that for all* $(f_1, ..., f_I)$, $\sum_{\omega \in \Omega} u_i(z_i, f_{-i}(\omega), \omega)\pi_i(\omega) = \bar{u}_i$. *Furthermore, suppose that if for any decision functions* $(f_1, ..., f_I)$, $\sum_{\omega \in \Omega} u_i(f(\omega), \omega)\pi_i(\omega) \geq \bar{u}_i$ *for all i, then* $f_j(\omega) = z_j$ *for all* $\omega \in \Omega, j = 1, ..., I$. *Then* Γ *has a unique equilibrium,* $(f_1^*, ..., f_I^*)$ *and* $f_i^*(\omega) = z_i$ *for all* $i = 1, ..., I$, *and all* $\omega \in \Omega$.

PROOF: Let $(f_1, ..., f_I)$ be an equilibrium. Fix f_j for all $j \neq i$, and look at the one-person decision problem this induces for player i. Clearly f_i must be an optimal plan for this decision problem. But if i had the trivial partition $Q_i(\omega) = \Omega$ for all $\omega \in \Omega$, he would be able to guarantee himself *ex ante* utility at least $\bar{u}_i$ by always playing $g_i(\omega) = z_i$. Hence by the principle that knowledge cannot hurt $\sum_{\omega \in \Omega} u_i(f(\omega), \omega)\pi_i(\omega) \geq \bar{u}_i$. Since this is true for all i, by hypothesis $f_i(\omega) = z_i$ for each i and $\omega \in \Omega$. QED

As an application of Theorem 4, consider the problem of the two sons and the envelopes. If both sons are risk neutral, ie interested in maximizing their expected money payoff, then it is easy to see that Theorem 4 applies. Any strategy of betting (as opposed to choosing the status quo of not betting) does nothing to increase the total quantity of money, at best rearranging it, but also frittering away 1 or 2 dollars per state in which one or both of the sons bet. Thus anything but the status quo must lower the ex ante expected utility of at least one son, and so by theorem 3 the only Bayesian Nash equilibrium is not to bet.

It is interesting to observe that the proof of theorem 3 does not require that the state space be finite. Yet consider again the sons and the two envelopes, except that now suppose that there is an infinity of possible envelopes, with no upper bound to the amount of money. For instance, suppose the father chooses any positive integer n with probability $1/2^n$, puts $\$10^n$ in one envelope, and $10^{(n+1)}$ in the other envelope. No matter what amount of money a son sees, he will prefer the other envelope, and indeed it can well be common knowledge that the sons will want to bet with each other after opening their envelopes. The proof of Theorem 4 breaks down only because the ex ante expected payoff is infinite to each player, so it is impossible to check the Pareto optimality condition, since that means comparing one infinity against another. Had there been a finite expectation to the amount of money that could appear in any envelope, which still allows for the case where there are unbounded dollar amounts, each

one higher than its predecessor, then once again for sufficiently high n a son would not want to bet and no betting would be the only Bayesian Nash equilibrium.

In striking contrast to the dictum "Differences in information make a horse race", when it is common knowledge that agents are rational, differences of information not only fail to generate a reason for trade on their own, but even worse, they inhibit trade which would have taken place had there been symmetric information. Take for example the two sons with their envelopes, in the original version of the story where there was a maximum of $10 million in the envelopes. Suppose now, however, that the sons are risk averse, instead of risk neutral. That is suppose they act to maximize the expectation of $u(x)$, where x is the final holding of money and u is a strictly concave and increasing function. Then it is easy to see that before the sons open their envelopes each has an incentive to bet, not the whole amount of his envelope against the whole amount of the other envelope, since that just replaces one lottery with another equivalent lottery and loses $1 for the betting fee, but to bet half his envelope against half of the other envelope. In that way each son guarantees himself the average of the two envelopes, which is a utility improvement for sufficiently risk averse utility u despite the $1 transaction cost. Once each son opens his envelope, however, there is no incentive to trade, precisely because of the difference in information! Each son must ask himself what the other son knows that he doesn't.

We can place this trading problem in a more general context as follows. The ex ante marginal utility to an agent i of an extra dollar in some state ω is the probability of that state $\pi_i(\omega)$ multiplied by the marginal utility of money in that state, $du_i(x(\omega))/dx$, evaluated at the level of i's consumption in that state. Thus two agents, like our two sons, who have the same priors but different levels of consumption in the same states will behave as if they have different priors. So let us consider the general question: given two risk neutral agents with different priors and different information, under what conditions can a bet be designed such that both agents will agree to it after seeing their information?

Notice that if at one extreme the agents have no information but different priors, then a bet can always be arranged that both will accept that exploits the difference in their beliefs. At the other extreme, Theorem 3 shows that where the differences of opinion are entirely the result of differences of information, the agents will never bet. The general case lies between these two extremes.

Following Harsanyi, we call two priors defined with respect to two information partitions consistent if there exists a third prior which gives the same conditional expectations in each partition cell in partition 1 that agent 1 had with his prior, and the same conditional expectation in each partition cell in partition 2 that agent 2 had with his prior. Evidently the behavior of each agent 1 and 2 after receiving his information is indistinguishable from the behavior of an agent who had the third prior and their information. The only difference between agents 1 and 2 is therefore their information, and since by Theorem 4 this does not lead to speculation, we see that agents with consistent priors will not bet. Conversely, an elementary application of Farkas' Lemma shows that whenever consistency is violated, there must be a bet that both agents will accept. (See Morris (1991).)

To see that two priors can be consistent, and yet very different, which shows that differences in information tend to suppress rather than encourage speculation when it is common knowledge that agents are rational, let us consider a simplified version of the envelopes problem. Let $\Omega = \{a, b, c\}$, let $P_1 = \{\{a\}, \{b, c\}\}$ and $P_2 = \{\{a, b\}, \{c\}\}$. Let the priors be strictly positive, but otherwise arbitrary, and designate them (a_1, b_1, c_1) and (a_2, b_2, c_2), respectively. Agents 1 and 2 may disagree about the probability of every event (except the trivial events Ω and ϕ), but their priors are consistent, hence after their information is revealed, they will not agree to bet on anything. Take $a_3 = d$, take $b_3 = db_2/a_2$, and take $c_3 = b_3c_1/b_1$, where d is chosen so that the sum of a_3, b_3, c_3 is 1. Note that the same kind of construction can be carried out for the envelopes problem, so without knowing the utility functions of the agents (except that they are concave), we can be sure they will not bet against each other.

VII. BOUNDED RATIONALITY: MISTAKES IN INFORMATION PROCESSING

Common knowledge of rationality (interpreted as Bayesian Nash equilibrium) has surprisingly strong consequences. It rules out the possibility that repeated interaction leads to more cooperation in the Prisoner's Dilemma, it implies that agents cannot agree to disagree, it implies that they cannot bet, and most surprising of all, it banishes speculation. Yet casual empiricism suggests that all of these are common phenomena in the world. In this section we explore two possible explanations: first, that it is not really common knowledge that agents

are rational, though in fact they are; second, that in fact agents make mistakes processing information, and this is common knowledge.

1. I Know that You Know ... Finitely Often

Common knowledge involves an infinite number of iterations of the knowledge operator. What happens if agents are rational, and know that they are rational, and know that they know that they are rational, but only a finite number of times? As we shall see, everything changes. We may explore this situation in the context of a Bayesian game by assigning moves to each player at each state of the world that are optimal for most of the states, but not all of them. At the actual state of the world every agent may be optimizing, and each may know that the others are optimizing, etc, but somewhere in the chain of reasoning about the other's reasoning one of the states at which one agent does not optimize may appear.

Consider the example of agreeing to disagree, which was our third main example, from Geanakoplos-Polemarchakis (1982). Recall that there are 9 equally probable states of the world, and that agent 1 has partition {{1, 2, 3}, {4, 5, 6}, {7, 8, 9}}, and agent 2 has the partition {{1, 2, 3, 4}, {5, 6, 7, 8}, {9}}, and that the random variable x takes on the values (17, -7, -7, -7, 17, -7, -7, -7, 17) as a function of the nine states, respectively. Suppose now that if 9 were the actual state, agent 2 would give as his expectation of x not 17, which is the optimal answer, but -1. Suppose that the actual state of the world is 1. Then at 1 it is common knowledge that agent has opinion 1, and agent 2 has opinion -1. The agents agree to disagree. Moreover, 1 and 2 are both rational, and 1 knows that 2 knows that 1 knows that 2 is rational, but 1 does not know that 2 knows that 1 knows that 2 knows that 1 knows that 2 is rational.

Suppose now that x represents the amount of money that 2 must pay 1, depending on the state, and suppose that the agents must say whether they agree to the bet or not, as in the example drawn from Sebenius-Geanakoplos. Again let us suppose that 2 would make a mistake in case the state turned out to be 9, agreeing to bet when his information should reveal that he is sure to lose. Then when the real state is 1, it is common knowledge that the agents want to bet, but as before, not quite common knowledge that 2 is rational.

The envelopes puzzle is a still more compelling example. The two sons have $10,000 and $1,000 in their envelopes respectively. Suppose it were common knowledge that had son 1

seen $10,000,000 in his envelope, then he would bet (even though he could only lose in that case). Despite the fact that this eventuality is irrelevant, and both players know that it is irrelevant, and that they each know that the other knows that it is irrelevant and so on, they would agree to bet. The probability of nonoptimization is only 1/12, and of course by extending the maximum amount we can make the prior probability of nonoptimal behavior arbitrarily small, and still guarantee that the sons always bet against each other.

The astute reader will realize that although the prior probability of error can be made as small as possible in the envelopes example, the size of the blunder grows bigger and bigger. Indeed the expected error cannot vanish. But the situation is quite different for repeated games.

Consider again the Prisoner' Dilemma we studied earlier. Let there be four states of the world, *SS, SN, NS, NN* and hence $4T$ time indexed states. State S refers to an agent being sane, and N to him not being sane. Thus *NS* means agent 1 is not sane, but agent 2 is sane. Each agent knows whether he is sane or not, but he never finds out about the other agent. Each agent is sane with probability 4/5, and insane with probability 1/5, and these states are independent across agents, so for example the chance of *NS* is 4/25. Each agent plays the tit for that strategy when insane, and plays the tit for tat strategy until time T, when he defects for sure, when sane. The reader can verify that if the actual state is *SS*, then both agents are optimizing at all time periods. In the second to last period agent 1 can Defect, instead of playing C as his strategy indicates, gaining in payoff from 5 to 6. But with probability 1/5 he was facing N who would have dumbly played C in the last period, allowing 1 to get a payoff of 6 by playing D in the last period, whereas by playing D in the second to last period 1 gets only 1 in the last period even against N. Hence by defecting in the second to last period, agent 1 would gain 1 immediately, then lose 5 with probability 1/5 in the last period, which is a wash.

Thus by adding the chance of nonoptimal behavior in the last period (i.e. only in states (*SNT, NST*)) we get optimizing agents to cooperate all the way until the last period, and the common sense view that repetition encourages cooperation seems to be borne out. Note that in the above example we could not reduce the probability of N below 1/5, for if we did it would no longer be optimal for S to cooperate in the second to last period. Kreps, Milgrom, Roberts, and Wilson (1980) showed that if the insane agent also behaved nonoptimally for periods t less than T, then it is possible to support cooperation between the optimizing agents

while letting the probability of N go to 0 as T goes to infinity. It suffices to let the total expected number of periods of nonoptimizing behavior stay bounded away from zero.

In the Prisoner's Dilemma game a nontrivial threat is required to induce the optimizing agents not to defect, and this is what bounds the irrationality just described from below. A stronger result can be derived when the strategy spaces of the agents are continuous. In Chou-Geanakoplos (1985) it is shown that for generic continuous games, like the Cournot game where agents chose the quantity to produce, an arbitrarily small probability of nonoptimizing behavior in the last round alone suffices to enforce cooperation. The "altruistic" behavior in the last round can give the agents an incentive for a tiny bit of cooperation in the second to last round. The last two rounds together give agents the incentive for a little bit more cooperation in the third to last round, and so on. By the time one is removed sufficiently far from the end, there is a tremendous incentive to cooperate, otherwise all the gains from cooperation in all the succeeding periods will be lost. The nonoptimizing behavior in the last period may be interpreted as a promise or threat made by one of the players at the beginning of the game. Thus we see the tremendous power in the ability to commit oneself to an action in the distant future, even with a small probability. One man, like a Gandhi, who can credibly commit himself to starvation might change the behavior of an entire nation.

2. Mistakes in Information Processing

When agents are shaking hands to bet, it seems implausible that the bet is not common knowledge. It might seem even less plausible that the agents do not fully realize that they are all trying to win, i.e. that it is common knowledge they are optimizing. In this last section we continue to examine the implications of common knowledge, by weakening the maintained hypothesis that agents process information perfectly, which has been subsumed in the assumption that knowledge has so far exclusively been described by a partition. We seek to answer the question: how much irrationality must be permitted before speculation and agreements to disagree emerge in equilibrium? We shall show that if agents are a little irrational, they will speculate, but not bet or agree to disagree. If they get still more irrational, they will bet, but not agree to disagree about the probability of an event. Finally, with still more irrationality they will agree to disagree.

There are a number of errors that are typically made by decision makers that suggest that we go beyond the orthodox Bayesian paradigm. We list some of them:[5]

1. Agents ignore the subtle information content of some signals, and perceive only their face value. For example, an order to "produce 100 widgets" might convey all kinds of information about the mood of the boss, the profitability of the widget industry, the health of fellow workers and so on, if the receiver of the message has the time and capacity to think about it long enough. Another important example involves prices. It is very easy to compute the cost of a basket of goods at the going prices, but it takes much longer to deduce what the weather must have been like all across the globe to explain those prices. In Bayesian decision making, it is impossible to perform the first calculation without also performing the second.
2. Agents often do not notice when nothing happens. For example, it might be that there are only two states of nature: either the ozone layer is disintegrating or it is not. One can easily imagine a scenario in which a decaying ozone layer would emit gamma rays. Scientists, surprised by the new gamma rays would investigate their cause, and deduce that the ozone was disintegrating. If there were no gamma rays, scientists would not notice their absence, since they might never have thought to look for them, and so might incorrectly be in doubt as to the condition of the ozone.
3. What one knows is partly a matter of choice. For example, some people are notorious for ignoring unpleasant information. Often there are other psychological blocks to processing information.
4. People often forget.
5. Knowledge derived from proofs is not Bayesian. A proposition might be true or false. If an agent finds a proof for it, he knows it is true. But if he does not find a proof, he does not know it is false.
6. People cannot even imagine some states of the world.

[5]Much of Part VII is taken from Geanakoplos (1989). In particular, unless explicitly stated otherwise, all the theorems of this section are taken from that paper.

3. General Possibility Correspondences

We can model some aspect of all of these non-Bayesian methods of information processing by generalizing the notion of partition from the usual Bayesian analysis. Let Ω, a finite set, represent the set of all possible states of the world. Let $P : \Omega \to 2^{\Omega}\backslash\phi$ be an arbitrary "possibility correspondence," representing the information processing capacity of an agent. For each $\omega \in \Omega$, $P(\omega)$ is interpreted to mean the collection of states the agent thinks are possible when the true state is ω. Let $\underline{P}$ denote the range of P, so $\underline{P} = \{R \subset \Omega \mid \exists\omega \in \Omega, R = P(\omega)\}$. Given an arbitrary event $A \subset \Omega$, we say that the agent *knows* A *at* ω if $P(\omega) \subset A$, since for any $\omega' \in \Omega$ which he regards as possible at ω, $\omega' \in A$.

Consider, for example, $\Omega = \{a, b\}$ as the state space. Let the possibility correspondence $P : \Omega \to 2^{\Omega}$ take $P(a) = \{a\}$ and $P(b) = \{a, b\}$. We can interpret $\omega = a$ to mean the ozone layer is disintegrating, or a horse is winning, or a proposition is true. Similarly, we can interpret $\omega = b$ to mean the ozone is not disintegrating, or the horse is losing, or the theorem is false.

Since $P(a) = \{a\}$, when $\omega = a$ the agent knows that his horse is winning, or that the ozone is disintegrating, or that the theorem is true. But when $\omega = b$ the agent has no idea whether his horse is winning or losing, or what is happening to the ozone, or whether the theorem is true or false. The reason for the asymmetry in the agent's information processing could be interpreted as any one of the above categories of errors. The agent might take notice of the gamma rays in state a, but not notice that there were no gamma rays in state b. In the horse racing interpretation of the model, the agent might not be able to face the unpleasant news that his favorite horse is damaged. Or he might not remember an event where nothing of interest happened to him.

When we write (Ω, P) and P is not a partition the question arises whether the agent "knows the model." For if he does, then he should refine P into a partition. The agent in state b should realize it is state b, since he hasn't seen the gamma rays signalling state a, and deduce the partition $\{\{a\}, \{b\}\}$. The fact is however, that people do not notice the dog that "didn't bark in the nighttime." So we have in mind that the agent does not know the model. After all, in any state ω, in order to act the agent does not need to know the model, but acts on the basis of the information he has at hand.

We now describe 3 axioms that a general possibility correspondence might satisfy.

DEFINITION: We say that P is *nondeluded* if $\omega \in P(\omega)$ for all $\omega \in \Omega$. Under this hypothesis the agent who processes information according to P always considers the true state as possible.

DEFINITION (Knowing that you know KTYK): When Knowledge is Self-Evident):

If for all $\omega \in \Omega$, and all $\omega' \in P(\omega)$, we have $P(\omega') \subseteq P(\omega)$, then we say that the agent *knows what he knows*. If the agent knows some A at ω, and can imagine ω', then he would know A at ω'. Bacharach (1985), Shin (1987), and Samet (1987) have all drawn attention to this property. If the agent can recognize circumstances which confine the possible states of the world to $R \in \underline{P}$, then whenever $\omega \in R$, so that these circumstances do indeed obtain, the agent must realize that.

DEFINITION: The event $E \subset \Omega$ is *self-evident* to the agent who processes information according to P if $P(\omega) \subset E$ whenever $\omega \in E$. A self-evident event can never occur without the agent knowing that it has occurred.

The axiom KTYK implies that every $R \in \underline{P}$ is self-evident to the agent.

Shin (1987) has suggested that KTYK and nondelusion are the only properties that need hold true for an agent whose knowledge was derived by logical deductions from a set of axioms.

DEFINITION: We say that P is *nested* if for all ω and ω', either $P(\omega) \cap P(\omega') = \phi$, or else $P(\omega) \subseteq P(\omega')$, or else $P(\omega') \subseteq P(\omega)$.

An example might make the significance of nondelusion, KTYK, and nested clearer. Let there be just two propositions of interest in the universe, and let us suppose that whether each is true or false is regarded as good or bad, respectively. The state space is then $\Omega = \{GG, GB, BG, BB\}$. One type of information processor P might always disregard anything that is bad, but remember anything that is good. Then $P(GG) = \{GG\}$, $P(GB) = \{GG, GB\}$, $P(BG) = \{GG, BG\}$, $P(BB) = \Omega$. (See Diagram 1a.) It is clear that P satisfies nondelusion and KTYK, but does not satisfy nested. Moreover, when the reports are GB the agent chooses to remember only the first, while if they are BG he chooses to remember only the last. Alternatively, consider an agent with possibility correspondence Q who can remember GG and BB because the pattern is simple, and can also remember when he sees GB that the first report was good whereas with BG he remembers nothing. Then $Q(GG) = \{GG\}$,

$Q(GB) = \{GB, GG\}$, $Q(BG) = \Omega$, $Q(BB) = \{BB\}$. (See Diagram 1b.) This does satisfy nested, as well as the other two conditions. Nondelusion in these examples means that the agent never mistakes a good report for a bad report, or vice versa. KTYK means that if an agent recalls some collection of reports, then whenever all those reports turn out the same way he must also recall them (and possibly some others as well). Nested means that the reports are ordered in the agent's memory. If he remembers some report, then he must also remember every report that came earlier on the list.

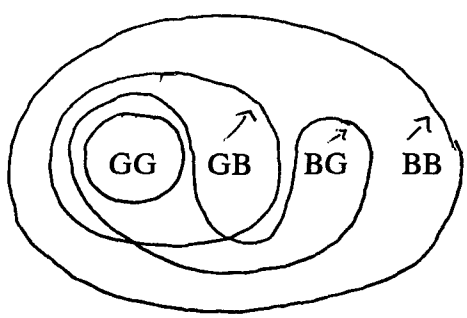


Diagram 1a
Nondeluded, KTYK, but not nested

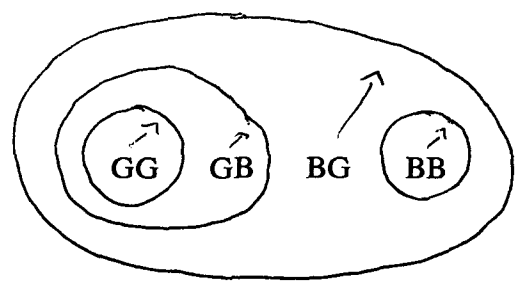


Diagram 1b
Nondeluded, nested, and KTYK

Note that nested and KTYK are independent properties. Let $\Omega = \{a,b,c\}$, and let $P(a) = P(c) = \{a,b,c\}$, while $P(b) = \{b,c\}$. Then P is nondeluded and nested, but P does not satisfy KTYK, since $c \in P(b)$ but $P(c) \subseteq P(b)$.

2. Decision Theory, Game Theory, and Common Knowledge with General Possibility Correspondence

Having explained that many kinds of errors in information processing can be captured by extending the notion of possibility correspondence beyond that of partitions, let us hasten to add that the notions of optimal (Bayesian) decision function, (Bayesian) Nash equilibrium, and common knowledge can be carried over from the previous sections with no formal change, simply by replacing the partition possibility correspondences with more general possibility correspondences. Thus $f_i : \Omega \to A_i$ is an optimal decision function for the decision problem $(\Omega, P_i, \pi_i, A_i, u_i)$ if and only if the same formal conditions

(1) $\quad [P_i(\omega) = P_i(\omega')] \Rightarrow [f_i(\omega) = f_i(\omega')] \quad \forall \omega \in \Omega$ and

$$(2) \qquad \sum_{\omega' \in P_i(\omega)} u_i(f_i(\omega), \omega')\pi_i(\omega') \geq \sum_{\omega' \in P_i(\omega)} u_i(a, \omega')\pi_i(\omega') \quad \forall a \in A_i, \forall \omega \in \Omega.$$

One consequence of this extension is that an agent does not necessarily "know what he is doing." At a given ω there are other states $\omega' \in P_i(\omega)$ which he imagines are possible, without realizing that if those states occurred he would choose to act differently.

EXAMPLE 1: Let $\Omega = \{a, b, c\}$, $P(a) = \{a, b\}$, $P(b) = \{b\}$, $P(c) = \{b, c\}$. (Note that (Ω, P) satisfies nondeluded and KTYK, but not nested.) Let $\pi(a) = \pi(c) = 2/7$ and $\pi(b) = 3/7$. Let the action set be $A = \{B, N\}$, corresponding to bet or not bet. Let the payoffs from not betting be $u(N, a) = u(N, b) = u(N, c) = 0$. Let the payoffs to betting be $u(B, a) = u(B, c) = -1$, while $u(B, b) = 1$. It is easy to calculate that $f(\omega) = B$ for all $\omega \in \Omega$ is optimal for (A, Ω, P, u, π). Yet,

$$\sum_{\omega \in \Omega} u(N, \omega)\pi(\omega) = 0 > -\frac{1}{7} = \sum_{\omega \in \Omega} u(B, \omega)\pi(\omega) .$$

Of course $g(\omega) = N$ for all $\omega \in \Omega$ is optimal for (A, Ω, Q, u, π) where $Q(\omega) = \Omega$ for all ω, so for this example a little knowledge is a bad thing. The sure-thing-principle is also violated.

EXAMPLE 2: Let $\Omega = \{a, b, c\}$, and let $P(a) = P(c) = \{a, b, c\}$, while $P(b) = \{b, c\}$. Then (Ω, P) is nondeluded and nested, but does not satisfy KTYK. Let $A = \{B, N\}$, let $\pi(\omega) = 1/3$, for all $\omega \in \Omega$, and let $u(N, \omega) = 0$ for all $\omega \in \Omega$, while $u(B, a) = -2$, $u(B, b) = -2$, $u(B, c) = 3$. Then $f = (f(a), f(b), f(c)) = (N, B, N)$ is optimal for (A, Ω, P, u, π), but

$$\sum_{\omega \in \Omega} u(N, \omega)\pi(\omega) = 0 > -\frac{2}{3} = \sum_{\omega \in \Omega} u(f(\omega), \omega)\pi(\omega) .$$

Once again let the coarse partition be $Q(\omega) = \Omega$ for all ω. Then $g(\omega) = N$ for all $\omega \in \Omega$ is optimal for (A, Ω, Q, u, π), and again more knowledge hurts. But this time the sure-thing-principle (as stated in Theorem 3) is not violated.

EXAMPLE 3: Let $\Omega = \{a, b\}$, let $P(a) = \{a\}$, $P(b) = \{a, b\}$. Then (Ω, P) satisfies all the properties nondeluded, nested, and KTYK. Let $A = \{B, N\}$, let $\pi(a) = \pi(b) = 1/2$. Let $u(N, a) = u(N, b) = 0$, let $u(B, a) = 1$, and $u(B, b) = -2$. Then $f(a) = B$, $f(b) = N$ is optimal for (A, Ω, P, u, π), and

$$\sum_{\omega\in\Omega} u(N, \omega)\pi(\omega) = 0 < -\frac{1}{2} = \sum_{\omega\in\Omega} u(f(\omega), \omega)\pi(\omega) .$$

In this case knowledge helps. Observe also that if we changed the payoff at (B, b) to $u(B, b) = -1$, then there would be a second optimal decision function $\tilde{f}(a) = B = \tilde{f}(b)$. In that case $\sum_{\omega\in\Omega} u(\tilde{f}(\omega), \omega)\pi(\omega) = 0$ is worse than the payoff arising from f, but still as good as the payoff arising from (A, Ω, Q, u, π) where $Q(\omega) = \Omega$ for all $\omega \in \Omega$.

In Examples 1 and 2, the agent was (ex ante) worse off knowing more because he did not process information coherently. One difference is that in Example 1, the agent always chose B, while in 2 he did not. In Example 3 the agent was not worse off, although he also did not process information correctly. In general we have:

THEOREM: *Let (Ω, P) satisfy nondeluded, nested, and KTYK. Let Q be a partition of Ω that is a coarsening of P. Let f, g be optimal for (A, Ω, P, u, π) and (A, Ω, Q, u, π) respectively. Then*

$$\sum_{\omega\in\Omega} u(g(\omega), \omega)\pi(\omega) \le \sum_{\omega\in\Omega} u(f(\omega), \omega)\pi(\omega) .$$

Conversely, suppose that (Ω, P) fails to satisfy one or more of the above hypotheses. Then there is a partition Q of Ω that is a coarsening of P, and an A, u, π such that f, g are optimal for (A, Ω, P, u, π), (A, Ω, Q, u, π), respectively, and yet the above inequality is strictly reversed.

Before turning to the sure-thing-principle, let us introduce two final properties for P_i, called positively balanced and balanced.

DEFINITION: The information processor (Ω, P) is *positively balanced* if for every self-evident set $E \subset \Omega$ iff there exists a function $\lambda : \underline{P} \to \mathbb{R}_+$, such that (letting χ_A be the characteristic function of any set $A \subset \Omega$)

$$\sum_{\substack{C\in\underline{P}\\ C\in E}} \lambda(C)\chi_C(\omega) = \chi_E(\omega) \quad \text{for } \textit{all} \quad \omega \in \Omega .$$

If the same holds true for some λ unrestricted in sign, $\lambda : \underline{P} \to \mathbb{R}$, then we say that (Ω, P) is *balanced.* QED

Balancedness gives a condition under which one can say that every element $\omega \in E$ is equally scrutinized by the information correspondence P. Every element $C \in \underline{P}$ has an intens-

ity $\lambda(C)$, and the sum of the intensities with which each $\omega \in E$ is considered possible by P is the same, namely 1. Balancedness is a generalization of partition. If E can be written as a disjoint union of elements of $\underline{P}$, then (Ω, P) is trivially balanced with respect to E.[6]

It can be shown that KTYK implies balanced, and that nested implies positively balanced (which in turn trivially implies balanced). Examples 2 and 3 are positively balanced, and example 1 is balanced but not positively balanced.

THEOREM: *The sure-thing-principle (as described in Theorem 3) applies to every decision problem* $(\Omega, P_i, \pi_i, A_i, u_i)$, *for fixed* P_i, *if and only if* P_i *is nondeluded and positively balanced.*

We can also extend common knowledge in a natural way when the P_i are general possibility correspondences. We say that the event E is common knowledge at ω among the information processors $(\Omega, P_i)_{i \in I}$ if and only if $M(\omega) \subset E$ where $M(\omega)$ is the smallest public event containing $\bigcup_{i \in I} P_i(\omega)$. Recall that R is a public event iff it is self-evident to every $i \in I$, $P_i(\omega') \subset R$ for all $\omega' \in R$. Clearly the intersection of public events is public, so $M(\omega)$ is well-defined. Shin (1987) and Samet (1987) have shown for the special case when all the P_i satisfy nondelusion and KTYK that this definition of common knowledge retains the sense of i knowing that j knows that i knows, etc.

The definition of (Bayesian) Nash equilibrium can be carried over directly from IV without formal change, by replacing partitions with arbitrary possibility correspondences. The following extensions to the nonspeculation and agreement theorems can be proved.

THEOREM: *The nonspeculation Theorem 3 holds in general if and only if each* P_i *satisfies nondeluded, KTYK, and nested. (The proof is derived from the conditions for which more knowledge cannot hurt.)*

THEOREM 17: *Betting is ruled out in general if and only if the strictly weaker hypotheses of nondeluded and positively balanced are satisfied by each* P_i. *(The proof relies on the generalized sure-thing-principle.)*

THEOREM 18: *The proposition that agents cannot agree to disagree about the probability of an event holds if and only if the yet strictly weaker hypotheses of nondelusion and balanced are satis-*

[6]Balancedness is similar to a concept (with the same name) that played an important role in the development of the theory of the core in cooperative game theory.

fied by every P_i. (Samet (1987) showed that nondelusion and KTYK imply no agreeing to disagree.)

Thus indeed we find that weakening the rationality assumptions to allow for information processing errors allows eventually for disagreements, common knowledge betting, and speculation. One surprising element of the story is that the amount of rationality needed to eliminate each phenomenon can be strictly ordered, so that, for example betting on whether an event will happen can occur (and be common knowledge) between agents who would not agree to disagree about the precise probability of the event in question. And two agents might speculate against each other in the market, yet not be willing to shake hands on a bet that becomes common knowledge.

REFERENCES

Aumann, R., "Subjectivity and Correlation in Randomized Strategies," *Journal of Mathematical Economics* (1974), 1:67-96.

_____, "Agreeing to Disagree," *The Annals of Statistics* (1976) 4:1236-1239.

_____, "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica* (1987), 55:1-18.

_____, "Irrationality in Game Theory," mimeo, 1988.

_____, "Interactive Epistemology," mimeo, 1989.

Bacharach, M., "Some Extensions of a Claim of Aumann in an Axiomatic Model of Knowledge," *Journal of Economic Theory* (1985), 37:167-190.

Binmore, K. and A. Brandenburger, "Common Knowledge and Game Theory," London School of Economics, 1988.

Boge, W. and th. Eisele, "On Solutions of Bayesian Games," *International Journal of Game Theory* (1979), 8(4):193-215.

Bollobás, B., ed., *Littlewood's Miscellany*. Cambridge: Cambridge University Press, 1953.

Brandenburger, A. and E. Deckel, "Common Knowledge with Probability 1," forthcoming in *Journal of Mathematical Economics* (1986).

Brandenburger, A., E. Deckel, and J. Geanakoplos, "Correlated Equilibrium with Generalized Information Structures," Yale University, 1988.

Cave, J., "Learning to Agree," *Economic Letters* (1983), 12:147-152.

Chow, C. and J. Geanakoplos, "The Power of Commitment," CFDP No. xxx, 1985.

Dubey, P., J. Geanakoplos, and M. Shubik, "The Revelation of Information in Strategic Market Games: A Critique of Rational Expectations Equilibrium," *Journal of Mathematical Economics* (1987), 16(2):105-138.

Geanakoplos, J. "Common Knowledge of Actions Negates Asymmetric Information," mimeo, Yale University, 1987.

Geanakoplos, J., "Common Knowledge, Bayesian Learning, and Market Speculation with Bounded Rationality," mimeo, Yale University, 1988.

_____, "Game Theory without Partitions, and Applications to Speculation and Consensus," Cowles Foundation Discussion Paper #914, Yale University, 1989.

Geanakoplos, J. and H. Polemarchakis, "We Can't Disagree Forever," *Journal of Economic Theory* (1982), 28:192-200.

Gilboa, I., "Information and Meta-Information," Tel Aviv Working Paper #3086, 1986.

Halpern, J. Y., "Reasoning about Knowledge: An Overview," IBM Research Report RJ-5001, 1986.

Halpern, J. and Y. Moses, "Knowledge and Common Knowledge in a Distributed Environment," in Proc. 3rd ACM Conference on Principles of Distributed Computing, 1984, pp. 50-61.

Kaneko, M., "Structural Common Knowledge and Factual Common Knowledge," RIEE Working Paper 87-27.

Kreps, D., "A Note on Fulfilled Expectations Equilibrium," *Journal of Economic Theory* (1977), 32-43.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," *Journal of Economic Theory* (1982), 27:245-23.

Kripke, S., "Semantical Analysis of Model Logic," *Z. Math Logik Grundlag der Math* (1963), 9:67-96.

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

McKelvey, R. and T. Page, "Common Knowledge, Consensus, and Aggregate Information," *Econometrica* (1986), 54:109-127.

Mertens, J. F. and S. Zamir, "Formation of Bayes Analysis for Games with Incomplete Information," *International Journal of Game Theory* (1985), 14:17-27.

Milgrom, P., "An Axiomatic Characterization of Common Knowledge," *Econometrica* (1981), 49:219-222.

_____ and N. Stokey, "Information, Trade, and Common Knowledge," *Journal of Economic Theory* (1982), 26:17-27.

Monderer, D. and D. Samet, "Approximating Common Knowledge with Common Beliefs," mimeo, Northwestern, 1988.

Morris, S., "The Role of Beliefs in Economic Theory," PhD. Dissertation, Yale University, 1991.

Nielsen, L., "Common Knowledge, Communication, and Convergence of Beliefs," *Mathematical Social Sciences* (1984), 8:1-14.

_____, A. Brandenburger, J. Geanakoplos, R. McKelvey, and T. Page, "Common Knowledge of an Aggregate of Expectations," mimeo, 1989.

Parikh, R. and P. Krasucki, "Communication, Consensus and Knowledge," mimeo, 1987.

Rubinstein, A. "The Electronic Mail Game: Strategic Behavior under 'Almost Common Knowledge'," *American Economic Review* (1989), 79(3):385-391.

Samet, D., "Ignoring Ignorance and Agreeing to Disagree," MEDS Discussion Paper, 1987, Northwestern University.

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

Sebenius J. and J. Geanakoplos, "Don't Bet On It: Contingent Agreements with Asymmetric Information," *Journal of the American Statistical Association* (1983), 78:424-426.

Shin, H., "Logical Structure of Common Knowledge, I and II," unpublished, Nuffield College, Oxford, 1987.

Tan, T. and S. Werlang, "The Bayesian Foundations of Rationalizable Strategic Behavior and Nash Equilibrium Behavior," unpublished, Department of Economics, Princeton University, 1984.

Tirole, J. "On the Possibility of Speculation under Rational Expectations," *Econometrica* (1982), 1163-1181.