Cowles Foundation Paper 828

Journal of Economic Perspectives—Volume 6, Number 4—Fall 1992—Pages 53–82

Common Knowledge

John Geanakoplos

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 think about what the others know and do not 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, "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." Later, Moriarty's limited mastery of interactive epistemology allowed Holmes and Watson to escape from the train at Canterbury, a mistake which ultimately led to Moriarty's death, because he 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. Any prospective stock buyer who has information suggesting the price will go up 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, the prospective buyer must ask whether buying is still a good idea.

Can rational agents agree to disagree? Is this question connected to whether rational agents will speculate in the stock market? How might the

■ *John Geanakoplos is Professor of Economics, Yale University, New Haven, Connecticut.*

degree of rationality of the agents, or the length of time they talk, influence the answer to this question?

The notion of common knowledge plays a crucial role in the analysis of these questions. An event is common knowledge among a group of agents if each one knows it, if each one knows that the others know it, if each one knows that each one knows that the others know it, and so on. Thus, common knowledge is the limit of a potentially infinite chain of reasoning about knowledge. A formal definition of common knowledge was introduced into the economics literature by Robert Aumann in 1976.

Public events are the most obvious candidates for common knowledge. But events that the agents create themselves, like the rules of a game or contract, can also plausibly be seen as common knowledge. Certain beliefs about human nature might also be taken to be common knowledge. Economists are especially interested, for example, in the consequences of the hypothesis that it is common knowledge that all agents are optimizers. Finally, it often occurs that after lengthy conversations or observations, 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. It will begin with several puzzles that illustrate the strength of the common knowledge hypothesis. It will then study how common knowledge can illuminate many problems in economics. In general, the discussion will show that a talent for interactive thinking is advantageous, but if everyone can think interactively and deeply all the way to common knowledge, then sometimes puzzling consequences may result.

Puzzles About Reasoning Based on the Reasoning of Others

The most famous example illustrating the ideas of reasoning about common knowledge can be told in many equivalent ways.[1] The earliest version that I could find appears in Littlewood's Miscellania, published in 1953, although he noted that it was already well-known and had caused a sensation in Europe some years before. The colonial version of the story begins with many cannibals married to unfaithful wives, and of course a missionary. I shall be content to offer a more prosaic version, involving a group of logical children wearing hats.

Imagine three girls sitting in a circle, each wearing either a red hat or a white hat. Suppose that all the hats are red. When the teacher asks if any

[1]This example is so well-known that it is difficult to find out who told it first. It appeared in Martin Gardner's collection (1984). It had already been presented by Gamow and Stern (1958) as the puzzle of the cheating wives. It was discussed in the economics literature by Geanakoplos and Polemarchakis (1982). It appeared in the computer science literature in Halpern and Moses (1984).

student can identify the color of her own hat, the answer is always negative, since nobody can see her own hat. But 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 can see two red hats in the room) then the answers change. The first student who is asked cannot tell, nor can the second. But the third will be able to answer with confidence that she is indeed wearing a red hat.

How? By following this chain of logic. If the hats on the heads of both children two and three were white, then the teacher's remark would allow the first child to answer with confidence that her hat was red. But she cannot tell, which reveals to children two and three that at least one of them is wearing a red hat. The third child watches the second also admit that she cannot tell her hat color, and then reasons as follows: "If my hat had been white, then the second girl would have answered that she was wearing a red hat, since we both know that at least one of us is wearing a red hat. But the second girl could not answer. Therefore, I must be 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 story contains several crucial elements: it is common knowledge that everybody can see two hats; the pronouncements of ignorance are public; each child knows the reasoning used by the others. Each student knew the apparently innocuous fact related by the teacher—that there was at least one red hat in the room—but the fact was not common knowledge between them. When it became common knowledge, the second and third children could draw inferences from the answer of the first child, eventually enabling the third child to deduce her hat color.

Consider a second example, also described by Littlewood, involving betting. An honest 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. 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. The second son finds 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 tells each son what his brother said and repeats the question. Again both sons say yes. The father relays the brothers' answers and asks each a third time whether he is willing to pay \$1 to switch envelopes. Again both say yes. But if the father relays their answers and asks each a fourth time, the son with \$1,000 will say yes, but the son with \$10,000 will say no.

It is interesting to consider a slight variation of this story. Suppose now that the very first time the father tells each of his sons that he can pay \$1 to switch envelopes it is understood that if the other son refuses, the deal is off and the father keeps the dollar. What would they do? Both would immediately say no, as we shall explain in a later section.

A third puzzle is more recent.[2] 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 a murder occurs, and the two detectives are ordered to conduct independent investigations. They promise not to share any data gathered from their research, and 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. They meet on the way to the station. 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. Hearing the other's opinion, each detective may change his mind and give another opinion. This may cause a further change in opinion.

If they talk long enough, however, then we can be sure that both detectives will 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, and so on. And if they had shared their clues, they might well have agreed on an entirely different suspect!

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 in the riddle of the detectives, common knowledge of each agent's action (what suspect is chosen, given the decision rules) negates asymmetric information about events (what information was actually gathered). At the end, the detectives are necessarily led to a decision which can be explained by a common set of clues, although in fact their clues might have been different, even allowing for the deductions each made from hearing the opinions expressed in the conversation. The lesson we shall draw is that asymmetric information is important only if it leads to uncertainty about what the other agents are doing.

Interactive Epistemology

To examine the role of common knowledge, both in these three puzzles and in economics more generally, the fundamental conceptual tool we shall use

[2]This story is due to Bacharach, perhaps somewhat embellished by Aumann, from whom I learned it. It caricatures the analysis in Aumann (1976), Geanakoplos and Polemarchakis (1982), and Cave (1983).

is the state of the world. Leibnitz first introduced this idea; it has since been refined by Kripke, Savage, Harsanyi, and Aumann, among others. A "state of the world" is very detailed. It specifies the physical universe, past, present, and future; it describes what every agent knows, and what every agent knows about what every agent knows, and so on; it specifies 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, and so on; it specifies 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, and so on; 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, and so on.

Let Ω be the set of all possible worlds, defined in this all-embracing sense. We model limited knowledge by analogy with a far-off observer who from his distance cannot quite distinguish some objects from others. For instance, the observer might be able to tell the sex of anyone he sees, but not who the person is. The agent's knowledge will be formally described throughout most of this survey by a collection of mutually disjoint and exhaustive classes of states of the world called cells that *partition* Ω. If two states of nature are in the same cell, then the agent cannot distinguish them.

As an 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 perceive that any of the even states is possible, while none of the odd states is possible. To remind us that what the agent thinks is possible depends on what the real state of the world is, we write $P_i(4) = \{2,4,6,8\}$ to denote the partition cell $\{2,4,6,8\}$ of states that i thinks are possible when the true state of the world is 4.

Any subset E contained in Ω is called an event. If the true state of the world is ω, and if $\omega \in E$, then we say that E occurs or is true. If every state that i thinks is possible (given that ω is the true state) entails E, which we write as $P_i(\omega) \subset E$, then we say that agent i "knows" E. Note that at some ω, i may know E, while at other ω, i may not. If whenever E occurs i knows E, that is if $P_i(\omega) \subset E$ for all states ω 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 we just used, let $E = \{1,2,3,5,7\}$. Then at $\omega = 3$ agent i knows ω is an odd number between 1 and 8 and hence in E. At $\omega = 2$ agent i knows ω is an even number between 1 and 8, and hence cannot tell whether or not ω is in E, since 2 is but 4, 6, and 8 are not. Hence E is not self-evident to i. The only self-evident events to i are $\{1,3,5,7\}$, $\{2,4,6,8\}$, and Ω itself.

We can model the process of learning by analogy to an observer getting closer to what he is looking at. Things which he could not previously distinguish, such as for example whether the people he is watching have brown hair or black hair, become discernible. In our framework, such an agent's partition becomes finer when he learns, perhaps containing four cells {female/brown hair}, {{female/black hair}, {male/brown hair}, {male/black hair}} instead of two.

Naturally, we can define the partitions of several agents, say i and j, simultaneously on the same state space. There is no reason that the two agents should have the same partitions. Indeed different people typically have different vantage points, and it is precisely this asymmetric information that makes the question of common knowledge interesting. In the example of integers from 1 to 8, suppose that we introduce a second partition $P_j = \{\{1, 2\},\{3, 4, 5, 6, 7, 8\}\}$. Note that the event $E = \{1, 2, 3, 5, 7\}$ is also not self-evident to agent j.

Suppose now that agent i knows the partition of j; that is, suppose that i knows what j is able to know, and vice versa. This does not mean that i knows what j knows. For example, at $\omega = 1$, i knows that E occurs, and j knows that E occurs, but i does not know that j knows that E occurs, since i thinks 3 is possible, in which case j would think all of $\{3, 4, 5, 6, 7, 8\}$ are possible, which includes states not in E.

Since the possibility correspondences are functions of the state of nature, the reader can see that in effect each state of nature ω specifies not only the physical universe, but also what each agent knows about the physical universe, and what each agent knows each agent knows about the physical universe and so on.

The Puzzles Reconsidered

With this framework, reconsider the puzzle of the three girls with red and white hats. A state of nature ω corresponds to the color of each child's hat. The table lists the eight possible states of nature.

STATES OF THE WORLD

		a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
	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*

In the notation we have introduced, the set of all possible states of nature Ω can be summarized as $\{a, b, c, d, e, f, g, h\}$, with a letter designating each state. Then, the partitions of the three 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 could know at the outset. Each can observe four cells, based on the hats the others are wearing: both red, both white, or two combinations of one of each. None can observe her own hat, which is why the cells come in groups of two states. For example, if the true state of the world is all red hats—that is,

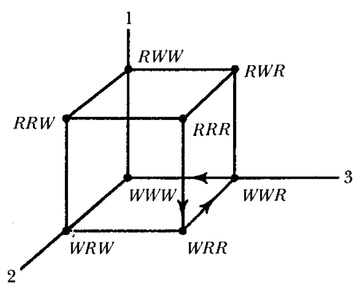


Figure 1a

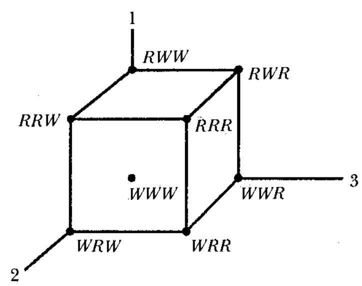


Figure 1b

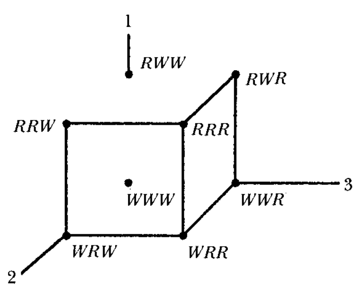


Figure 1c

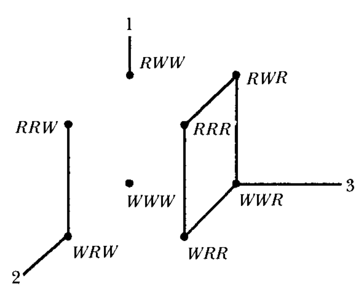


Figure 1d

$\omega = a = RRR$—then agent 1 is informed of $P_1(a) = \{a, e\}$, and thus knows that the true state is either $a = RRR$, or $e = WRR$. In the puzzle, agent i "knows" her hat color only if the color is the same in all states of nature ω which that agent regards as possible.

In using this model of knowledge to explain the puzzle of the hats, it helps to represent the state space as the vertices of a cube, as in Figure 1*a*.[3] Think of R as 1 and W as 0. Then every corner of a cube has three coordinates which are either 1 or 0. Let the i^{th} coordinate denote the hat color of the i^{th} agent. For each agent i, 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 i^{th} axis, so that if the vertical axis is designated as 1, the four vertical sides of the cube correspond to the four cells in agent 1's partition.

An agent i knows her 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, no agent knows her hat color in any state.

Note that 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

[3]This has been pointed out in Fagin, Halpern, Moses, and Vardi (1988).

WRR is possible. But at *WRR*, agent 2 thinks *WWR* is possible. And at *WWR* agent 3 thinks *WWW* is possible. In short, at *RRR* agent 1 thinks that agent 2 might think that agent 3 might think that *WWW* is possible. In other words, *WWW* is reachable from *RRR*. This chain of thinking is indicated in the diagram by the path marked by arrows.

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. When the teacher announces that there is at least one red hat in the room, that is tantamount to declaring that the actual state is not *WWW*. This can be captured pictorially by dropping all the edges leading out of the state *WWW*, as seen in Figure 1*b*. (Implicitly, we are assuming that had all the hats been white the teacher would have said so.) Each of the girls now has a finer partition than before, that is, some states that were indistinguishable before have now become distinguishable. There are now two connected components to the graph: 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 she does not know her hat color, she reveals that the state could not be *RWW*, since if it were, she would also be able to deduce the state from her own information and the teacher's announcement and therefore would have known her hat color. We can capture the effect of the first student's announcement on every other agent's information by severing all the connections between the set {*WWW*, *RWW*} and its complement. Figure 1*c* now has three different components, and agents 2 and 3 have finer partitions.

The announcement by student 2 that she still does not know her hat color reveals that the state cannot be any of {*WWW*, *RWW*, *RRW*, *WRW*}, since these are the states in which the above figure indicates student 2 would have the information (acquired in deductions from the teacher's announcement and the first student's announcement) to unambiguously know her hat color. Conversely, if 2 knows her hat color, then she 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 students' information partitions by severing all connections between the set {*WWW*, *RWW*, *RRW*, *WRW*} and its complement, producing Figure 1*d*. Notice now that the diagram has four separate components.

In this final situation, after hearing the teacher's announcement, and each of student 1 and student 2's announcements, student 3 knows her hat color at all the states. Thus no more information is revealed, even when student 3 says she knows her hat color is red.

If, after student 3 says yes, student 1 is asked the color of her hat again, she will still say no, she cannot tell. 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, their responses will be "common knowledge": 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, and so on. By logic alone the students come to a common understanding of what must happen in the future. Note also that at the final state of information, the three girls have different information.

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. For example, the analysis shows 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 her hat. In fact, we will argue later that one student must eventually realize her 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.

The second puzzle, about the envelopes, can be explored along similar lines, as a special case of the analysis in Sebenius and Geanakoplos (1983); it is closely related to Milgrom and Stokey (1982). For that story, take the set of all possible worlds Ω to be the set of ordered pairs (m, n) with m and n integers between 1 and 7; m and n differ by one, but either could be the larger. At state (m, n), agent 1 has 10^m dollars in his envelope, and agent 2 has 10^n dollars in his envelope.

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						■	

The partitions divide the state space into two components, namely those states reachable from $(1, 2)$ and those states reachable from $(2, 1)$. In one connected component of mutually reachable states, agent 1 has an odd number and 2 has an even number, and this is "common knowledge"—that is, 1 knows it and 2 knows it and 1 knows that 2 knows it, and so on. For example, the state $(4, 3)$ is reachable from the state $(2, 1)$, because at $(2, 1)$, agent 1 thinks the state $(2, 3)$ is possible, and at $(2, 3)$ agent 2 would think the state $(4, 3)$ is possible. In the other component of mutually reachable states, the even/odd 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 start by believing that all states are equally likely. Thus, at $\omega = (4, 3)$ 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 tells each of them the other's previous answer, however, the situation changes. Neither son would bet if he had the maximum $10 million in his envelope, so when the sons learn that the other is willing to bet, it becomes "common knowledge" that neither number is 7. The state space is now divided 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 he would get only $100,000. 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) 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 the next time his father asks him, he voids the bet.

If from the beginning the sons had to ante a dollar knowing that they could not recover it if the other son refused to bet, then both of them would say that they did not want the bet on the very first round. We explain this later.

Here is a third example, reminiscent of the detective story. Suppose, following Aumann (1976) and Geanakoplos and 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. 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 the set of all possible worlds be $\Omega = \{1, 2, \ldots, 9\}$, and 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 a random variable x takes on the following values as a function of the state:

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.

$$1\overline{\ldots}2\overline{\ldots}3\ldots4\overline{\qquad}5\overline{\ldots}6\ldots7\overline{\ldots}8\overline{\qquad}9$$

Suppose that $\omega = 1$. Agent 1 calculates his opinion about the expectation of x by averaging the values of x over the three states 1, 2, 3 that he thinks are possible, and equally likely. When agent 1 declares that his opinion of the expected value of x is 1, he reveals nothing, since no matter what the real state of the world, his partition would have led him to the same conclusion. But when agent 2 responds with his opinion, he is indeed revealing information. For if he thinks that $\{1, 2, 3, 4\}$ are possible, and equally likely, his opinion about the expected value of x is -1. Similarly, if he thought that $\{5, 6, 7, 8\}$ were possible and equally likely, he would say -1, while if he knew only $\{9\}$ was possible, then he would say 17. Hence when agent 2 answers, 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:

$$1\overline{\ldots}2\overline{\ldots}3\ldots4\overline{\qquad}5\overline{\ldots}6\ldots7\overline{\ldots}8 \qquad 9$$

If agent 1 now gives his opinion again, he will reveal new information, even if he repeats the same number 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:

$$1\overline{\ldots}2\overline{\ldots}3\ldots4\overline{\qquad}5\overline{\ldots}6 \qquad 7\overline{\ldots}8 \qquad 9$$

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. Depending on whether he says -1, 5, or -7, agent 1 will learn something different, and so the partitions become:

$$1\overline{\ldots}2\overline{\ldots}3\ldots4 \qquad 5\overline{\ldots}6 \qquad 7\overline{\ldots}8 \qquad 9$$

Similarly if 1 responds a third time, he will yet again reveal more information, even if his 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 $(1, -1), (1, -1), (1, 1)$.

Although this example suggest that the partitions of the agents will converge, this is not necessarily true—all that must happen is that the opinions about expectations converge. Consider the state space below, and suppose that agents assign probability $1/4$ to each state. As usual, 1 cannot distinguish states in the same row and 2 cannot distinguish states in the same column.

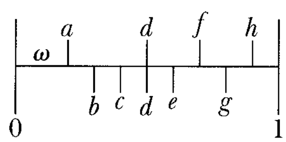
$$\begin{matrix} a & b \\ c & d \end{matrix}$$

Let $x(a) = x(d) = 1$, and $x(b) = x(c) = -1$. Then at $\omega = a$, both agents will say that their expectation of x is 0, and agreement is reached. But the information of the two agents is different. If asked why they think the expected value of x is 0, they would give different explanations, and if they shared their reasons, they would end up agreeing that the expectation should be 1, not 0.

As pointed out in Sebenius and Geanakoplos (1983), if instead of giving their opinions of the expectation of x, the agents in the last two examples instead were called upon to agree to bet, 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 of betting and opining are quite different, though they both result in a kind of agreement.

Characterizing Common Knowledge of Events and Actions

To this point, the examples and discussion have used the term common knowledge rather loosely, as simply meaning a fact that everyone knows, that everyone knows that everyone knows, and so on. An example may help to give the reader a better grip on the idea.



The whole interval [0, 1] represents Ω. The upper subintervals with endpoints $\{0, a, d, f, h, 1\}$ represent agent 1's partition. The lower subintervals with endpoints $\{0, b, c, d, e, g, 1\}$ represent 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 might think $[0, a]$ is possible or $[a, d]$ is possible. But at ω, nobody need think outside $[0, d]$. Note that $[0, d]$ is the smallest event containing ω that is both the union of partition cells of agent 1 (and hence self-evident to 1) and also the union of partition cells of player 2 (and hence self-evident to 2).

How can we formally capture the idea of i reasoning about the reasoning of j? For any event F, denote by $P_j(F)$ the set of all states that j might think are possible if the true state of the world were somewhere in F. That is, $P_j(F) = \bigcup_{\omega' \in F} P_j(\omega')$. Note that F is self-evident to j if and only if $P_j(F) = F$. Recall that for any ω, $P_i(\omega)$ is simply a set of states, that is, it is itself an event. Hence we can write formally that at ω, i knows that j knows that the event G occurs if and only if $P_j(P_i(\omega)) \subset G$. The set $P_i(\omega)$ contains all worlds ω' that i

believes are possible when the true world is ω, so i cannot be sure at ω that j knows that G occurs unless $P_j(P_i(\omega)) \subset G$.

The framework of Ω and the partitions (P_i) for the agents $i \in I$ also permits us to formalize the idea that at ω, i knows that j knows that k knows that some event G occurs by the formula $P_k(P_j(P_i(\omega))) \subset G$. (If $k = i$, then we say that i knows that j knows that i knows that G occurs.) Clearly there is no limit to the number of levels of reasoning about each others' knowledge that our framework permits by iterating the P_i correspondences. In this framework we say that the state ω' is reachable from ω if and only if there is a sequence of agents $i, j, \ldots, k$ such that $\omega' \in P_k \ldots (P_j(P_i(\omega)))$, and we interpret that to mean that i thinks that j may think that... k may think that ω' is possible.

> *Definition*: The event $E \subset \Omega$ is *common knowledge* among agents $i = 1, \ldots, I$ at ω if and only if for any n and any sequence $(i_1, \ldots, i_n)$, $P_{i_n}(P_{i_{n-1}} \cdots (P_{i_1}(\omega))) \subset E$.

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 an infinite number of conditions must be checked to verify that E is common knowledge. 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 (see also Milgrom, 1981). Recall that an event E is self-evident to i if and only if $P_i(E) = E$, and hence if and only if E is the union of some of i's partition cells. Since there are comparatively few such unions, the collection of self-evident events to a particular agent i is small. The collection of events that are simultaneously self-evident to all agents i in I is much smaller still. This can be expressed as the:

> *Characterizing Common Knowledge Theorem*: Let P_i, $i \in I$, be possibility correspondences representing the (partition) knowledge of individuals $i = 1, \ldots, I$ defined over a common state space Ω. Then the event E is common knowledge at ω if and only if $M(\omega) \subset E$, where $M(\omega)$ is set of all states reachable from ω. Moreover, $M(\omega)$ can be described as the smallest set containing ω that is simultaneously self-evident to every agent $i \in I$.

Since self-evident sets are easy to find, it is easy to check whether an event E is common knowledge at ω. In our three puzzles, the set $M(\omega)$ appears as the connected component of the graph that contains ω. The event E is common knowledge at ω if and only if it contains $M(\omega)$.

A state of nature so far has described the prevailing physical situation; it also describes what everybody knows, and what everybody knows about what everybody knows, and so on. We now allow each state to describe what everybody does. Indeed, in the three puzzles given so far, each state did specify at each time what each agent does. Consider the opinion puzzle. For all ω

between 1 and 8, at first agent 2 thought the expectation of x was -1, while at $\omega = 9$, he thought the expectation of x was 17. By the last time period, he thought at ω between 1 and 3, the expectation of x was 1, at $\omega = 4$ it was -7 and so on. We now make the dependence of action on the state explicit. Let A_i be a set of possible actions for each agent i. Each ω thus specifies an action $a_i = f_i(\omega)$ in A_i for each agent i in I.

Having associated actions with states, it makes sense for us to rigorously describe whether at ω i knows what action j is taking. Let a_j be in A_j, and let E be the set of states at which agent j takes the action a_j. Then at ω, i knows that j is taking the action a_j if and only if at ω i knows that E occurs. Similarly, we say that at ω it is common knowledge that j is taking the action a_j if and only if the event E is common knowledge at ω.

Let us close this section by noting that we can think of the actions an agent takes as deriving from an action *rule* that prescribes what to do as a function of any information situation that agent might be in. The first girl could not identify her hat color because she thought both *RRR* and *WRR* were possible states. Had she thought that only the state *RRR* was possible, she would have known her hat color. The second detective expected x to be -1 because that was the average value x took on the states $\{1, 2, 3, 4\}$ that he thought were possible. Later, when he thought only $\{1, 2, 3\}$ were possible, his expectation of x became 1. Both the girl and the detective could have answered according to their action rule for any set of possible states.

Common Knowledge of Actions Negates Asymmetric Information about Events

The decision rules in our three puzzles all satisfy the sure-thing principle, which runs like this for the opinion game: If the expectation of a random variable is equal to a conditional on the state of nature lying in E, and similarly if the expectation of the same random variable is also a conditional on the state lying in F, and if E and F are disjoint, then the expectation of the random variable conditional on $E \cup F$ is also a. Similarly, if the expectation of a random variable is positive conditional on E, and it is also positive conditional on a disjoint set F, then it is positive conditional on $E \cup F$.[4] In the hat example, the sure-thing principle sounds like this: An agent who cannot tell her hat color if she is told only that the true state of nature is in E, and similarly if she is told it is in F, will still not know if she is told only that the true state is in $E \cup F$. Similarly if she could deduce from the fact that the state lies in E that her hat color is red, and if she could deduce the same thing from the knowledge that the state is in F, then she could also deduce this fact from the knowledge that the state is in $E \cup F$ (even if $E \cap F \neq \varnothing$).

[4]Or in other terms, we say that a decision rule $\psi : 2^{\Omega}/\phi \to A$ satisfies the sure-thing-principle if and only if $\psi(A) = \psi(B) = a$, $A \cap B = \phi$ implies $\psi(A \cup B) = a$.

An Agreement Theorem follows from this analysis, *that common knowledge of actions negates asymmetric information about events*.[5] If agents follow action rules satisfying the sure-thing principle, and if with asymmetric information the agents i are taking actions a_i, and if those actions are common knowledge, then there is symmetric information that would lead to the same actions. Furthermore, if all the action rules are the same, then the agents must be taking the same actions, $a_i = a$ for all i.

To see why the theorem is true, consider the previous diagram in which at ω the information of agent 1, $[0, a]$, is different from the information of agent 2, $[0, b]$. This difference in information might be thought to explain why agent 1 is taking the action a_1 whereas agent 2 is taking action a_2. But if it is common knowledge that agent 1 is taking action a_1 at ω, then that agent 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 that agent 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 $[0, d]$. So the symmetric information $[0, d]$ explains both actions. Furthermore, if the action rules of the two agents are the same, then with the same information $[0, d]$, they must take the same actions, hence $a_1 = a_2$.

The agreement theorem has the very surprising consequence that whenever logically sophisticated agents come to common knowledge (agreement) about what each shall do, the joint outcome does not use in any way the differential information about events they each possess. This theorem 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. Players with common priors and the same information would not bet against each other. The agreement theorem then assures us that even with asymmetric information it cannot be common knowledge that they want to bet (Milgrom and Stokey, 1982).

Similarly, agents who have the same priors will not agree to disagree about the expectation of a random variable. Conditional expectations satisfy the sure-thing principle. Agents with identical priors and the same information would have the same opinion. Hence the agreement theorem holds that they must have the same opinion, even with different information, if those opinions are common knowledge (Aumann, 1976).

[5]A special case of the theorem was proved by Aumann (1976), for the case where the decision rules $\psi_i = \psi =$ 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 $\psi_i = \psi$. However the aspect which I emphasize here (following Geanakoplos (1987)) is that even when the ψ_i are different, and the actions are different, they can all be explained by the same information E.

We now come to the question of how agents reach common knowledge of actions. Recall that each of our three puzzles illustrated what could happen when agents learn over the course of time from the actions of the others. These examples are special cases of a *getting to common knowledge theorem.* Suppose that Ω is finite, and that there are a finite number of agents whose knowledge is defined over Ω, but suppose that time goes on indefinitely. Suppose that at each time period t at least one agent acts, and his action is seen by at least one other agent. Then at some finite time period t^* it will be common knowledge at every ω that each agent can already predict every action he will see in the future. If all the agents see all the actions, then at t^* it will be common knowledge at every ω what all the agents are going to do in the future.

The logic of the getting to common knowledge theorem is illustrated by our examples. Over time the partitions of the agents evolve, getting finer and finer as they learn more. But if Ω is finite, there is an upper bound on the cardinality of the partitions (they cannot have more cells than there are states of nature). Hence after a finite time the learning must stop.

Apply this argument 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 theorem that common knowledge of actions negates asymmetric information, at that point the two agents 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. Hence eventually someone must say no (Sebenius and Geanakoplos, 1983).

The same argument can be applied to the detectives' conversation, or to people expressing their opinions about the probability of some event (Geanakoplos and Polemarchakis, 1982). Eventually it becomes common knowledge what everyone is going to say. At that point they must all say the same thing, as long as the decision rules satisfy the sure-thing principle.

Let us show that the convergence to common knowledge theorem also clarifies the puzzle about the hats. Suppose *RRR* is the actual state and that it is common knowledge (after the teacher speaks) that the state is not *WWW*. Let the children speak in any order, perhaps several at a time, and suppose that each speaks at least every third period, and every girl is heard by everyone else. Then it must be the case that eventually one of the girls knows her hat color. For if not, then by the above theorem it would become common knowledge at *RRR* by some time t^* that no girl was ever going to know her hat color. This means that at every state ω reachable from *RRR* with the partitions that the agents have at t^*, no girl knows her hat color at ω. But since 1 does not know her hat color at *RRR*, she must think *WRR* is possible. Hence *WRR* is reachable from *RRR*. Since 2 does not know her hat color at any state reachable from *RRR*, in particular she does not know her hat color at *WRR*, and so she must think *WWR* is possible there. But then *WWR* is reachable from *RRR*. But then 3 must not know her hat color at *WWR*, hence she must think *WWW* is possible there. But this implies that *WWW* is reachable from *RRR* with the partitions the agents have at time t^*, which contradicts the fact that it is common knowledge at *RRR* that *WWW* is not the real state.

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 theorem that convergence to common knowledge will eventually occur is clearly false. We shall discuss the implications of an infinite state space below.

Bayesian Games

The analysis has considered decision rules which depend on what the agent knows, but not on what the agent expects the other agents to do. This framework was sufficient to analyze the puzzles about the hat color, and the expectation of the random variable x, and also for betting when each son assumed that the bet would be taken (perhaps by the father) as long as he himself gave the OK. But in the envelopes puzzle when the first son realizes that the bet will be taken only if the other son also accepts it, he must try to anticipate the other son's action before deciding on his own action, or else risk that the bet comes off only when he loses. To take this into account we now extend our model of interactive epistemology to include payoffs to each agent depending on any hypothetical action choices by all the agents.

So far the states of nature describe the physical universe, the knowledge of the agents, and the actions of the agent. Implicitly in our examples the states of nature also described the payoffs to the agents of their actions, for this is the motivation for why they took their actions. We now make this motivation more explicit. At each ω, let us associate with every vector of actions of all the I players a payoff to each agent i. In short, each ω defines a game among the I agents. Since the players do not know the state ω, we must say more before we can expect them to decide which action to take. We suppose, in accordance with the Bayesian tradition, that every agent has a prior probability on the states of nature in Ω, and that at ω the agent updates his prior to a posterior by conditioning on the information that ω is in $P_i(\omega)$. This defines a Bayesian game. The agents then choose at each ω the actions which maximize their expected utility with respect to these posterior probabilities, taking the decision rules of the others as given. If the mapping of states to actions satisfies this optimizing condition, then we refer to the entire framework of states, knowledge, actions, payoffs, and priors as a Bayesian Nash equilibrium.

For example, in the last version of the envelopes puzzle the payoffs to the sons depend on what they *both* do. Below we list the payoffs to each son at a state (m, n), depending on whether each decides to bet (B) or to stick with his own envelope and not bet (N).

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

Recall that each son has equal prior probability on every state and that the first son does not know whether the state is $(m, m - 1)$ or $(m, m + 1)$ unless $m = 1$ or $m = 7$. If $m = 7$, the first son knows that he has the highest envelope, $(m > n)$, and therefore in any Bayesian Nash equilibrium he will choose not to switch—that is he will choose N. A similar argument applies to the second son.

But now it is easy to see that in any Bayesian Nash equilibrium of the envelopes game, both sons will refuse to bet at every state (m, n). Since at state $(6, 7)$ the second son must play N in any Bayesian Nash equilibrium, the first son must play N whenever he sees 6, namely at states $(6, 5)$ and $(6, 7)$. But it then follows that the second son must play N whenever he sees 5, and so on. This explains why the very first time the father asks his boys whether they will bet on changing envelopes, knowing that the bet is off unless it is agreed upon by both sides, they will both immediately say no.

At this point it is worth emphasizing that the structure of Bayesian Nash equilibrium extends the framework of interactive epistemology that we developed earlier. For example, we can turn the hats puzzle into a Bayesian game by specifying that the payoff to player i if she correctly guesses her hat color is 1, and if she says she does not know her payoff it is 0, and if she guesses the wrong hat color it is $-\infty$. Similarly, in the opinion game (in which the random variable x that the players are guessing about is given) we can define the payoff at ω to any player i if he chooses the action a to be $-(a - x(\omega))^2$. It is well-known from elementary statistical decision theory that a player minimizes the expected squared error by guessing the conditional expectation of the random variable. Hence these payoffs motivate the players in the opinion game to behave as we have described them in our previous analysis.

Bayesian optimal decisions (that is, maximizing expected utility) satisfy the sure-thing principle. Hence an argument similar to that given in the last section proves the following *Agreement Theorem for Bayesian Games*: Suppose that in Bayesian Nash equilibrium it is common knowledge at some ω what actions the players are each taking. Then we can replace the partitions of the agents so that at ω all the agents have the same information, without changing anything else including the payoffs and the actions of the agents at every state, and still be at a Bayesian Nash equilibrium. In particular, any vector of actions that can be common knowledge and played as part of a Bayesian Nash equilibrium with asymmetric information can also be played as part of a Bayesian Nash equilibrium with symmetric information.

The implications of this theorem for game theory are similar to those indicated in the last section. If a game theorist attributes some Bayesian Nash equilibrium behavior to asymmetric information, it must be that the actions the agents are taking are not common knowledge. For if they were, then asymmetric information about events could not give rise to new phenomena.

We now return to the problem of the envelopes. Observe that we showed that there could not be betting in any Bayesian Nash equilibrium of the envelopes example, even when it might not have been common knowledge that the bet would come off. This result cannot be derived from the above theorem, which presumes common knowledge of actions. Yet the result is so general that we indicate why it must be true. To this end we invoke a second property of Bayesian optimal decisions, namely that more information cannot hurt. In Bayesian Nash equilibrium a player at a state ω need only consider the consequences of action at states ω' in $P_i(\omega)$. Nevertheless, we as the outside observer, can calculate what each player's expected payoff is by taking the average payoff of the player over all the states, weighted by the player's prior probability. It is easy to see that if we replaced player i's partition by the "know nothing" partition $\{\Omega\}$, then (assuming that the other player's actions stayed fixed) the best i could do would not increase i's expected payoff. (If by choosing some constant action i improved his expected payoff, then it would follow that for at least one partition cell he should have been choosing that action instead of what he did choose).

In the envelopes puzzle, the sum of the payoffs to the two players in any state is uniquely maximized by the action choice (N, N) for both players. A bet wastes at least one dollar, and only transfers money from the loser to the winner. It follows that the sum of the two players' expected payoffs is uniquely maximized when both players choose (N, N) at every state. Now consider any Bayesian Nash equilibrium of the envelopes game. Agent 1 could have chosen N at every ω. Since this was possible with no information, it follows that such a choice (assuming that the other player stuck to his choices at each ω) could not have increased 1's expected utility. But note that the payoff to agent 1 when he plays N does not depend on what the other players do. Hence the expected payoff to player 1 must not go up if all the players switch to N for all ω. But the same argument applied to agent 2 shows that his expected payoff must also not go up if all the players switched to N at every state. Combining these two statements, we see that the sum of the expected payoffs to the two players at the Bayesian Nash equilibrium must be at least as high as it would be if the players always played N, a contradiction, unless the Bayesian Nash equilibrium is to play (N, N) at each state.

Speculation

The cause of financial speculation and gambling has long been put down to differences of opinion. Since the simplest explanation for differences of opinion is differences in information, it was natural to conclude that such differences could explain gambling and speculation. Yet, we now see that such a conclusion was premature.

To understand why, begin by distinguishing between two kinds of speculation. One involves two agents who agree on some contingent transfer of money, perhaps using a handshake or a contract to give some 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 to risk money before knowing what the odds may be (as at a horse race) 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.

We also distinguish speculation from investing or trading. With an investment, both the borrower and lender can benefit. When two people bet with each other, one wins what the other loses. An agent who buys a stock from another will win if the stock price rises, while 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 (perhaps because 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, one agent is borrowing and the other agent is investing. Even with symmetric information, both parties might think they are benefiting from the trade. This is not speculation. It appears, however, that only a small proportion of the trades on the stock market can be explained by such savings/investment reasons. Similarly if one agent trades out of dollars into yen, while another agent is trading yen for dollars, it might be because the first agent plans to travel to Japan and the second agent needs dollars to buy American goods. But since the volume of trade on the currency markets is orders of magnitude greater than the money purchases of goods and services, it would seem that speculation and not transactions demand explains much of this activity.

In this discussion, speculation will mean actions taken purely on account of differences of information. To formalize this idea, suppose that each agent has a status quo action, which does not take any knowledge to implement, and which guarantees that agent a utility in each state independent of what actions the others choose. Suppose also that if every agent pursued the status quo action in every state the resulting utilities would be Pareto optimal. A typical Pareto optimal situation might arise as follows. Risk averse agents (possible with different priors) trade a complete set of Arrow-Debreu state contingent claims for money; one agent promising to deliver in some states and receive money in others, and so on. At the moment the contracts are signed, the agents do not know which state is going to occur, although they will recognize the state once it occurs in order to carry out the payments. After the signing of all the contracts for delivery, but before the state has been revealed, the status quo action of refusing all other contracts is well known to be Pareto optimal.

At a Pareto optimum there can be no further trade, if agents have symmetric information. But now suppose that each agent receives additional information revealing something about which state will occur. If different agents get different information, that would appear to create opportunities for betting, or speculative trade. But if it is common knowledge that the agents want to trade, as occurs when agents bet against each other, then our theorem that common knowledge of actions negates asymmetric information about events implies that the trades must be zero. But even if the actions are not common knowledge, the argument given in the last section proves that there will be no more trade. Since the actions are not common knowledge, what is? Only the facts that the agents are rational—that is, their knowledge is given by partitions, and that they are optimizing, and that the status quo is Pareto optimal. We can therefore summarize our argument from the previous section by stating the *Non-Speculation Theorem*: Common knowledge of rationality and of optimization eliminates speculation.

When it is common knowledge that agents are rational and optimizing, 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. For example, take the two sons with their envelopes, with a maximum of $10 million in one of the envelopes. However, suppose now that the sons are risk averse, instead of risk neutral. Then 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—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 bettors, despite the $1 transaction cost. Once each son opens his envelope, however, the incentive to trade disappears, precisely because of the difference in information! Each son must ask himself what the other son knows that he doesn't.

More generally, consider the envelopes problem where the sons may be risk neutral, but they have different priors on Ω. In the absence of information, many bets could be arranged between the two sons. But it can easily be argued that no matter what the priors, as long as each state got positive probability, after the sons look at their envelopes they will not be able to agree on a bet. The reason is that the sons act only on the basis of their conditional probabilities, and given any pair of priors with the given information structure it is possible to find a single prior, the same for both sons, that gives rise to the conditional probabilities each son has at each state of nature. The priors are then called consistent with respect to the information structure. Again, the message is that adding information tends to suppress speculation, rather than encouraging it, when it is common knowledge agents are rational.

Infinite State Spaces and Knowledge about Knowledge to Level *N*

There are two reasons why we might be led to consider an infinite state space. First, we have assumed so far that each agent i knows the partition of every other agent j. One could easily imagine that i does not know which of several partitions j has. This realistic feature could be incorporated into our framework by expanding the state space, so that each new state specifies the original state and also the kind of partition that j has over the original state space. By defining i's partition over this expanded state space, we allow i not only to be uncertain about what the original state is, but also about what j's partition over the original state space is. (The same device also can be used if i is uncertain about what prior j has over the original state space.) Of course it may be the case that j is uncertain about which partition i has over this expanded state space, in which case we could expand the state space once more. We could easily be forced to do this an infinite number of times. But in a series of papers—including Mertens and Zamir (1985), Fagin, Geanakoplos, Halpern, and Vardi (1992), Gilboa (1986), and Kaneko (1987)—it has been shown that a countable number of expansions is sufficient for describing whether an event contained in the original state space is common knowledge.[6]

A second reason we might consider an infinite state space is that messages can sometimes be lost. Suppose that two airplane fighter pilots are transmitting to each other over their radios. If every message has a chance of being lost, and if the states of the world are meant to be complete descriptions of everything that might happen, then there must be at least as many states as there are messages. If we allow for an arbitrarily large number of messages, then we need an infinite state space.

The assumption that the state space Ω is finite played a crucial role in the theorem that common knowledge of actions must eventually be reached. With an infinite state space, common knowledge of actions may never be reached, and one wonders whether that calls into question our conclusions about agreement, betting, and speculation. The answer is that it does not.

Consider for example the envelopes problem, but with no upper bound to the amount of money the father might put in an envelope. More precisely, suppose that the father chooses $m > 0$ with probability $1/2^m$, and puts $\$10^m$ in one envelope and $\$10^{m+1}$ in the other, and randomly hands them to his sons. Then no matter what $m > 1$ he sees in his own envelope, each son calculates the odds are $1/3$ that he has the lowest envelope, and that therefore in expected terms he can gain from switching. This will remain the case no matter how long the father talks to him and his brother. At first glance this seems to reverse our previous findings. But in fact it has nothing to do with the state space being infinite. Rather it results because the expected number of dollars in each envelope (namely the infinite sum of $(1/2^m)(10^m)$) is infinite. The same proof we gave before shows that with an infinite state space, even if the maximum amount of money in each envelope is unbounded, as long as the expected number of dollars is finite, betting cannot occur. (In the case with an

[6]Curiously, to represent that agent i knows that an event is not common knowledge between agents j and k requires more than a countable expansion of the original state space.

unbounded maximum but finite expected number of dollars in each envelope, one can show that for some large but finite $m = M$ it would not be worth switching envelopes, and then we would be in a situation analogous to the finite case considered earlier where there is a maximum amount of money in any envelope.)

Similarly, it can be shown that even with an infinite state space, if agents talk long enough about the expectation of a random variable (that has finite expectation), then their opinions will converge, although they might not become exactly equal in finite time. (This is a consequence of the martingale convergence theorem.)

However, one consequence of a large state space is that it permits states of the world at which a fact is known by everybody, and it is known by all that the fact is known by all, and it is known by all that it is known by all that the fact is known by all, up to N times, without the fact being common knowledge. When the state space is infinite, there could be for each N a (different) state at which the fact was known to be known N times, without being common knowledge.

The remarkable thing is that iterated knowledge up to level N does not guarantee behavior that is anything like that guaranteed by common knowledge, no matter how large N is. The agreement theorem assures us that if actions are common knowledge, then they could have arisen from symmetric information. But this is far from true for actions that are N-times known, where N is finite. For example, in the opinion puzzle taken from Geanakoplos and Polemarchakis, at state $\omega = 1$, agent 1 thinks the expectation of x is 1, while agent 2 thinks it is -1. Both know that these are their opinions, and they know that they know these are their opinions, so there is iterated knowledge up to level 2, and yet these opinions could not be common knowledge because they are different. Indeed they are not common knowledge, since the agents don't know that they know that they know that these are their respective opinions.

Similarly, in the envelopes puzzle, at the state $(4, 3)$ each son wants to bet, and each son knows that the other want to bet, and each knows that the other knows that they each want to bet, so their desires are iterated knowledge up to level 2. But since they would lead to betting, these desires cannot be common knowledge, and indeed they are not, since the state $(6, 7)$ is reachable from $(4, 3)$, and there the second son does not want to bet. It is easy to see that by letting the maximum envelope contain $\$10^{5+N}$, we could build a state space in which there is iterated knowledge to level N that both agents want to bet at the state $(4, 3)$.

Recall the infinite state space version of the envelopes example just described, where the maximum dollar amount is unbounded. At any state (m, n) with $m > 1$ and $n > 1$, agent 1 believes the probability is $1/3$ that he has the lowest envelope, and agent 2 believes that the probability is $2/3$ that agent 1 has the lowest envelope! If $m > N + 1$, and $n > N + 1$, then it is iterated knowledge at least N times that the agents have these different opinions. Thus, for every N there is a state at which it is iterated knowledge N times that the agents disagree about the probability of the event that 1 has the lowest dollar amount in his envelope. But of course for no state can this be common knowledge.

Another example illustrates the difficulty in coordinating logically sophisticated reasoners. Consider again our airplane fighter pilots, and suppose that the first pilot radios a message to the second pilot telling him where to coordinate their attack. If there is a probability p that any message between pilots is not properly transmitted, then even if the second pilot receives the message, he will know where to attack, but the first pilot will not know that the second pilot knows where to attack, since the first pilot cannot be sure that the message arrived. If the first pilot proceeds with the plan of attacking, then with probability $(1 - p)$ the attack is coordinated, but with probability p he flies in with no protection. Alternatively, the first pilot could ask the second pilot for an acknowledgement of his message. If the acknowledgement comes back, then both pilots know where to attack, and both pilots know that the other knows where to attack, but the second pilot does not know that the first pilot knows that the second pilot knows where to attack. The potential level of iterated knowledge has increased, but has the degree of coordination improved?

Let us consider the strategy of the first pilot. If he attacks whether or not he receives the acknowledgement, then in effect he is ignoring the message and we are back in the previous situation where the second pilot sent no acknowledgement. It would anyway be silly for him to attack if he did not get an acknowledgement, since, based on that fact (which is all he has to go on), the odds are more likely (namely, p compared to $p(1 - p)$) that it was his original message that got lost, rather than the acknowledgement. Presumably the first pilot will attack if he gets the acknowledgement, and not otherwise. The second pilot will attack if he got the message, and not otherwise. The chances are now $(1 - p)^2$ that the attack is coordinated, and $(1 - p)p$ that the second pilot attacks on his own, and there is probability p that neither pilot attacks. (If a message is not received, then no acknowledgement is sent.)

Compared to the original plan of sending one message there is no improvement. In the original plan the first pilot could simply have flipped a coin and with probability p sent no message at all, and not attacked, and with probability $(1 - p)$ sent the original message without demanding an acknowledgement. That would have produced precisely the same chances for coordination and one-pilot attack as the two message plan. (Of course the vulnerable pilot in the two message plan is the second pilot, whereas the vulnerable pilot in the one message plan is the first pilot, but from the social point of view, that is immaterial. It may explain however why tourists who write to hotels for reservations demand acknowledgements about their reservations before going.)

Increasing the number of required acknowledgements does not help the situation. No strategy pair for the two pilots can yield a different outcome from the one message scenario where the first pilot sometimes withholds the message, except to change the identity of the vulnerable pilot. Furthermore, if the pilots are self-interested and do not want to attack if the odds are less than even that the other pilot will also be attacking the same spot, then there is a unique Bayesian Nash equilibrium strategy pair, in which each pilot attacks at the designated spot if and only if he has received every scheduled message. To see this, note that if to the contrary one pilot were required to attack with a threshold of messages received well below the other pilot's threshold, then there would be cases where he would know that he was supposed to attack and that the other pilot was not going to attack, and he would refuse to follow the plan. There is also a difficulty with a plan in which each pilot is supposed to attack once some number k less than the maximum number of scheduled messages (but equal for both pilots) is received. For if the second pilot gets k messages but not the $(k + 1)st$, he would reason to himself that it was more likely that his acknowledgement that he received k messages got lost and that therefore the first pilot only got $(k - 1)$ messages, rather than that the first pilot's reply to his acknowledgement got lost. Hence in case he got exactly k messages, the second pilot would calculate that the odds were better than even that the first pilot got only $k - 1$ messages and would not be attacking, and he would therefore refuse to attack. This confirms that there is a unique Bayesian Nash equilibrium. In that equilibrium, the attack is coordinated only if all the scheduled messages get through. One pilot flies in alone if all but the last scheduled message get though. If there is an interruption anywhere earlier, neither pilot attacks. The chances for coordinated attack decline exponentially in the number of scheduled acknowledgements.

The most extreme plan is where the two pilots agree to send acknowledgements back and forth indefinitely. The unique Bayesian Nash equilibrium is for each pilot to attack in the designated area only if he has gotten all the messages. But since with probability one, some message will eventually get lost, it follows that neither pilot will attack. This is exactly like the situation where only one message is ever expected, but the first pilot chooses with probability one not to send it.

Note that in the plan with infinite messages (studied in Rubinstein, 1989), for each N there is a state in which it is iterated knowledge up to level N that both pilots know where to attack, and yet they will not attack, whereas if it were common knowledge that they knew where to attack, they would indeed attack. This example is reminiscent of the example in which the two brothers disagreed about the probability of the first brother having the lowest envelope. Indeed, the two examples are isomorphic. In the example of the pilots, the states of the world can be specified by where to attack, and by an ordered pair (m, n) designating the number of messages each pilot received. Each pilot knows the number of messages he received, but cannot tell which of two numbers the other pilot received, giving the same staircase structure to the states of the world we saw in the earlier example.

The upshot is that when coordinating actions, there is no advantage in sending acknowledgements unless one side feels more vulnerable, or unless the acknowledgement has a higher probability of successful transmission than the previous message. Pilots acknowledge each other once, with the word "roger," presumably because a one word message has a much higher chance of successful transmission than a command, and because the acknowledgement puts the commanding officer in the less vulnerable position.

Bounded Rationality: Mistakes in Information Processing

Common knowledge of rationality and optimization (interpreted as Bayesian Nash equilibrium) has surprisingly strong consequences. It implies that agents cannot agree to disagree; it implies that they cannot bet (when the bet is common knowledge); and most surprising of all, it banishes speculation. Yet casual empiricism suggests that all of these are common phenomena. This section explores the implications of common knowledge by weakening the maintained hypothesis that agents process information perfectly, which has been subsumed so far in the assumption that knowledge has 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?[7]

There are a number of errors that are typically made by decision-makers that suggest we go beyond the orthodox Bayesian paradigm. Agents often forget, or ignore unpleasant information, or grasp only the superficial content of signals. Many of these mistakes turn on the idea that agents often do not know that they do not know. For example, in the story of Silver Blaze, Sherlock Holmes draws the attention of the inspector to "the curious incident of the dog in the night-time." "The dog did nothing in the night-time," protested the inspector, to which Holmes replied, "That was the curious incident." As another 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.

We can model some aspects of non-Bayesian methods of information processing by generalizing the notion of information partition. We begin as usual with the set Ω of states of nature, and a possibility correspondence P mapping each element ω in Ω into a subset of Ω. As before, we interpret $P(\omega)$ to be the set of states the agent considers possible at ω. But now P may not be derived from a partition. For instance, following the ozone example we could imagine $\Omega = \{a, b\}$ and $P(a) = \{a\}$ while $P(b) = \{a, b\}$. A perfectly rational agent who noticed what he did not know would realize when he got the signal

[7]Much of this section is taken from Geanakoplos (1989), which offers a fuller description of possible types of irrationality and how they affect behavior.

$\{a, b\}$ that he had not gotten the signal $\{a\}$ that comes whenever a is the actual state of the world, and hence he would deduce that the state must be b. But in some contexts it is more realistic to suppose that the agent is not so clever, and that he takes his signal at face value.

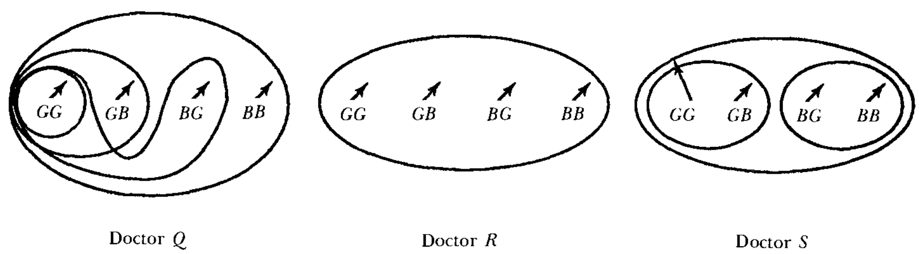
When an agent's information is described by a partition, whether that agent takes information at face value or looks for subtle inferences makes no difference. Partition information is the only situation in which the superficial content of the information is all that there is. Of course the source of an agent's partition information may be that the agent has already teased out all of the subtle messages and interpretations from earlier information that did have subtle content.

Imagine a doctor Q who initially assigns equal probability to all the four states describing whether each of two antibodies is in his patient's blood (which is good) or not in the blood (which is bad). If both antibodies are in the blood, i.e. if the state is GG, the operation the doctor is contemplating will succeed and the payoff will be 3. But if either is missing—if the state is any of GB, BG, or BB—the operation will fail and the doctor will lose 2. Suppose that laboratory assistants are looking for the antibodies in blood samples. The doctor does not realize that if an antibody is present, the lab will find it, whereas if it is missing, the lab will never conclusively discover that it is not there. The knowledge of doctor Q is therefore described by Figure 2. The laboratory never makes an error, always reporting correct information. However, though the doctor does not realize it, the way his laboratory and he process information is tantamount to recognizing good news, and ignoring unpleasant information.

A lab technician comes to the doctor and says that he has found the first antibody in the blood, and his proof is impeccable. Should the doctor proceed with the operation? Since doctor Q takes his information at face value, the doctor will assign equal probability to GG and GB, and decide to go ahead. Indeed for each of the signals that the doctor could receive from the actual states GG, GB, BG the superficial content of the doctor's information would induce going ahead with the operation. (If the actual state of the world is BB, doctor Q will get no information from the lab and will decide not to do the operation.) Yet a doctor R who had no lab and knew nothing about which state was going to occur, would never choose to do the operation. Doctor R's information is also given by the partition in Figure 2.

Suppose that doctors Q and R know each other's laboratories, and know that they each take their laboratory tests at face value. After looking at their private laboratory tests, the two doctors would be willing to make a wager in which doctor R pays doctor Q the net value the operation turns out to be worth if doctor Q performs it. At each of the states GG, GB, BG doctor Q will decide to perform the operation, and therefore the bet will come off, yet each doctor would feel that he was likely to come out ahead. The uninformed but rational doctor R would in fact come out ahead, since 2 out of every 3 times the operation is performed he will receive 2, while 1 out of every 3 times he will lose 3.

Figure 2



Imagine now another doctor S, contemplating the same operation, but with a different laboratory. Doctor S's lab reveals whether or not the first antibody is in the blood of his patient, except when both antibodies are present, in which case the experiment fails and reveals nothing at all. If doctor S takes his laboratory results at face value, then his information is described by Figure 2. The superficial content of doctor S's information is also impeccable. But if taken at face value, it would lead him to undertake the operation if the state were GB (in which case the operation would actually fail), but not in any other state.

If doctors S and R understood each other's laboratories and the way they drew conclusions from their experiments (but S did not realize that there was more to his information than the superficial message), then doctors S and R would also be willing to sign a bet in which R paid S the net value of the operation if doctor S decides to perform it. Again the rational but less well informed doctor R would come out ahead.

S's information processing errors are less severe than Q's. It is not known by doctor R that the bet is going to come off when doctors S and R or when doctors Q and R, set their wager. Doctor R is put in a position much like that of a speculator who places a buy order, but does not know whether it will be accepted. One can show that there is no doctor R' (with partition information, or even one who made the same kinds of errors as doctor S) who doctor S would bet with and with whom the bet would come off whenever R' agreed. But the same is not true for doctor Q. Consider $R' = \{\{GG, GB, BG\}, \{BB\}\}$. Then R' would agree to bet with Q in precisely the states GG, GB, BG, that Q would agree to bet against R'. We could think of Q and R' shaking hands on a bet, rather than as speculators placing buy orders.

Furthermore, the reader might like to prove to himself that the scientist P worried about the ozone, whose information possibility correspondence was described earlier, would not get lured into any unfavorable bets (provided that the ozone layer was the only issue on which he made information processing errors). Furthermore, it can be shown that none of the four agents P, Q, R, S would agree to disagree with any of the others about the probability of some event.

Geanakoplos (1989) establishes necessary and sufficient conditions for the information processing errors captured by the nonpartitional possibility correspondences to allow for speculation, betting, and agreeing to disagree. There is a hierarchy here. Agents can be a little irrational, and still not speculate, agree to bet, or agree to disagree. But if agents are a little more irrational, they will speculate, but not agree to bet or agree to disagree. If they get still more irrational, they will speculate and agree to bet, but not agree to disagree about the probability of an event. Finally, with still more irrationality, they will speculate, agree to bet, and agree to disagree.

■ *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 also wish to acknowledge funding from Intelligent Systems grant IRI-9015570.

Finally I wish to acknowledge helpful advice on early drafts of this paper from Barry Nalebuff, Timothy Taylor, Carl Shapiro, and Adam Brandenburger.

References

Aumann, R., "Agreeing to Disagree," *The Annals of Statistics*, 1976, *4*, 1236–39.

Aumann, R., "Correlated Equilibrium as an Expression of Bayesian Rationality," *Econometrica*, January 1987, *55*, 1–18.

Aumann, R., "Irrationality in Game Theory," mimeo, 1988.

Aumann, R., "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*, 167–90.

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

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

Fagin, R., J. Geanakoplos, J. Halpern, and M. Vardi, "The Expressive Power of the Hierarchical Approach to Modeling Knowledge and Common Knowledge." From The Proceedings of the 4th Conference on Theoretical Aspects of Reasoning About Knowledge. San Mateo: Morgan Kauffman, 1992, 229–44.

Fagin, R., J. Halpern, Y. Moses, and M. Vardi, "Reasoning about Knowledge," mimeo, IBM, San Jose, California, 1988.

Gamow, G. and M. Stern, "Forty Unfaithful Wives," *Puzzle Math*. The Viking Press, New York, 20–23, 1958.

Gardner, M., *Puzzle from other Worlds*. New York: Vintage Books, 1984.

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.

Geanakoplos, J., "Game Theory without Partitions, and Applications to Speculation and Consensus," Cowles Foundation Discussion Paper 914, Yale University, 1989, forthcoming *Journal of Economic Theory*.

Geanakoplos, J. and H. Polemarchakis, "We Can't Disagree Forever," *Journal of Economic Theory*, October 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 Proceedings of the 3rd Association of Computing Machinery Conference on Principles of Distributed Computing, 1984, 50-61.

Kaneko, M., "Structural Common Knowledge and Factual Common Knowledge," RIEE Working Paper 87-27, 1987.

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," *Econometric*, 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.

Milgrom, P. and N. Stokey, "Information, Trade, and Common Knowledge," *Journal of Economic Theory*, February 1982, *26*:1, 17-27.

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

Parikh, R. and P. Krasucki, "Communication, Consensus and Knowledge," *Journal of Economic Theory*, October 1990, *52*:1, 178-89.

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," *Journal of Economic Theory*, October 1990, *52*:1, 190-207.

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*, June 1983, *78*:382, 424-26.

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

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