Cowles Foundation Paper 882

Chapter 40

COMMON KNOWLEDGE

JOHN GEANAKOPLOS*

Yale University

Contents

1. Introduction 1438
2. Puzzles about reasoning based on the reasoning of others 1439
3. Interactive epistemology 1441
4. The puzzles reconsidered 1444
5. Characterizing common knowledge of events and actions 1450
6. Common knowledge of actions negates asymmetric information about events 1453
7. A dynamic state space 1455
8. Generalizations of agreeing to disagree 1458
9. Bayesian games 1461
10. Speculation 1465
11. Market trade and speculation 1467
12. Dynamic Bayesian games 1469
13. Infinite state spaces and knowledge about knowledge to level N 1476
14. Approximate common knowledge 1480
15. Hierarchies of belief: Is common knowledge of the partitions tautological? 1484
16. Bounded rationality: Irrationality at some level 1488
17. Bounded rationality: Mistakes in information processing 1490

References 1495

*About 60% of the material in this survey can be found in a less technical version "Common Knowledge" that appeared in the *Journal of Economic Perspectives*. 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 with to acknowledge funding from computer science grant IRI-9015570. Finally I wish to acknowledge helpful advice on early drafts of this paper from Barry Nalebuff, Tim Taylor, Carl Shapiro, Adam Brandenburger, and Yoram Moses.

Handbook of Game Theory, Volume 2, Edited by R.J. Aumann and S. Hart

1. 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 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 Moriarity'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? In this question connected to whether rational agents will speculate in the stock market? How might the 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, each one knows that the others know it, 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. This definition of common knowledge was suggested by the philosopher D. Lewis in 1969. 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. The main conclusion is that an apparently innocuous assumption of common knowledge rules out speculation, betting, and agreeing to disagree. To try to restore the conventional understandings of these phenomena we allow for infinite state spaces, approximate common knowledge of various kinds including knowledge about knowledge only upto level n, and bounded rationality. We begin this survey with several puzzles that illustrate the strength of the common knowledge hypothesis.

2. 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. The earliest version that I could find appears in Littlewood's Miscellania, (edited by Bollobás) 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.[1]

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 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

[1] These versions are so well-known that it is difficult to find out who told them first. The hats version 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–Polemarchakis (1982). It appeared in the computer science literature in Halpern–Moses (1984).

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 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

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

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 the action plans of the other agents.

3. 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 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. For each $\omega \in \Omega$, we define $P_i(\omega) \subset \Omega$ as all states that agent i cannot distinguish from ω.

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.

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 $\omega \in K_i(E)$. 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. 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) $K_i K_i(A) = K_i(A)$. 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 later. Note that axiom (5) implies axiom (4); $K_i(K_iA) = K_i(-(-K_iA)) = K_i(-(K_i(-(K_iA))) = -K_i(-K_iA)) = -(-K_iA)) = K_iA$.

The most interesting events in the knowledge operator approach are the fixed point events E that satisfy $K_i(E) = E$. From axiom (4), these events make up the range of the $K_i: 2^\Omega \to 2^\Omega$ operator. Axioms (1)–(4) are analogous to the familiar properties of the "interior operator" defined on topological spaces, where Int E is the union of all open sets contained in E. To verify that $(\Omega, \text{Range } K_i)$ is a topological space, we must check that Ω itself is in Range K_i [which follows from axiom (1)], that the intersection of any two elements of Range K_i is in Range K_i [which follows from axiom (2)], and that the arbitrary union $E = \bigcup_{\alpha \in I} E_\alpha$ of sets E_α in Range K_i is itself in Range K_i. To see this, observe that by axiom (2), for all $\alpha \in I$,

$$E_\alpha = K_i(E_\alpha) = K_i(E_\alpha \cap E) = K_i(E_\alpha) \cap K_i(E) \subset K_i(E)$$

hence $E = \bigcup_{\alpha \in I} E_\alpha \subset K_i(E)$, and therefore by axiom (3), $E = K_i(E)$. Thus we have confirmed that $(\Omega, \text{Range } K_i)$ is a topological space, and that for any event $A \subset \Omega$, $K_i(A)$ is the union of all elements of Range K_i that are contained in A.

Axiom (5) gives us a very special topological space because it maintains that if E is a fixed point of K_i, then so is $-E$. The space Range K_i is a complete field, that is, closed under complements and arbitrary intersections. Thus the topological space $(\Omega, \text{Range } K_i)$ satisfies the property that every open set is also closed, and vice versa. In particular, this proves that an arbitrary intersection of fixed point events of K_i is itself a fixed point event of K_i. Hence the minimal fixed point events of K_i form a partition of Ω.

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. For all $\omega \in \Omega$, define $P_i(\omega)$ as the intersection of all fixed point events of the operator K_i that contain ω. By our analysis of the topology of fixed point events, $P_i(\omega)$ is the smallest fixed point event of the K_i operator that contains ω. It follows that the sets $P_i(\omega)$, $\omega \in \Omega$, form a partition of Ω. We must now check that P_i generates K_i, that is we must show that for any $A \subset \Omega$, $K_i(A) = \{\omega \in A: P_i(\omega) \subset A\}$. Since $K_i(A)$ is the union of all fixed point events contained in A, $\omega \in K_i(A)$ if and only if there is a fixed point event E with $\omega \in E \subset A$. Since $P_i(\omega)$ is the smallest fixed point event containing ω, we are done.

We can model an agent's 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, {{female}, {male}}.

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.

Suppose now that agent i knows the partition of j, i.e., suppose that i knows what j is able to know, and vice versa. (This does not mean that i knows what j knows; i may know that j knows her hair color without knowing it himself.) Since the possibility correspondences are functions of the state of nature, 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.

4. The puzzles reconsidered

With this framework, let us 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
PLAYER	1	R	R	R	R	W	W	W	W
	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: $P_1 = \{\{a, e\}, \{b, f\}, \{c, g\}, (d, h\}\}$, $P_2 = \{\{a, c\}, \{b, d\}, (\{e, g\}, \{f, h\})$, $P_3 = \{\{a, b\}, \{c, d\}, \{e, f\}, \{g, h\}\}$.

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 $\omega = a = RRR$ 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 agent i 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 Diagram 1a.[3] Think of R

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

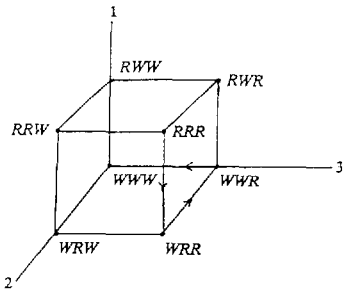


Diagram 1a

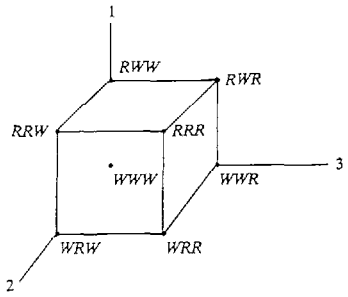


Diagram 1b

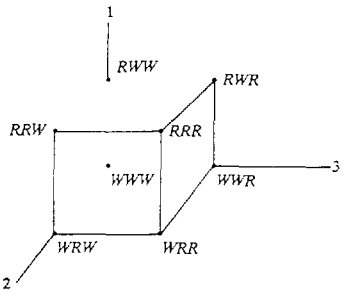


Diagram 1c

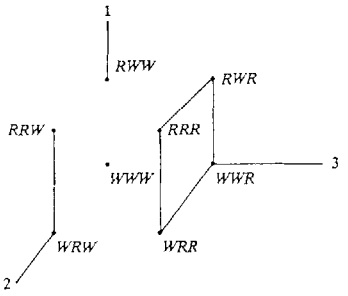


Diagram 1d

as 1 and W as 0. Then every corner of a cube has three coordinates which are either 1 or 0. Let the ith coordinate denote the hat color of the ith 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 ith 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 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 Diagram 1b. (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. Diagram 1c 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 diagram 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 student's information partitions by severing all connections between the set $\{WWW, RWW, RRW, WRW\}$ and its complement, producing Diagram 1d. 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

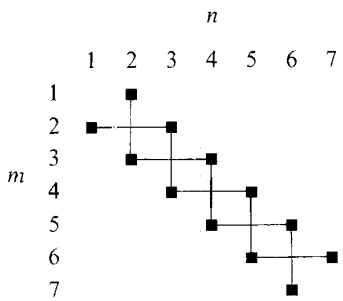
knows what each student is going to say, and so on. By logic alone the students come to a common understanding of what must happen in the future. Note also that at the final stage of information, the three girls have different information.

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–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.

The partitions divide the state space into two components, namely those states reachable from (2, 1) and those states reachable from (1, 2). In one connected component of mutually reachable states, agent 1 has an even number and 2 has an odd 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. This component of the state space is highlighted by the staircase where each step connects two states that agent 1 cannot distinguish, and each rising connects two states that agent 2 cannot distinguish. 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 realize that he would be giving up \$1 million for only \$100 000. Hence if the bet still stands after the second instant, both sons conclude that the state does 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 in choosing to bet the sons had to ante a dollar knowing that the bet would be cancelled and the dollar lost if the other son refused to bet in the same round, 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 Geankoplos 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.

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

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:

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

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:

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

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:

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

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:

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

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 suggests 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.

$a \quad b$

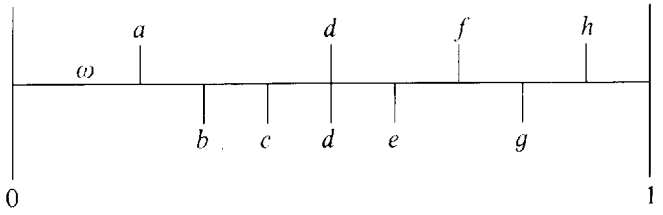
$c \quad d$

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 Geanakoplos and Sebenius (1983), if instead of giving their opinions of the expectation of x, the agents in the last two examples 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. When there are three agents, a bet can be represented by a vector $x(\omega) = (x_1(\omega), x_2(\omega), x_3(\omega))$, denoting the payoffs to each agent, such that $x_1(\omega) + x_2(\omega) + x_3(\omega) \leqslant 0$. If each agent i is asked in turn whether the expectation of x_i is positive, one agent will eventually say no. Thus eventually the agents will give different answers to different questions, as in the hats example. Nevertheless, in the next three sections we shall show how to understand all these examples in terms of a general process of convergence to "agreement."

5. 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 $(0, 1]$ represent Ω. The upper subintervals with endpoints $\{0, a, d, f, h, 1\}$ represent agent 1's partition. The lower subintervals with endpoints $\{0, b, c. d, e, g, 1\}$ 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 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 iff $P_j(P_i(\omega)) \subset G$. 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 $P_j(P_j(\omega)) \subset G$.

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 iff there is a sequence of agents $i, j, \ldots, k$ such that $\omega' \in P_k \cdots (P_j(P_i(\omega)))$, and we interpret that to mean that i thinks that j may think that $\ldots k$ 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 $(i_1, \ldots, i_n)$, $P_{i_n}(P_{i_{n-1}} \cdots (P_{i_1}(\omega))) \subset E$, or equivalently, $\omega \in K_{i_1}(K_{i_2} \cdots (K_{i_n}(E)))$.

This formal definition of common knowledge was introduced 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) showed 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 iff $P_i(E) = E$, and hence iff 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. An event that is simultaneously self-evident to all agents i in I is called a public event. The collection of public events is much smaller still.

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 $i \in I$. In short, E is common knowledge at ω if and only if there is a public event occurring at ω that entails E.

Proof. Let $M(\omega) = \bigcup_n \bigcup_{i_1,\ldots,i_n} P_{i_1} P_{i_2} \cdots P_{i_n}(\omega)$, where the union is taken over all strings $i_1, \ldots, i_n \in I$ of arbitrary length. Clearly E is common knowledge at ω if and only if $M(\omega) \subset E$. But notice that for all $i \in I$, $P_i(M(\omega)) = P_i(\bigcup_n \bigcup_{i_1,\ldots,i_n} P_{i_1} P_{i_2} \cdots P_{i_n}(\omega)) = \bigcup_{n+1} \bigcup_{i_1,\ldots,i_n} P_i P_{i_1} P_{i_2} \cdots P_{i_n}(\omega) \subset M(\omega)$, so $M(\omega)$ is self-evident for each i. □

Before leaving the characterization of common knowledge we define the meet M of the partitions $(P_i, i \in I)$ as the finest partition that is coarser than every P_i. (M is coarser than P_i if $P_i(\omega) \subset M(\omega)$ for all $\omega \in \Omega$; M is finer if the reverse inclusion holds.) To see that the meet exists and is unique, let us define the complete field $\mathcal{F}$ associated with any partition Q as the collection of all self-evident events, that is, the collection of all unions of the cells in Q. [A complete field is a collection of subsets of Ω that is closed under (arbitrary) intersections and complements.] Every complete field $\mathcal{F}$ defines a partition Q where $Q(\omega)$ is the intersection of all the sets in $\mathcal{F}$ that include ω. Given the partitions $(P_i, i \in I)$, let the associated complete fields be $(\mathcal{F}_i, i \in I)$, and define $\mathcal{F} = \bigcap_{i \in I} \mathcal{F}_i$ as the collection of public events. Since the intersection of complete fields is a complete field, $\mathcal{F}$ is a complete field and associated with $\mathcal{F}$ is the meet M of the partition $(P_i, i \in I)$. Clearly $M(\omega)$ is the smallest public event containing ω. Hence we have another way of saying that the event E is common knowledge at ω: at ω the agent whose knowledge is the meet M of the partitions $(P_i, i \in I)$ knows E.

Since self-evident sets are easy to find, it is easy to check whether the event E is common knowledge at ω. In our three puzzles, the public event $M(\omega)$ appears as the connected component of the graph that contains ω. An event E is common knowledge at ω iff 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 etc. 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 was ready to announce the expectation of x was -1, while at $\omega = 9$, he was ready to announce the expectation of x was 17. By the last time period, he was ready to announce 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 iff 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 iff the event E is common knowledge at ω.

Let us close this section by noting that we can think of the actions an agent i takes as deriving from an *external action rule* $\psi_i: 2^{\Omega}/\phi \to A_i$ that prescribes what to do as a function of any information situation he 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.

6. Common knowledge of actions negates asymmetric information about events

The external action 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 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 his hat color if he is told only that the true state of nature is in E, and similarly if he is told it is in F, will still not know if he is told only that the true state is in $E \cup F$. Similarly if he could deduce from the fact that the state lies in E that his hat color is red, and if he could deduce the same thing from the knowledge that the states is in F, then he could also deduce this fact from the knowledge that the state is in $E \cup F$. (Note that we did not use the fact that E intersection F is empty).

An *Agreement Theorem* follows from this analysis, *that common knowledge of actions negates asymmetric information about events.* If agents follow action rules satisfying the sure-thing principle, and if with asymmetric information the agents

[4]Or in other terms, we say that an external action rule $\psi: 2^{\Omega}/\phi \to A$ satisfies the sure-thing principle iff $\psi(A) = \psi(B) = a$, $A \cap B = \phi$ implies $\psi(A \cup B) = a$. If Ω is infinite we require that $\psi(\bigcup_{\alpha} E_{\alpha}) = a$ whenever the E_{α} are disjoint, and $\psi(E_{\alpha}) = a$ for all α in an arbitrary index set. The sure-thing principle could have this interpretation in the detectives example: if a detective would have arrested the butler if the blood type turned out to be A, given his other clues, and if he also would have arrested the butler if the blood type turned out to be O, given those other clues, then he should arrest the butler as soon as he finds out the blood type must be A or O, given those other clues.

i are taking actions a_i, then if those actions are common knowledge, 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.

Theorem. *Let $(\Omega, (P_i, A_i, f_i)_{i\in I})$ be given, where Ω is a set of states of the world, P_i is a partition on Ω, A_i is an action set, and $f_i: \Omega \to A_i$ specifies the action agent i takes at each $\omega\in\Omega$, for all $i\in I$. Suppose that f_i is generated by an action rule $\psi_i: 2^\Omega \to A_i$ satisfying the sure-thing-principle. [Thus $f_i(\omega) = \psi_i(P_i(\omega))$ for all $\omega\in\Omega$, $i\in I$.] If for each i it is common knowledge at ω that f_i takes on the value a_i, then there is some single event E such that $\psi_i(E) = a_i$ for every $i\in I$.*[5]

Corollary. *Under the conditions of the theorem, if $\psi_i = \psi$ for all i, then $a_i = a$ for all i.*

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

To illustrate the theorem, 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 or more players with common priors want to bet with each other, even though they have different information. Choosing to bet (which

[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 is that even when the ψ_i are different, and the actions are different, they can all be explained by the same information E.

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 (1986)].

7. A dynamic state space

We now come to the question of how agents reach common knowledge of actions. Recall that each of our three puzzle 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*, which we state loosely as follows. Suppose that the state space 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. If all the agents see all the actions, then at some finite time period 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 be 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, since opining is an action rule which satisfies 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 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 in the next section, and then again later.

It is useful to describe the dynamic state space formally. Let T be a discrete set of consecutive integers, possibly infinite, denoting calendar dates. We shall now consider an expanded state space $\bar{\Omega} = \Omega \times T$. 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 $\bar{E}$ contained in $\bar{\Omega}$ now specifies what happens at various dates. The simplest events are called dated events and they take the form $\bar{E} = E \times \{t\}$ for some calendar time y, 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 $\bar{P}_i$ of $\bar{\Omega}$. We shall always suppose that agent i is aware of the time, i.e., we suppose that if (ω', t') is in $\bar{P}_i(\omega, t)$, 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., $P_{it}(\omega) = \{\omega' \in \Omega : (\omega', t) \in \bar{P}_i(\omega, t)\}$. 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 at that at some (ω, t), agent i knows that a (dated) event $E = E \times \{t'\}$ 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 E occurs at time t' iff $E = \{\omega: (\omega, t') \in E\}$ is common knowledge with respect to the information partitions P_{it}, i in I.

We now describe how actions and knowledge co-evolve over time. Let A_i be the action space of agent i, for each $i \in I$. Let S_i be the signal space of agent i, for each i in I. Each player $i \in I$ receives a signal at each time $t \in T$, depending on all the actions taken at time t and the state of nature, given by the function $\sigma_{it}: A_1 \times \cdots \times A_I \times \Omega \to S_i$. At one extreme σ_{it} might be a constant, if i does not observe any action. At the other extreme, where $\sigma_{it}(a_1, \ldots, a_I, \omega) = (a_1, \ldots, a_I)$, i observes every action. If some action is observed by all the players at every state, then we say the action is public. If a_{it} depends on the last term ω, without depending at all on the actions, then agent i does not observe the actions, but he does learn something about the state of the world. If each agent whispers something to the person on his left, then $\sigma_{it}(a_1, \ldots, a_I, \omega) = a_{i+1}$ (take $I + 1 = 1$).

Agents take actions $f_{it}: \Omega \to A_i$ depending on the state of nature. The actions give rise to signals which the agents use to refine their information. On the other hand, an agent must take the same action in two different states that he cannot distinguish.

We say that $(\Omega, (A_i, S_i), (\sigma_{it}, P_{it}, f_{it}))_{i \in I}^{t \in T}$ is a dynamically consistent model of action and knowledge (DCMAK) iff for all $t \in T$, $i \in I$

(1) $[P_{it}(\omega) = P_{it}(\omega')] \to [f_{it}(\omega) = f_{it}(\omega')]$ and

(2) $[\omega' \in P_{it+1}(\omega)] \Leftrightarrow [\{\sigma_{it}(f_{1t}(\omega), \ldots, f_{It}(\omega), \omega) = \sigma_{it}(f_{1t}(\omega'), \ldots, f_{It}(\omega'), \omega')\}$ and $\omega' \in P_{it}(\omega)]$.

Condition (1) says that an agent can take action only on the basis of his current stock of knowledge. Condition (2) says that an agent puts together what he knows at time t and the signal he observes at time t to generate his knowledge at time $t + 1$.

We can describe condition (2) somewhat differently. Let $g: \Omega \to S$, where g is any function and S any set. Then we say that g generates the partition G of Ω defined by $\omega' \in G(\omega)$ iff $g(\omega') = g(\omega)$. Furthermore, given partitions Q and G of Ω, we define their join $Q \vee G$ by

$$[Q \vee G](\omega) = Q(\omega) \cap G(\omega).$$

If we have a family of partitions $(Q_i, i \in I)$, we define their join $J = \bigvee_{i \in I} Q_i$ by

$$J(\omega) = [\bigvee_{i \in I} Q_i](\omega) = \bigcap_{i \in I} Q_i(\omega),$$

provided that Ω is finite. Note that J is a finer partition than each Q_i in the sense that $J(\omega) \subset Q_i(\omega)$ for all i and ω. But any partition R that is also finer than each Q_i must be finer than J; so J is the coarsest common refinement of the Q_i.

Let Σ_{it} be the partition generated by the function $\omega \to \sigma_{it}(f_{1t}(\omega), \ldots, f_{It}(\omega), \omega)$. Then condition (2) becomes

$$P_{it+1} = P_{it} \vee \Sigma_{it}.$$

Noticc that over time the partitions grow finer, that is, cach cell of $P_{i\tau}$ is the disjoint union of cells in P_{it} if $\tau < t$.

We now state a rigorous version of the getting to common knowledge theorem. Let $\#P_{i1}$ denote the number of cells in the first partition P_{i1}.

Theorem. *Let $(\Omega, (A_i, S_i), (\sigma_{it}, P_{it}, f_{it}))_{i\in I}^{t\in T}$ be a dynamically consistent model of action and knowledge. Let $T^* = \sum_{i\in I} (\#P_{i1} - 1)$. Suppose T^* is finite and suppose $T > T^*$. Suppose for all $i\in I$ and $t\in T$, σ_{it} does not depend on ω. Then there is some $t \leqslant T^*$ at which it is common knowledge at every $\omega\in\Omega$ that every agent i knows the signal he will observe in period t. If T is infinite, then there is some finite period t^* at which it is common knowledge at every $\omega\in\Omega$ that each agent already knows all the signals he will receive for all $t \geqslant t^*$. In particular, if some agent's actions are always public, and T is infinite, then at some time t^* it will already be common knowledge what action that agent will take for all $t \geqslant t^*$.*

8. Generalizations of agreeing to disagree

To conclusion that agents with common priors who talk long enough will eventually agrce can be generalized to infinite state spaces in which the opinions may never become common knowledge. Moreover, the convergence does not depcnd on every agent hearing every opinion.

Let π be a probability measure on Ω, and let $x: \Omega \to \mathbb{R}$ be a random variable. For convenience, lct us temporarily assumc that Ω is finite and that $\pi(\omega) > 0$ for all $\omega\in\Omega$. Given any partition P on Ω, we define the random variable $f = E(x|P)$ by $f(\omega) = [1/\pi(P(\omega))]\sum_{\omega'\in P(\omega)} x(\omega')\pi(\omega')$. Notice that if F is the partition generated by f, then $E(f|Q) = f$ if and only if Q is finer than F. If so, then we say f is measurable wrt Q. If Q is finer than P, then $E(E(x|P)|Q) = E(x|P) = E(E(x|Q)|P)$.

A *martingale* (f_t, P_t), $t = 1, 2, \ldots$ is a sequence of random variables and partitions such that f_t is measurable wrt P_t for all t, and P_{t+1} is finer than P_t for all t, and such that for all t,

$$E(f_{t+1}|P_t) = f_t.$$

The martingale convergence theorem guarantces that thc martingale functions must converge, $f_t(\omega) \to f(\omega)$ for all ω, for some function f. The classic case of a martingale occurs when x and the increasingly finer partitions P_t are given, and f_t is defined by $E(x|P_t)$. In that case $f_t \to f = E(x|P_\infty)$ where P_∞ is the join of the partitions $(P_t, t = 1, 2, \ldots)$. Furthermore, if (f_t, P_t) is a martingalc and if F_t is the partition generated by $(f_1, \ldots, f_t)$, then (f_t, F_t) is also a martingale.

The foregoing definitions of conditional expectation, and the martingale convergece theorcm, can be extended without change in notation to infinite state spaces Ω provided that we think of partitions as σ-fields, and convergence $f_t \to f$ as convergence π-almost evcrywhere, $f_t(\omega) \to f(\omega)$ for all $\omega\in A$ with $\pi(A) = 1$. (We

must also assume that the f_t are all uniformly bounded, $|f_t(\omega)| \leqslant M$ for all t, and π-almost all ω.) The join of a family of σ-fields, $\mathscr{F}_i$, $i \in I$, is the smallest σ-field containing the union of all the $\mathscr{F}_i$. We presume the reader is familiar with σ-fields. Otherwise he can continue to assume Ω is finite.

We can reformulate the opinion dialogue described in Geanakoplos and Polemarchakis (1982) in terms of martingales, as Nielsen (1984) showed. Let the DCMAK $(\Omega, (A_i, S_i), (\sigma_{it}, P_{it}, f_{it}))_{i\in I}^{t\in T=N}$ be defined so that for some random variable $x: \Omega \to \mathbb{R}$ and probability π,

$$f_{it} = E(x|P_{it}),$$
$$A_i = S_i = \mathbb{R},$$
$$\sigma_{it}(a, \ldots, a_I, \omega) = (a_1, \ldots, a_I).$$

It is clear that (f_{it}, P_{it}) is a martingale for each $i \in I$. Hence we can be sure that each agent's opinion converges, $f_{it}(\omega) \to f_{i\infty}(\omega) = E[x|P_{i\infty}]$ where $P_{i\infty} = \vee_{t\in T} P_{it}$.

Let F_{it} be the σ-field generated by the functions $f_{i\tau}$, for $\tau \leqslant t$. Then (f_{it}, F_{it}) is also a martingale, and $f_{i\infty} = E(x|F_{i\infty})$ where $F_{i\infty} = \vee_{t\in T} F_{it}$. If agent j hears agent i's opinion at each time t, then for $\tau > t$, $P_{j\tau}$ is finer than F_{it}. Hence for $\tau > t$,

$$E(f_{j\tau}|F_{it}) = E(E(x|P_{j\tau})|F_{it}) = E(x|F_{it}) = f_{it}.$$

Letting $t \to \infty$, we get that $E(f_{j\infty}|F_{i\infty}) = f_{i\infty}$, from which it follows that the variance $\mathrm{Var}(f_{j\infty}) > \mathrm{Var}(f_{i\infty})$ unless $f_{j\infty} = f_{i\infty}$ (π-almost everywhere). But since i hears j's opinion at each period t, the same logic shows also that $\mathrm{Var}(f_{i\infty}) > \mathrm{Var}(f_{j\infty})$ unless $f_{i\infty} = f_{j\infty}$. We conclude that for all pairs, $f_{i\infty} = f_{j\infty}$. Thus we have an alternative proof of the convergence theorem in Geanakoplos–Polemarchakis, which also generalizes the result to infinite state space Ω.

The proof we have just given does not require that the announcements of opinions be public.

Following Parikh and Krasucki (1990), consider $I < \infty$ agents sitting in a circle. Let each agent whisper his opinion (i.e., his conditional expectation of x) in turn to the agent on his left. By our getting to common knowledge theorem if Ω is finite, then after going around the circle enough times, it will become common knowledge that each agent knows the opinion of the agent to his immediate right. (Even if Ω is infinite, the martingale property shows that each agent's own opinion converges.) It seems quite possible, however, that an agent might not know the opinion of somebody two places to his right, or indeed of the agent on his left to whom he does all his speaking but from whom he hears absolutely nothing. Yet all the opinions must eventually be the same, and hence eventually every agent does in fact know everybody else's opinion.

To see this, observe that if in the previous proof we supposed $\sigma_{it}(a_1, \ldots, a_I, \omega) = a_{i+1,t}$, then we could still deduce that $E(f_{i\infty}|F_{i+1,\infty}) = f_{i+1,\infty}$, and hence $\mathrm{Var}(f_{i\infty}) > \mathrm{Var}(f_{i+1,\infty})$ unless $f_{i\infty} = f_{i+1,\infty}$. But working around the circle, taking $I + 1 = 1$, we get that $\mathrm{Var}(f_{1\infty}) > \cdots > \mathrm{Var}(f_{1\infty})$ unless all the $f_{i\infty}$ are the same.

The reader may wonder whether the convergence holds when the conversation proceeds privately around a circle if the actions f_{it} are not conditional expectations, but are derivable from external action rules $\psi_i: 2^{\Omega} \to A_i$ satisfying the sure-thing principle. Parikh and Krasucki show that the answer is no, even with a finite state space. When Ω is finite, then convergence obtains if the action rule satisfies $A_i = \mathbb{R}$ and if $E \cap F = \phi$, $\psi_i(E \cup F) = \lambda\psi_i(E) + (1-\lambda)\psi_i(F)$ for some $0 < \lambda < 1$.

Following McKelvey and Page (1986), suppose that instead of whispering his opinion to the agent on his left, each agent whispers his opinion to a poll-taker who waits to hear from everybody and then publicly reveals the average opinion of the I 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 Nielsen et al. (1990). Continuing with our martingale framework, let $\sigma_{jt}(a_1, \ldots, a_I, \omega) = (1/I)\sum_{i \in I} a_i$. Let $f_t(\omega) = (1/I)\sum_{i \in I} f_{it}(\omega)$.

From the getting to common knowledge theorem for finite Ω, or the martingale convergence theorem for infinite Ω, we know that $E(x|P_{it}) \to_t f_{i\infty} \equiv E(x|P_{i\infty})$ for all $i \in I$, π-almost everywhere. Hence $f_t \to f_\infty = (1/I)\sum_{i \in I} f_{i\infty}$ π-almost everywhere. Note that f_t is measurable with P_{it+1}, hence $P_{i\infty} = \vee_{t=0}^{\infty} P_{it}$ is finer than the partition $\mathcal{F}$ generated by f_∞ for all $i \in I$. Then

$$\begin{aligned} &E((x - f_\infty)(f_{i\infty} - f_\infty)|\mathcal{F}) \\ &= E(E((x - f_\infty)(f_{i\infty} - f_\infty)|P_{i\infty})|\mathcal{F}) \\ &= E((f_{i\infty} - f_\infty)^2|\mathcal{F}) \geqslant 0, \end{aligned}$$

with equality holding only if $f_{i\infty} = f_\infty$ π-almost everywhere.

It follows that

$$\begin{aligned} 0 &\leqslant \frac{1}{I}\sum_{i \in I} E((x - f_\infty)(f_{i\infty} - f_\infty)|\mathcal{F}) \\ &= E\left((x - f_\infty)\left(\frac{1}{I}\sum_{i \in I}(f_{i\infty} - f_\infty)\right)|\mathcal{F}\right) \\ &= 0, \end{aligned}$$

where equality holds only if $f_{i\infty} = f_\infty$ π-almost everywhere for each $i \in I$. But that is exactly what we wanted to prove.

9. Bayesian games

The analysis so far has considered action rules which depend on what the agent knows, but not on what he 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 agents. 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 $(a_1,\dots,a_I)$ of all the I players a payoff to each agent i. In short, each ω defines a game $\Gamma(\omega)$ 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 each 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 maximizes their expected utility with respect to these posterior probabilities, taking the action 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.

Formally, a (Bayesian) game is a vector $\Gamma = (I, \Omega, (P_i, \pi_i, A_i, u_i)_{i\in I})$ where $I = \{1,\dots I\}$ 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: $A \times \Omega \to \mathbb{R}$, 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 $f = (f_1,\dots,f_I)$ where $\forall i$, f_i: $\Omega \to A_i$ and

(1) $[P_i(\omega) = P_i(\omega')] \to [f_i(\omega) = f_i(\omega')]$, $i = 1,\dots,I$ 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) (called ex post optimality) means that at each ω to which agent i assigns positive probability 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. [Condition (2) is deliberately vacuous when $\pi_i(P_i(\omega)) = 0$.] 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. This is not the place to explain how BNE arises, and hence how player i comes to know f_{-i}.

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 $\omega = (m, n)$, depending on whether each decides to bet (B) or to stick with his own envelope and not bet (N). Note also the dependence of $\Gamma(\omega)$ on ω.

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

We consider two more examples of Bayesian games in which $\Gamma(\omega)$ does not depend on ω. The first is based on the payoff matrix G, called Matching Pennies, given below:

	Left	Right
Top	$1, -1$	$-1, 1$
Bottom	$-1, 1$	$1, -1$

We know that there is a unique mixed strategy Nash equilibrium to G 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 a state space Ω with four elements arranged in a 2×2 matrix. The first player has a partition of the state space consisting of the two rows of Ω. Similarly the second player has a partition of Ω given by the two columns of Ω. Both players have prior 1/4 on each state. Let $\Gamma(\omega) = G$ for all $\omega = \Omega$. This defines the Bayesian game of Matching Pennies. The Bayesian Nash equilibrium for Matching Pennies is for each player to play the move corresponding to what he sees: if player 1 sees Top, he plays Top, etc.

When the games $\Gamma(\omega) = G$ are independent of the state, and there is a common prior $\pi = \pi_i$ on Ω given by a product of individual priors, then a Bayesian Nash equilibrium for Γ gives a slightly different interpretation to behavior from the usual mixed strategy Nash equilibrium for G. 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 and Boge (1979).

When there is a common prior $\pi = \pi_i$, and the games $\Gamma(\omega) = G$ are independent of the state 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, which provides our third example of Bayesian Nash equilibrium.

Each of two agents sees the color of his own light. There are four states: (green, green), (green, red), (red, green), and (red, red). Both players assign prior probability 1/2 to (green, red) and to (red, green) and probability zero to the other two states. In every state the choices (stop and go) and the payoffs are the same:

	Stop	Go
Stop	(1, 1)	(1, 2)
Go	(2, 1)	(0, 0)

This describes the Bayesian Nash game. The Bayesian Nash equilibrium actions for each state are symmetric for each player: Stop if he sees red, Go if he sees green.

In a Bayesian Nash equilibrium it is tautological (and hence common knowledge at every state ω) that each agent's knowledge is described by a partition, and that each agent has a prior probability over the states of the world. I refer to the partition/individual prior representation of knowledge as Bayesian rationality. In a Bayesian Nash equilibrium agents are always optimizing, that is choosing their actions to maximize their conditional expected utility, hence this must be common knowledge. In short, we may describe the situation of Bayesian Nash equilibrium as common knowledge of Bayesian rationality, and of optimization. The Harsanyi doctrine asserts that all agents must have the same prior. (We briefly discuss the merits of this doctrine in a later section.) Accepting the Harsanyi doctrine, let us suppose that the game $\Gamma(\omega) = G$ is the same for all $\omega \in \Omega$. Then, as Aumann (1987) pointed out, common knowledge of rationality and optimization is tantamount to correlated equilibrium.

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 is 0, and if she guesses the wrong hat color her payoff is -infinity. 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.

Nowadays it is conventional wisdom to assert that many phenomena can only be explained via asymmetric information. A buyer and seller of a house may make

peculiar seeming bids and offers, it is suggested, because they have different private information: each knows what the house is worth to him, but not to the other. But our analysis shows that this wisdom depends on there being uncertainty about the actions of the players. If their actions were common knowledge (for example if the bid and offer were common knowledge) then asymmetric information would have no explanatory power. Bayesian optimal decisions (i.e., maximizing expected utility) satisfy the sure-thing principle. Hence an argument similar to that given in the section on common knowledge of actions 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.

Theorem. *Let $(f_1,\ldots,f_I)$ be a Bayesian Nash equilibrium for the Bayesian game $\Gamma=(I,\Omega,(P_i,\pi_i,A_i,u_i)_{i\in I})$. Suppose at ω it is common knowledge that $(f_1,\ldots,f_I)=(a_1,\ldots,a_I)$. Then there are partitions $\hat{P}_i$ of Ω such that $\hat{P}_i(\omega)=\hat{P}_j(\omega)$ for all $i,\ j\in I$ and such that $(f_1,\ldots,f_I)$ is a Bayesian Nash equilibrium for the Bayesian game $\Gamma=(I,\Omega,(\hat{P}_i,\pi_i,A_i,u_i)_{i\in I})$.*

This theorem is surprising and it explains the puzzles discussed earlier. Of course, its application to Bayesian games is limited by the fact that the actions need not be common knowledge in a Bayesian Nash equilibrium (and in these games asymmetric information does have explanatory power. We return to this question later when we discuss games in extensive form). Consider again the Bayesian Nash game with the envelopes. One common knowledge component of the state space Ω consists of all (m,n) with m even and n odd. (The other common knowledge component reverses the parity.) Hence the agreement theorem for Bayesian Nash equilibrium assures us that there cannot be a Bayesian Nash equilibrium in which both brothers *always* choose to bet when m is even and n odd, for if there were, then the brothers would bet against each other with the same information, which is impossible. (Looked at from the point of view of identical information, both would agree that one brother had an expected payment at least as high as the other, so that taking into account the one dollar betting fee, one brother would not want to bet.) On the other hand, this is a trivial result, since we know at a glance that if the second brother sees that he has the maximum number of dollars in his envelope, he will not bet. A much stronger result would be that there is only one Bayesian Nash equilibrium. Since there is one Bayesian Nash equilibrium in which each brother chooses not to bet at every state of the world, this would rule out any Bayesian Nash equilibrium of the envelopes game in which both brothers bet

in even one state. Such a result indeed is true, and we shall prove it later when we discuss speculation. But it cannot be directly derived from the agreement theorem, which itself depends only on the sure-thing principle. It must be derived from another property of Bayesian optimal decisions, namely that more information cannot hurt.

10. 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 speculation from investing. With an investment, there are gains to trade for all parties that can be perceived by all sides when they have the same information. An agent who buys a stock from another will win if the stock price rises dramatically, 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. 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 him a utility 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. In other words, suppose that it is common knowledge that the status quo is Pareto optimal. At a Pareto optimum there can be no further trade, if agents have symmetric information. A typical Pareto optimal situation might arise as follows. Risk averse agents (possibly 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.

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.

Here we must distinguish 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, and that the payoffs do not depend on their own future actions. 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, the payoffs depend partly on the actions of the speculators, and what the agents are doing is not common knowledge. We reserve the term betting for (the first kind of) common knowledge speculation.

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, there will be no more trade. Since the actions are not common knowledge, what is? Only the facts that the agents are rational, i.e., their knowledge is given by partitions, and that they are optimizing, and that the status quo is Pareto optimal.

Nonspeculation theorem. *Common knowledge of rationality and of optimization eliminates speculation. Let* $\Gamma = (I, \Omega, (P_i, \pi_i, A_i, u_i)_{i\in I})$ *be a Bayesian game. Suppose each player* i *in* I *has an action* $z_i \in A_i$ *such that for all* $(f_1, \ldots, f_I)$, $\sum_{\omega\in\Omega} u_i(z_i, f_{-i}(\omega), \omega)\pi_i(\omega) = \bar{u}_i$. *Furthermore, suppose that* $(z_1, \ldots, z_I)$ *yields a Pareto optimal outcome in the sense that if any* $(f_1, \ldots, f_I)$ *satisfies* $\sum_{\omega\in\Omega} u_i(f(\omega), \omega)\pi_i(\omega) \geq \bar{u}_i$ *for all* $i \in I$, *then* $f_j(\omega) = z_j$ *for all* $\omega \in \Omega$, $j \in I$. *Then* Γ *has a unique Bayesian Nash equilibrium* $(f_1^*, \ldots, f_I^*)$ *and* $f_i^*(\omega) = z_i$ *for all* $\omega \in \Omega$, $i \in I$.

Proof. The following lemma needs no proof. We emphasize, however, that it relies on the properties of partitions. □

Lemma (Knowledge never hurts a Bayesian optimizer). *Consider two single-player Bayesian Nash games* $\Gamma_A = (I = \{i\}, \Omega, P_i, \pi_i, A_i, u_i)$ *and* $\Gamma_B = (I = \{i\}, \Omega, Q_i, \pi_i, A_i, u_i)$ *that differ only in that* P_i *is finer than* Q_i. *Let* f_i *be a Bayesian Nash equilibrium for* Γ_A, *and let* g_i *be a Bayesian Nash equilibrium for* Γ_B. *Then*

$$\sum_{\omega\in\Omega} u_i(f_i(\omega), \omega)\pi_i(\omega) \geq \sum_{\omega\in\Omega} u_i(g_i(\omega), \omega)\pi_i(\omega).$$

Indeed the above inequality holds for any g satisfying $[P_i(\omega) = P_i(\omega')] \rightarrow [g_i(\omega) = g_i(\omega')]$ *for all* $\omega, \omega' \in \Omega$.

Proof of nonspeculation theorem. Let $(f_1, \ldots, f_I)$ be a Bayesian Nash equilibrium for Γ. Fix $f_j, j \neq i$, and look at the one-person Bayesian game this induces for player i. Clearly f_i must be a Bayesian Nash equilibrium for this one-person game. From the fact that knowledge never hurts a Bayesian optimizer we conclude that i could not do better by ignoring his information and playing $f_i^*(\omega) = z_i$ for all $\omega \in \Omega$. Hence

$$\sum_{\omega \in \Omega} u_i(f(\omega), \omega)\pi_i(\omega) \geqslant \sum_{\omega \in \Omega} u_i(z_i, f_{-1}(\omega), \omega)\pi_i(\omega) = \bar{u}_i.$$

But this holds true for all $i \in I$. Hence by the Pareto optimality hypothesis, $f_i = f_i^*$ for all $i \in I$. □

In the envelopes example the action z_i corresponds to not betting N. (We are assuming for now that the agents are risk neutral.) The sum of the payoffs to the players in any state is uniquely maximized by the action choice (N, N) for both players. A bet wastes at least a dollar, and only transfers money from the loser to the winner. It follows that the sum of the two players' ex ante *expected* payoffs is uniquely maximized when the two players (N, N) at every state. Hence by the nonspeculation theorem, the unique Bayesian Nash equilibrium of the envelope game involves no betting (N, N) at every state.

11. Market trade and speculation

We define an economy $E = (I, \mathbb{R}_+^L, \Omega, (P_i, U_i, \pi_i, e_i)_{i \in I})$ by a set of agents I, a commodity space $\mathbb{R}_+^L$, a set Ω of states of nature, endowments $e_i \in \mathbb{R}_+^{L\Omega}$ and utilities $U_i: \mathbb{R}_+^L, \times \Omega \rightarrow \mathbb{R}$ for $i = 1, \ldots, I$, and partitions P_i and measures π_i for each agent $i = 1, \ldots, I$. We suppose each U_i is strictly monotonic, and strictly concave.

Definition. A *rational expectations equilibrium* (REE) $(p, (x_i)_{i \in I})$ for $E = (I, \mathbb{R}_+^L, (P_i, U_i, \pi_i, e_i)_{i \in I})$ is a function $p: \Omega \rightarrow \mathbb{R}_{+}^{L}{}_{+}$, such that for each $i \in I$, $x_i \in \mathbb{R}_+^{L\Omega}$ and if $z_i = x_i - e_i$, then

(i) $\sum_{i=1}^{I} z_i = 0$.

(ii) $p(\omega) z_i(\omega) = 0$, for all $i = 1, \ldots, I$, and all $\omega \in \Omega$.

(iii) $[P_i(\omega) = P_i(\omega')$ and $p(\omega) = p(\omega')] \rightarrow [z_i(\omega) = z_i(\omega')]$ for all $i = 1, \ldots, I$, and all $\omega, \omega' \in \Omega$.

(iv) Let $Q(p) = \{\omega: p(\omega) = p\}$. Then $\forall \omega \in \Omega$, and all i, if $e_i(\omega') + y \in \mathbb{R}_+^L$, $\forall \omega' \in P_i(\omega) \cap Q(p(\omega))$, and $p(\omega) y = 0$, then

$$\sum_{\omega' \in P_i(\omega) \cap Q(p(\omega))} U_i(x_i(\omega'), \omega')\pi_i(\omega') \geqslant \sum_{\omega' \in P_i(\omega) \cap Q(p(\omega))} U_i(e_i(\omega') + y, \omega')\pi_i(\omega').$$

The reference to rational in REE comes from the fact that agents use the subtle information conveyed by prices in making their decision. That is, they not only use the prices to calculate their budgets, they also use their knowledge of the function p to learn more about the state of nature. If we modified (iv) above to

(iv') $\sum_{\omega' \in P_i(\omega)} U_i(x_i(\omega'), \omega')\pi_i(\omega') \geqslant \sum_{\omega' \in P_i(\omega)} U_i(e_i(\omega') + y, \omega')\pi_i(\omega')$ for all $i = 1, \ldots, I$, for all $\omega \in \Omega$ and all $y \in \mathbb{R}^L$ with $p(\omega)y = 0$ and $e_i(\omega') + y \geqslant 0\ \forall \omega' \in P_i(\omega)$

then we would have the conventional definition of competitive equilibrium (CE). The following nonspeculation theorem holds for REE, but note for CE. For an example with partition information in which agents do not learn from prices, and so speculate, see Dubey, Geanakoplos and Shubik (1987). We say that there are only speculative reasons to trade in E if in the absence of asymmetric information there would be no perceived gains to trade. This occurs when the initial endowment allocation is ex ante Pareto optimal, that is if $\sum_{i=1}^{I} y_i(\omega) \leqslant \sum_{i=1}^{I} e_i(\omega)$ for all $\omega \in \Omega$, and if for each $i = 1, \ldots, I$, $\sum_{\omega \in \Omega} u_i(y_i(\omega), \omega)\pi_i(\omega) \geqslant \sum_{\omega \in \Omega} u_i(e_i(\omega), \omega)\pi_i(\omega)$, then $y_i = e_i$ for all $i = 1, \ldots, I$.

Theorem (Nonspeculation in REE). *Let $E = (I, \mathbb{R}^L_+, \Omega, (P_i, U_i, \pi_i, e_i)_{i \in I})$ be an economy, and suppose the initial endowment allocation is ex ante Pareto optimal. Let $(p, (x_i)_{i \in I})$ be a rational expectations equilibrium. Then, $x_i = e_i$ for all $i = 1, \ldots, I$.*

This theorem can be proved in two ways. A proof based on the sure-thing principle was given by Milgrom and Stokey (1982). Proofs based on the principle that more knowledge cannot hurt were given Kreps (1977), Tirole (1982), Dubey, Geanakoplos and Shubik (1987).

First proof. Let $A_i = \{B, N\}$, and define $u_i(B, B, \ldots, B, \omega) = U_i(x_i(\omega), \omega)$, and $u_i(a, \omega) = U_i(e_i(\omega), \omega)$ for $a \neq (B, B, \ldots, B)$. This gives a Bayesian Nash game $\Gamma = (I, \Omega, (P_i, \pi_i, A_i, u_i)_{i \in I})$ which must have a Nash equilibrium in which $f_i(\omega) = B\ \forall i \in I$, $\omega \in \Omega$. Since each $f_i = B$ is common knowledge, by the agreement theorem each player would be willing to play B even if they all had the same information, namely knowing only that $\omega \in \Omega$. But that means each agent (weakly) prefers x_i ex ante to e_i, which by the Pareto optimality hypothesis is impossible unless $x_i = e_i$.

A **second proof** based on the principle that knowledge cannot hurt is given by ignoring the fact that the actions are common knowledge, and noting that by playing N at each ω, without any information agent i could have guaranteed himself $e_i(\omega)$. Hence by the lemma that knowledge never hurts a Bayesian optimizer, x_i is ex ante at least as good as e_i to each agent i, and again the theorem follows from the Pareto optimality hypothesis on the e_i. □

It is interesting to consider what can be said if we drop the hypothesis that the endowments are ex ante Pareto optimal. The following theorem is easily derived from the theorem that common knowledge of actions negates asymmetric information about events.

Theorem. *Let* $E = (I, \mathbb{R}^L_+, \Omega, (P_i, U_i, \pi_i, e_i)_{i\in I})$ *be an economy, and suppose* $(p, (x_i)_{i\in I})$ *is a rational expectations equilibrium. Suppose at some* ω *that the net trade vector* $z_i(\omega) = x_i(\omega) - e_i(\omega)$ *is common knowledge for each* i. *Then* P_i *can be replaced by* $\hat{P}_i$ *for each* i *such that* $\hat{P}_i(\omega) = \hat{P}_j(\omega)$ *for all* $i, j \in I$, without disturbing the equilibrium.

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. 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 does not.

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 original (distinct) priors are then called consistent with respect to the information structure. Again, the message is that adding asymmetric information tends to suppress speculation, rather than encouraging it, when it is common knowledge that agents are rational. [See Morris (1991).]

12. Dynamic Bayesian games

We have seen that when actions are common knowledge in (one-shot) Bayesian Nash equilibrium, asymmetric information becomes irrelevant. Recall that a dynamically consistent model of action and knowledge $(\Omega, (A_i, S_i), (\sigma_{it}, P_{it}, f_{it}))^{t\in T}_{i\in I}$ specifies what each agent will do and know at every time period, in every state of nature. Over time the players will learn. From our getting to common knowledge theorem for DCMAK we know that if Ω is finite and the time horizon is long enough, there will be some period t^* at which it is common knowledge what the players will do that period. If the time period is infinite, then there will be a finite time period t^* when it will become common knowledge what each player will do at every future time period t. One might therefore suppose that in a Bayesian

Nash equilibrium of a multiperiod (dynamic) game with a finite state space, asymmetric information would eventually become irrelevant. But unlike DCMAK, dynamic Bayesian Nash equilibrium must recognize the importance of contingent actions, or *action plans* as we shall call them. Even if the immediately occurring actions become common knowledge, or even if all the future actions become common knowledge, the action *plans* may not become common knowledge since an action plan must specify what a player will do if one of the other players deviates from the equilibrium path. Moreover, in dynamic Bayesian games it is common knowledge of action plans, not common knowledge of actions, that negates asymmetric information.[6] The reason is that a dynamic Bayesian game can always be converted into a Bayesian game whose action space consists of the action plans of the original dynamic Bayesian game.

We indicate the refinement to DCMAK needed to describe dynamic Bayesian games and equilibrium. An action plan is a sequence of functions $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{it} \ldots)$ such that $\alpha_{i1} \in A_i$, and for all $t > 1$, α_{it}: $\times_{\tau=1}^{t-1} S_i \to A_i$. At time t, agent i chooses his action on the basis of all the information he receives before period t. Denote by $\mathscr{A}_i$ the space of action plans for agent $i \in I$.

Action plans $(\alpha_i, i \in I)$ generate signals $s(\omega) \in (S_1 \times \cdots \times S_I)^T$ and actions $a(\omega) \in (A_1 \times \cdots \times A_I)^T$ for each $\omega \in \Omega$ that can be defined recursively as follows.

Let $a_{i1}(\omega) = \alpha_{i1}$, and let $s_{i1}(\omega) = \sigma_{i1}(a_{i1}(\omega), \ldots, a_{I1}(\omega), \omega)$ for $\omega \in \Omega$, $i \in I$.

For $t > 1$, let $a_{it}(\omega) = \alpha_{it}(s_{i1}(\omega), \ldots, s_{it-1}(\omega))$ and $s_{it}(\omega) = \sigma_{it}(a_{1t}(\omega), \ldots, a_{It}(\omega), \omega)$ for $\omega \in \Omega$, $i \in I$.

Define payoffs u_i that depend on any sequence of realized actions and the state of the world: u_i: $(A_1 \times \cdots \times A_I)^T \times \Omega \to \mathbb{R}$. We say that the payoffs are additively separable if there are functions v_{it}: $A_1 \times \cdots \times A_I \times \Omega \to \mathbb{R}$ such that for any $a \in (A_1 \times \cdots \times A_I)^T$,

$$u_i(a, \omega) = \sum_{t \in T} v_{it}(a_{1t}, \ldots, a_{It}, \omega).$$

A strategy is a function $\tilde{\alpha}_i$: $\Omega \to \mathscr{A}_i$ such that $[P_{i1}(\omega) = P_{i1}(\omega')]$ implies $[\tilde{\alpha}_i(\omega) = \tilde{\alpha}_i(\omega')]$. We may write $\tilde{\alpha}_i \in \tilde{\mathscr{A}}_i = \mathscr{A}_i^{\text{Range } P_{i1}}$.

Given a probability π_i on Ω, the strategies $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_I)$ give rise to payoffs $U_i(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_I) = \sum_{\omega \in \Omega} u_i(a(\omega), \omega) \pi_i(\omega)$ where $a(\omega)$ is the outcome stemming from the action plans $(\alpha_1, \ldots, \alpha_I) = (\tilde{\alpha}_1(\omega), \ldots, \tilde{\alpha}_I(\omega))$.

A dynamic Bayesian game is given by a vector $\Gamma = (I, T, \Omega, (P_{i1}, \pi_i, A_i, u_i, \sigma_i)_{i \in I})$ A (dynamic) Bayesian Nash equilibrium is a tuple of strategies $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_I)$ such that for each $i \in I$, $\tilde{\alpha}_i \in \text{Arg Max}_{\beta \in \tilde{\mathscr{A}}_i} U_i(\tilde{\alpha}_1, \ldots, \beta, \ldots, \tilde{\alpha}_I)$. Clearly any (dynamic) Bayesian Nash equilibrium gives rise to a dynamically consistent model of action and knowledge. In particular, P_{it} for $t > 1$ can be derived from the agent's action plans and the signals σ_{it}, as explained in the section on dynamic states of nature.

[6] Yoram Moses, among others, has made this point.

Any dynamic Bayesian game Γ and any Bayesian Nash equilibrium $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_I)$ for Γ defines for each t the truncated dynamic Bayesian game $\Gamma_t = (I, T_t, \Omega, (P_{it}, \pi_i, A_i, \bar{u}_i, \sigma_i)_{i \in I})$ where T_t begins at t and the P_{it} are derived from the Bayesian Nash equilibrium. The payoffs $\bar{u}_i : (A_1 \times \cdots \times A_I)^{T_t} \to \mathbb{R}$ are defined on any $b \in (A_1 \times \cdots \times A_I)^{T_t}$ by

$$\bar{u}_i(b, \omega) = u_i(a_1(\omega), \ldots, a_{t-1}(\omega), b, \omega),$$

where $a_1(\omega), \ldots, a_{t-1}(\omega)$ are the Bayesian Nash equilibrium actions played at ω that arise from the Bayesian Nash equilibrium $\tilde{\alpha}$.

We say that a dynamic Bayesian Nash equilibrium of a Bayesian game Γ does not depend on asymmetric information at ω if we can preserve the BNE and replace each P_{i1} with $\hat{P}_{i1}$ in such a way that $\hat{P}_{i1}(\omega)$ is the same for all $i \in I$. (We say the same thing about Γ_t if $\hat{P}_{it}(\omega)$ is the same for all $i \in I$.)

One can imagine an extensive form Bayesian game which has a Bayesian Nash equilibrium in which it is common knowledge at some date t what all the players are going to do in that period, and yet it is not common knowledge at t what the players will do at some subsequent date. In such a game one should not expect to be able to explain the behavior at date t on the basis of symmetric information. The classic example is the repeated Prisoner's Dilemma with a little bit of irrationality, first formulated by Kreps et al. (1982).

The two players have two possible moves at every state, and in each time period, called cooperate (C) and defect (D). The payoffs are additively separable, and 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. An action plan 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 an action plan, called grim, is to defect at all times, no matter what. Tit for tat is to play C at $t = 1$ and for $t > 1$ to play what the other player did at $t - 1$. Trigger is the action plan in which a player plays C until the other player has defected, and then plays D for ever after. Other actions plans typically involve more complicated history dependence in the choices.

It is well-known that the only Nash equilibrium for the T-repeated Prisoner's Dilemma is defection in every period.

Consider again the Prisoner's Dilemma, but now let there be four states of exogenous uncertainty, SS, SN, NS, NN. 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 types are independent across agents, so for example the chance of NS is 4/25. The payoff to a sane agent is as before, but the payoff to an insane agent is 1 if his actions for $1 \leqslant t \leqslant T$ are consistent with the action plan trigger, and 0 otherwise. A strategy must associate an action plan to each partition cell. Let each agent play trigger when insane, and play trigger until time T, when he defects for sure, when sane. The reader can verify that this is a Bayesian Nash equilibrium. For example, let $\omega = SS$. 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.

The getting to common knowledge theorem assures us that so long as $T > (\#P_1 - 1) + (\#P_2 - 1) = (2-1) + (2-1) = 2$, in any Bayesian Nash equilibrium there must be periods t at which it is common knowledge what the agents are going to do. Observe that in this Bayesian Nash equilibrium it is already common knowledge at $t = 1$ what the players are going to do for all $t \leqslant T - 1$, but not at date T. Yet as we have noted, we could not explain cooperative behavior at period 1 in state SS on the basis of symmetric information. If both players know the state is SS, then we are back in the standard repeated Prisoner's Dilemma which has a unique Nash equilibrium – defect in each period. If neither player knows the state, then in the last period by defecting a player can gain 1 with probability 4/5, and lose at most 1 with probability 1/5. Working backwards we see again there can be no cooperation in equilibrium. Thus we have a game where asymmetric information matters, because some *future* actions of the players do not become common knowledge before they occur.

By adding the chance of crazy behavior in the last period alone (the only period N's actions differ from S's actions), plus asymmetric information, we get the sane 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 (1982) showed that if the insane agent is given a strategy that differs from the sane agent's strategy 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. However, as the probability of irrationality goes to zero, the number of periods of nonoptimizing (when N and S differ) behavior must go to infinity.

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 and Geanakoplos (1988) it is shown that

for generic continuous games, like the Cournot game where agents choose 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 credibly committed himself to starvation, might change the behavior of an entire nation.

Even if it is common knowledge at $t = 1$ what the agents will do at every time period $1 \leqslant t \leqslant T$, asymmetric information may still be indispensable to explaining the behavior, if $T > 1$. Suppose for example that player 1 chooses in the first period which game he and player 2 will play in the second period. Player 1 may avoid a choice because player 2 knows too much about that game, thereby selecting a sequence of forcing moves that renders the actions of both players common knowledge. The explanation for 1's choice, however, depends on asymmetric information. Consider the Match game. Let $\Omega = \{1, \ldots, 100\}$ where each $\omega \in \Omega$ has equal probability. Suppose at $t = 1$, agent i must chose L or R or D. If i chooses L, then in period 2 player j must pick a number $n \in \Omega$. If player j matches and $n = \omega$, then player j gets 1 and player i gets -1. Otherwise, if $n \neq \omega$, then player j gets -1 and player i gets 1. If i chooses R, then again player j must choose $n \in \Omega$, giving payoff $n - 100$ to j, and $100 - n$ to agent i, for all ω. If i chooses D, then in period 2 i must choose $n \in \Omega$; if $n = \omega$, then i gets 2 and j gets -1, while if $n \neq \omega$, then i gets -1 and j gets 1.

Suppose finally that $P_{i1} = \{\Omega\}$, while j knows the states, $P_{j1} = \{\{\omega\}, \omega \in \Omega\}$. A Bayesian Nash equilibrium is for i to play R, and for player j to choose $n = 100$ if R, and to choose $n = \omega$ if L. There can be no other outcome in Bayesian Nash equilibrium. Note that it is common knowledge at each state ω before the first move what actions all the agents will take at $t = 1$ and $t = 2$. But i does not know the action plan of agent j. Without knowing the state, i cannot predict what j *would* do if i played L.

Asymmetric information is crucial to this example. If agent j were similarly uninformed, $P_{j1} = \{\Omega\}$, then i would choose L and get an expected payoff of 98/100. If both parties were completely informed, i would choose D and get an expected payoff of 2. Symmetric information could not induce i to choose R.

Despite these examples to the contrary, there are at least two important classes of Bayesian games in extensive form where common knowledge of actions (rather than action plans) negates asymmetric information about events: nonatomic games and separable two-person zero-sum games.

Suppose the action plans of the agents are independent of the history of moves of any single agent. For example, the action plans may be entirely history independent. Or they may depend on a summary statistic that is insensitive to any single agent's action. This latter situation holds when there is a continuum of agents and the signal is an integral of their actions. In any BNE, once it becomes common knowledge at some date t^* what all the agents will do thereafter, the partitions P_{it^*} can be replaced by a common, coarser partition $\hat{P}_{it^*} = \hat{P}_{t^*}$ and each player will still have enough information to make the same responses to the signals he expects to see along the equilibrium path. However, he may no longer have the information to respond according to the BNE off the equilibrium path. But in the continuum of agents situations, no single agent can, by deviating, generate an off-equilibrium signal anyway. Hence if there was no incentive to deviate from the original equilibrium, by the sure-thing principle there can be no advantage in deviating once the information of all the agents is reduced to what is common knowledge. Without going into the details of defining nonatomic (i.e., continuum) games, these remarks can serve as an informal proof of the following informal theorem:

Theorem (Informal). *For nonatomic Bayesian games in extensive form where the state space is finite, if the time horizon is infinite, there will be a time period t^* such that the whole future of the equilibrium path can be explained on the basis of symmetric information. If the time horizon is finite but long enough, and if the payoffs are additively separable between time periods, then there will be a finite period t^* whose equilibrium actions can be explained on the basis of symmetric information in a one-period game.*

The three puzzles with which we began this paper can all be recast as nonatomic games with additively separable payoffs. We can simply replace each agent by a continuum of identical copies. The report that each agent gives will be taken and averaged with the report all his replicas give, and only this average will be transmitted to the others. Thus in the opinion game, each of the type 1 replicas will realize that what is transmitted is not his own opinion of the expectation of x, but the average opinion of all the replicas of type 1. Since a single replica can have no effect on this average, he will have no strategic reason not to maximize the one-shot payoff in each period separately. Similarly we can replace each girl in the hats puzzle with a continuum of identical copies. We put one copy of each of the original three girls in a separate room (so that copies of the same girl cannot see each other). Each girl realizes that when she says "yes, I know my hat color" or "no, I do not know may hat color," her message is not directly transmitted to the other girls. Instead the proportion of girls of her type who say yes is transmitted to the other girls. A similar story could be told about the boys who say yes or no about whether they will bet with their fathers (or with each other).

All three puzzles can be converted into nonatomic games in which the Bayesian

Nash equilibrium generates exactly the behavior we described. The reason this is possible for these three puzzles, but not for the repeated Prisoner's Dilemma or the Match game, is that the behavior of the agents in the puzzles was interpersonally myopic; no agent calculated how changes in his actions at period t might affect the behavior of others in future periods. This interpersonal myopia is precisely what is ensured by the nonatomic hypothesis. By contrast, the repeated Prisoner's Dilemma with a little bit of irrationality hinges entirely on the sane player's realization that his behavior in early periods influences the behavior of his opponent in later periods. In contrast to the puzzles, in the repeated Prisoner's Dilemma game and in the Match game, asymmetric information played an indispensable role even after the actions of the players became common knowledge.

Consider now a sequence of two-person zero-sum games in which the payoff to each of the players consists of the (separable, discounted) sum of the payoffs in the individual games. The game at each time t may depend on the state of nature, and possibly also t. The players may have different information about the state of nature. We call this the class of repeated zero-sum Bayesian games. In the literature on repeated games, the game played at time t is usually taken to be independent of t. We have the same basic theorem as in the nonatomic case:

Theorem. *Consider a (pure strategy) Bayesian Nash equilibrium of a repeated zero-sum Bayesian game with a finite set of states of the world. If the time horizon T is infinite, there will be a time period t^* such that the whole future of the equilibrium path can be explained on the basis of symmetric information. If the time horizon T is finite but $T > T^* = \#P_1 - 1 + \#P_2 - 1$, then there must be some period $t \leqslant T^*$ whose actions can be explained on the basis of symmetric information.*

Proof. In any Bayesian equilibrium the equilibrium strategies define a Bayesian Nash equilibrium for the truncated Bayesian game obtained by considering the time periods from $t+1$ onward beginning with the equilibrium partitions P_{it+1}. Since the games are zero-sum, and the payoffs are additively separable, the fact that player 2 cannot improve his payoff from period $t+1$ onward if 1 sticks to his equilibrium strategy means that player 1 can guarantee his payoff from period $t+1$ onward by sticking to his equilibrium strategy no matter what he does in period t, provided that he does not reveal additional information to player 2. Hence we deduce from the fact that we began with a Bayesian Nash equilibrium, that player 1 cannot find an action function $b(\omega)$ at any time t that improves his expected payoff at time t and that uses (and hence reveals) only information that both players already had at time t. (The reader should note that there could be actions that agent 1 can take on the basis of his own information at time t that would improve his time t payoff that he will not undertake, because those actions would reveal information to player 2 that could be used against player 1 in subsequent periods.)

From our getting to common knowledge theorem we know that there must be

some time $t \leqslant T^*$ such that the actions of the players are common knowledge before they occur, at every state of the world. We can thus find a partition P of the state space that is coarser than the partition P_{it} of each of the agents at time t, such that the action functions $a_{it}(\omega)$ at time t of each of the players i is measurable with respect to P.

It follows from the last two paragraphs that there is some time $t \leqslant T^*$ and an information partition P such that if all the agents had the same information P, their actions a_{it} would form a Bayesian Nash equilibrium for the one-shot game defined at time t. This proves the second part of the theorem.

If the game is infinitely repeated, then there must be a t^* such that at t^* all the current and future equilibrium actions are common knowledge. Hence restricting both the players' actions to some common partition P for all periods t^* and onward will not disturb the equilibrium. □

Aumann and Maschler (1966) considered infinite repeated zero-sum games in which agent i has a finer partition P_{i1} than agent j's partition P_{j1}. They supposed that $\sigma_{kt}(a_i, a_j, \omega) = (a_i, a_j)$, for all $k \in \{1, 2\}$, $t \in T$, and $(a_i, a_j, \omega) \in A_i \times A_j \times \Omega$. They took as the payoffs the limit of the average of the one-shot payoffs, which has the consequence that payoffs in any finite set of time periods do not influence the final payoff.

Consider a (pure strategy) Bayesian Nash equilibrium of an Aumann–Maschler game. At each t, $P_{it} = P_{i1}$, while P_{jt} is intermediate between P_{j1} and P_{i1}. Once t^* is reached at which all subsequent moves are common knowledge, $P_{jt} = P_{jt^*}$ for all $t \geqslant t^*$. From the foregoing theorem, we know that if we replaced P_{it^*} with P_{jt^*}, we would not affect the equilibrium. In fact, since t^* is finite, this equilibrium gives the same payoffs as the game in which $\hat{P}_{i1} = \hat{P}_{j1} = P_{jt^*}$. In effect, player i chooses how much information P_{jt^*} to give player j, and then the two of them play the symmetric information game with partitions P_{jt^*}.

13. Infinite state spaces and knowledge about knowledge to level N

If we allow for random (exogenous) events at each date $t \in T$, such as the possibility that a message (or signal) might fail to be transmitted, 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 time periods. If we allow for an arbitrarily large number of faulty 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. We have already seen via martingale theory that when agents are discussing the expectation of a random variable, their opinions must converge even with an infinite state space. We now turn to betting.

Consider 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 amount he sees in his own envelope, each son calculates the odds are at least 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. On close examination, 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.

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–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 do not know that they know that they know that these are their respective opinions.

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 lower envelope, and agent 2 believes that the probability is 2/3 that agent 1 has the lower envelope! (If $m = 1$, then agent 1 knows he has the lower envelope, and if $n = 1$, agent 2 knows that agent 1 does not have the lower 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 lower dollar amount in his envelope. Moreover, not even the size of the disagreement depends on N. But of course for no state can this be common knowledge.

Similarly, in our original finite state envelopes puzzle, at the state (4, 3) each son wants to bet, and each son knows that the other wants 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 expanding the state space and letting the maximum envelope contain $\$10^{5+N}$, instead of $\$10^7$, 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).

Another example illustrates the difficulty in coordinating logically sophisticated reasoners. Consider two airplane fighter pilots, and suppose that the first pilot radios a message to the second pilot telling him they should attack. If there is a probability $(1-p)$ that any message between pilots is lost, then even if the second pilot receives the message, he will know they should attack, but the first pilot will not know that the second pilot knows they should 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 p the attack is coordinated, but with probability $(1-p)$ 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 they should attack, and both pilots know that the other knows they should attack, but the second pilot does not know that the first pilot knows that the second pilot knows they should attack. The potential level of iterated knowledge has increased, but has the degree of coordination improved? We must analyze the dynamic Bayesian game.

Suppose the pilots are self-interested, so each will attack if and only if he knows they should attack and the odds are at least even that the other pilot will be attacking. Suppose furthermore that the first pilot alone is able to observe whether they should attack. In these circumstances there is a trivial Bayesian Nash equilibrium where neither pilot ever attacks because each believes the other will not attack. If it were common knowledge whether they should attack, then there would be another BNE in which they would both attack when they should, and not when they should not. Unfortunately for the pilots, it can never be common knowledge that they should attack. The only other Bayesian Nash equilibrium is where each pilot attacks if and only if every possible message he might have gotten telling him to attack is received.

The second pilot clearly will not attack if he gets no message, for without the message he could not know that they should attack. At best, he will attack if he gets the message, and not otherwise. He will indeed be willing to do that if he

expects the first pilot to attack if he gets the second pilot's acknowledgement (assuming that $p > 1/2$). Given the second pilot's strategy, the first pilot will indeed be willing to attack if he gets the acknowledgement, since he will then be sure the second pilot is attacking. Thus there is a BNE in which the pilots attack if every message is successfully transmitted. Notice that the first pilot will not attack if he does not get the acknowledgement, since, based on that fact (which is all he has to go on), the odds are more likely [namely $(1-p)$ versus $(1-p)p$] that it was his original message that got lost, rather than the acknowledgement. The chances are now p^2 that the attack is coordinated, and $(1-p)p$ that the second pilot attacks on his own, and there is probability $(1-p)$ 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 $(1-p)$ sent no message at all, and not attacked, and with probability 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. Aside from the trivial BNE, there is a unique Bayesian Nash equilibrium, 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 non-trivial 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 through. If there is an interruption anywhere earlier, neither pilot attacks. The outcome is the same as the one message scenario where the first pilot sometimes

withholds the message, except to change the identity of the vulnerable pilot. 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 non-trivial 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 should attack, and yet they will not attack, whereas if it were common knowledge that they should 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 pilots example, the states of the world can be specified by ordered integer pairs (m, n), with $n = m$ or $n = m - 1$, and $n \geqslant 0$. The first entry m designates the number of messages the first pilot received from the second pilot, plus one if they should attack. The second entry n designates the number of messages the second pilot received from the first pilot. Thus if $(m, n) = (0, 0)$, there should be no attack, and the second pilot receives no message. If $m = n > 0$, then they should attack, and the nth acknowledgement from the second pilot was lost. If $m = n + 1 > 0$, then they should attack, and the nth message from the first pilot was lost. Let $\text{Prob}(0, 0) = 1/2$, and for $m \geqslant 1$, $\text{Prob}(m, n) = \frac{1}{2}p^{m+n-1}(1 - p)$. 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.

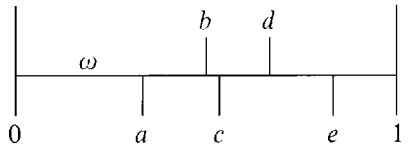
14. Approximate common knowledge

Since knowledge up to level N, no matter how large N is, does not guarantee behavior that even approximates behavior under common knowledge, we are left to wonder what is approximate common knowledge?

Consider a Bayesian game $\Gamma = (I, \Omega, (P_i, A_i, \pi_i, u_i)_{i \in I})$, and some event $E \subset \Omega$

and some $\omega \in \Omega$. If $\pi_i(\omega) > 0$, then we say that i p-believes E at ω iff the conditional probability $[\pi_i(P_i(\omega) \cap E)]/[\pi_i(P_i(\omega))] \geq p$, and we write $\omega \in B_i^p(E)$. Monderer and Samet (1989) called an event E p-self-evident to i iff for all $\omega \in E$, i p-believes E; an event E is p-public iff it is p-self-evident to every agent $i \in I$. Monderer and Samet called an event C p-common knowledge at ω iff there is some p-public event E with $\omega \in E \subset \bigcap_{i \in I} B_i^p(C)$.

We can illustrate this notion with a diagram.



The only public events are ϕ and Ω. But $[0, b)$ is p-public where $p = \text{Prob}[a, b)/\text{Prob}[a, c)$. Any event C containing $[0, b)$ is p-common knowledge at ω.

In our first theorem we show that if in a Bayesian game with asymmetric information the players' actions are p-common knowledge at ω, then we can define alternative partitions for the players such that the information at ω is symmetric, and such that with respect to this alternative information, the same action function for each player is "nearly" optimal at "nearly" every state ω', including at $\omega' = \omega$, provided that p is nearly equal to 1.

Theorem. *Let $(f_1, \ldots, f_I)$ be a Bayesian Nash equilibrium for the Bayesian game $\Gamma = (I, \Omega, (P_i, A_i, \pi_i, u_i)_{i \in I})$. Suppose $\sup_{i \in I} \sup_{a, a' \in A} \sup_{\omega, \omega' \in \Omega} [u_i(a, \omega) - u_i(a', \omega')] \leq M$. Suppose $\pi_i(\omega) > 0$ for all $i \in I$, and suppose that at ω it is p-common knowledge that $(f_1, \ldots, f_I) = (a_1, \ldots, a_I)$. Then there is a Bayesian game $\hat{\Gamma} = (I, \Omega, (\hat{P}_i, A_i, \pi_i, u_i)_{i \in I})$ with symmetric information at ω, $\hat{P}_i(\omega) = E$ for all $i \in I$, and sets $\omega \in E \subset \Omega_i \subset \Omega$ with $\pi_i(\Omega_i) \geq p$ such that for all $\omega' \in \Omega_i$ with $\pi_i(\omega') > 0$, and all $b_i \in A_i$,*

$$\frac{1}{\pi_i(\hat{P}_i(\omega'))} \sum_{s \in \hat{P}_i(\omega')} [u_i(f(s), s) - u_i(b_i, f_{-i}(s), s)] \pi_i(s) \geq -M \frac{(1-p)}{p}.$$

Proof. Let E be a p-public event with $\omega \in E \subset \bigcap_{i \in I} B_i^p(F) \subset F \equiv \{\omega' \in \Omega : f(\omega') = a\}$. Define

$$\hat{P}_i(\omega') = \begin{cases} E, & \text{if } \omega' \in E, \\ -E \cap P_i(\omega'), & \text{if } \omega' \notin E. \end{cases}$$

Then $\hat{P}_i(\omega) = E \; \forall i \in I$. Note that since $f_i(s) = a_i$ for all $s \in E$, f_i is a feasible action function given the information $\hat{P}_i$.

Consider any ω' such that $P_i(\omega') \cap E = \phi$. Then $\hat{P}_i(\omega') = P_i(\omega')$, so $f_i(\omega')$ is optimal for i. Consider $\omega' \in E$. Then since $(f_1, \ldots, f_I)$ is a BNE,

$$\sum_{s \in P_i(\omega') \cap E} [u_i(f(s), s) - u_i(b_i, f_{-i}(s), s)]\pi_i(s)$$
$$\geqslant - \sum_{s \in P_i(\omega') \setminus E} [u_i(f(s), s) - u_i(b_i, f_{-i}(s), s)]\pi_i(s)$$
$$\geqslant - M\pi_i(P_i(\omega') \setminus E)$$

so

$$\frac{1}{\pi_i(E)} \sum_{s \in E} [u_i(f(s), s) - u_i(b_i, f_{-i}(s), s)]\pi_i(s)$$
$$\geqslant \frac{-M\pi_i(P_i(E) \setminus E)}{\pi_i(E)} \geqslant \frac{-M(1-p)}{p},$$

where $P_i(E) = \bigcup_{\omega' \in E} P_i(\omega')$.

Finally, the set $P_i(E) \setminus E$ on which i may not be optimizing even approximately has π_i probability at most $1 - p$. So let $\Omega_i = E \cup (\Omega \setminus P_i(E))$. □

As an immediate corollary we deduce a proposition in Monderer and Samet (1989) that if it is p-common knowledge that two agents with the same priors believe the probabilities of an event G are q_i, respectively, then $|q_i - q_j| \leqslant 2\sqrt{(1-p)/p}$. To see this, note that the optimal action for i at ω in $\hat{\Gamma}$ is to choose $r = \pi(G \cap E)/\pi(E)$. Since with the loss function $u_i(a_i, a_{-i}, \omega) = -[a_i - \chi_G(\omega)]^2$, $M = 1$, we know that q_i cannot do worse than r by any more than $(1-p)/p$. Hence $(q_i - r)^2 \leqslant (1-p)/p$, hence $|q_i - q_j| \leqslant 2\sqrt{(1-p)/p}$. Thus as $p \to 1$, the agents must nearly agree. This result stands in contrast to the example in the last section where the opinions 2/3 and 1/3 stayed bounded away from each other no matter how many levels of knowledge about knowledge were reached.

An alternative definition of approximate common knowledge which allows for more p-public events suggests itself. We could say that an event E with $\pi_i(E) > 0$ is weakly p-self-evident to agent i iff

$$\frac{1}{\pi_i(E)} \sum_{\omega' \in E} \frac{\pi_i(P_i(\omega') \cap E)}{\pi_i(P_i(\omega'))} \pi_i(\omega') \geqslant p.$$

Instead of requiring at every $\omega' \in E$ that agent i should think that the probability of E is at least p, this requires the same thing only on average. In the previous diagram the event $[0, c)$ is weakly p-self-evident to each player, but not p-self-evident to the first agent. Notice that under this more generous definition of weakly p-self-evident (and hence weakly p-public and weakly p-common knowledge) exactly the same proof can be used to prove the preceding theorem.

The preceding theorem can be generalized in a second way. Suppose that the

action spaces A_i are compact metric spaces, and that the utilities u_i are continuous in A. Then we can replace the hypothesis that (it is weakly p-common knowledge that) the actions are $(a_1,\ldots,a_I)$ with the hypothesis that it is weakly p-common knowledge that the actions are within ε of $(a_1,\ldots,a_I)$. This explains why different agents' opinions must converge to each other in an infinite state space even though the opinions do not become common knowledge in finite time, but for brevity we omit the details.

The preceding theorem says that any BNE at which the actions are weakly p-common knowledge is an approximate BNE with symmetric information about events. The converse is also of interest. The following theorem [adapted from Monderer and Samet (1989)] shows that any conventional Nash equilibrium (which by definition can be achieved with symmetric information that the game is G) can be approximately achieved whenever it is p-common knowledge that the game is G.

Let M be as in the previous theorem, the maximum payoff difference for any player at any ω. Suppose now that the action spaces A_i are convex and compact, and that u_i is continuous in a, and concave in a_i for any fixed a_{-i} and ω.

Theorem. *Let $(\hat{f}_1,\ldots,\hat{f}_I)$ be a BNE for the Bayesian game $\hat{\Gamma}=(I,\Omega,\hat{P}_i,A_i,\pi_i,u_i)_{i\in I}$. Suppose that at some $\omega\in\Omega$, with $\pi_i(\omega)>0$ for all $i\in I$, $\hat{P}_i(\omega)=E$ for all $i\in I$, and $\hat{\Gamma}(\omega')=G$ for all $\omega'\in E$. Suppose that in the Bayesian game $\Gamma=(I,\Omega,(P_i,A_i,\pi_i,u_i)_{i\in I})$ E is p-common knowledge at ω. Then there exists $(f_1,\ldots,f_I)$ such that $f_i(\omega')=a_i=\hat{f}_i(\omega)$ for $\omega'\in E$ and all $i\in I$, and such that for all $\omega'\in\Omega$, $b_i\in A_i$,*

$$\frac{1}{\pi_i(P_i(\omega'))}\sum_{s\in P_i(\omega')}[u_i(f_i(s),f_{-i}(s),s)-u_i(b_i,f_{-i}(s),s)]\pi_i(s)\geqslant -M(1-p).$$

Proof. Define $f_i(\omega')=a_i$ if $P_i(\omega')\cap E\neq\phi$. Having fixed these actions, the Bayesian game Γ with these actions fixed defines a restricted Bayesian Game Γ^*. By our hypothesis on A_i and u_i, Γ^* must have a BNE $(f_1,\ldots,f_I)$. Observe that for ω' with $P_i(\omega')\cap E=\phi$, $f_i(\omega')$ is optimal in Γ. For ω' with $P_i(\omega')\cap E\neq\phi$,

$$\begin{aligned}&\frac{1}{\pi_i(P_i(\omega'))}\sum_{s\in P_i(\omega')}[u_i(a_i,f_{-i}(s),s)-u_i(b_i,f_{-i}(s),s)]\pi_i(s)\\&\quad\geqslant\frac{1}{\pi_i(P_i(\omega'))}\sum_{s\in P_i(\omega')\cap E}[u_i(a,s)-u_i(b_i,a_{-i},s)]\pi_i(s)+-M(1-p)\\&\quad\geqslant 0+-M(1-p).\quad\square\end{aligned}$$

The two theorems explain the coordinated attack problem. Suppose p is close to 1, so messages are quite reliable. Recalling our description from the last section, let $E=\{(m,n):m\geqslant 1\}$. For $(m,n)\geqslant(1,1)$, $\{\pi(E\cap P_i(m,n))\}/\{\pi(P_i(m,n))\}=1$. Only in the very unlikely state $(1,0)\in\bar{E}$ where the first message to the second pilot failed

to arrive can it be true that it is appropriate to attack but pilot 2 does not know it. Hence E is weakly p-public, but not p-public. We conclude first that the BNE of never attacking, in which the actions are common knowledge but there is asymmetric information, can be (approximately) achieved when there is symmetric information and $P_i(\omega) = E$ for all $i \in I$ and $\omega \in E$. And indeed, not attacking is a (Pareto inferior) Nash equilibrium of the coordinated attack problem when $P_i(\omega) = E$ for all i. On the other hand, although attacking is a (Pareto superior) Nash equilibrium of the common information game where $P_i(\omega) = E$ for all i, because in the asymmetric information attack game E is only *weakly* p-common knowledge, attacking is not even an approximate BNE in the asymmetric information game.

15. Hierarchies of belief: Is common knowledge of the partitions tautological?

Our description of reasoning about the reasoning of others (and ultimately of common knowledge) is quite remarkable in one respect which has been emphasized by Harsanyi (1968), in a Bayesian context. We have been able to express a whole infinite hierarchy of beliefs (of the form i knows that j knows that m knows, etc.) with a finite number of primitive states $\omega \in \Omega$ and correspondences P_i. One might have been tempted to think that each higher level of knowledge is independent of the lower levels, and hence would require another primitive element.

The explanation of this riddle is that our definition of i's knowledge about j's knowledge presupposes that i knows how j thinks; more precisely, i knows P_j. Our definition that i knows that j knows that m knows that A is true at ω, presupposes that i knows P_j, j knows P_m, and i knows that j knows P_m. Thus the model does include an infinite number of additional primitive assumptions, if not an infinite number of states. We refer to these additional assumptions collectively as the hypothesis of mutual rationality.

In order to rigorously express the idea that an event is common knowledge we apparently must assume mutual rationality and take as primitive the idea that the information partitions are "common knowledge." This raises two related questions. Are there real (or actually important) situations for which mutual rationality is plausible? Is mutual rationality an inevitable consequence of universal individual rationality?

As for the first question, the puzzles we began with are clear situations where it is appropriate to assume common knowledge of knowledge operators. Each child can readily see that the others know his hat color, and that each of them knows that the rest of them know his hat color and so on. In a poker game it is also quite appropriate to suppose that players know their opponents' sources of information about the cards. But what about the even slightly more realistic settings, like horse races? Surely it is not sensible to suppose that every bettor

knows what facts each other bettor has access to? This brings us to the second question.

One influential view, propounded first by Aumann (1976) along lines suggested by Harsanyi (1968), is that mutual rationality is a tautological consequence of individual rationality once one accepts the idea of a large enough state space. 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. One wonders whether the process would ever stop. The Harsanyi–Aumann doctrine asserts that it does. However, if it does, the states become descriptions of partition cells of the state space, which would seem to be an inevitable self-referential paradox requiring the identification of a set with all its subsets.

Armbruster and Boge (1979), Boge and Eisele (1979), and Mertens and Zamir (1985) were the first to squarely confront these issues. They focused on the analogous problem of probabilities. For each player i, each state is supposed to determine a conditional probability over all states, and over all conditional probabilities of player j, etc., again suggesting an infinite regress. Following Armbruster and Boge, Boge and Eisele and Mertens and Zamir, a large literature has developed attempting to show that these paradoxes can be dealt with. [See for example, Tan and Werlang (1985), Brandenburger and Dekel (1987), Gilboa (1988), Kaneko (1987), Shin (1993), Aumann (1989), and Fagin et al. (1992).]

The most straightforward analysis of the Harsanyi–Aumann doctrine (which owes much to Mertens and Zamir) is to return to the original problem of constructing the (infinite) hierarchy of partition knowledge to see whether at some level the information partitions are "common knowledge" at every ω, that is defined tautologically by the states themselves.

To be more precise, if $\Omega_0 = \{a, b\}$ is the set of payoff relevant states, we might be reluctant to suppose that any player $i \neq j$ knows j's partition of Ω_0, that is whether j can distinguish a from b. So let us set $\Omega^1 = \Omega_0 \times \{y_1, n_1\} \times \{y_2, n_2\}$. The first set $\{y_1, n_1\}$ refers to when player 1 can distinguish a from b (at y_1), and when he cannot (n_1). The second set $\{y_2, n_2\}$ refers to the second player. Thus the "extended state" $(a, (y_1, n_2))$ means that the payoff relevant state is a, that player 1 knows this, $y_1(a) = \{a\}$, but player 2 does not, $n_2(a) = \{a, b\}$. More generally, let Ω_0 be any finite set of primitive elements, which will define the payoff relevant universe. An element $\omega_0 \in \Omega_0$ might for example specify what the moves and payoffs to some game might be. For any set A, let $\mathbf{P}(A)$ be the set of partitions of A, that

is $\mathbf{P}(A) = \{P: A \to 2^A \mid \omega \in P(\omega)$ for all $\omega \in A$ and $[P(\omega) = P(\omega')]$ or $[P(\omega) \cap P(\omega')] = \phi$ for all $\omega, \omega' \in A\}$. For each player $i = 1, \ldots, I$, let $\Omega_{1i} = \mathbf{P}(\Omega_0)$ and let $\Omega_1 = \times_{i=1}^{I} \Omega_{1i}$. Then we might regard $\Omega^1 = \Omega_0 \times \Omega_1$ as the new state space.

The trouble of course is that we must describe each player's partition of Ω^1. If for each player i there was a unique conceivable partition of Ω^1, then we would say that the state space Ω^1 tautologically defined the players' partitions. However, since Ω^1 has greater cardinality than Ω_0 it would seem that there are more conceivable partitions of Ω^1 than there were of Ω_0. But notice that each player's rationality restricts his possible partitions. In the example, if $\omega' = (a, (y_1, n_2))$ then player 1 should recognize that he can distinguish a from b. In particular, if P is player 1's partition of Ω^1, then $(c, (z_1, z_2)) \in P(a, (y_1, n_2))$ should imply $z_1 = y_1$ and $c = a$. (Since player 1 might not know 2's partition, z_2 could be either y_2 or n_2.) Letting Proj denote projection, we can write this more formally as

$$\text{Proj}_{\Omega 1 i} P(a, (y_1, n_2)) = \{y_1\} \text{ and } \text{Proj}_{\Omega_0} P(a, (y_1, n_2)) = y_1(a).$$

In general, suppose we have defined Ω_0, and $\Omega_k = \Omega_{k1} \times \cdots \times \Omega_{kI}$ for all $0 < k < n$. This implicitly defines $\Omega^{n^*} = \times_{0 \leq k < n} \Omega_k$, and for each $k < n$, $\Omega^k = \times_{0 \leq r \leq k} \Omega_r$. Define $\Omega_{ni} = \{P_{ni} \in \mathbf{P}(\Omega^{n^*}): \forall (\omega_0, \ldots, \omega_k, \ldots) \in \Omega^{n^*}, \forall k < n,$

(1) $\text{Proj}_{\Omega_{ki}} P_{ni}(\omega_0, \ldots, \omega_k, \ldots) = \{\omega_{ki}\}$,

(2) $\text{Proj}_{\Omega^k} P_{ni}(\omega_0, \ldots, \omega_k, \ldots) = \{\omega_{k+1,i}(\omega_0, \ldots, \omega_k)\}$.

Condition (1) says that i knows his partitions at lower levels, and condition (2) says that he uses his information at lower levels to refine his partition at higher levels.

Let $\Omega_n = \times_{i \in I} \Omega_{ni}$. By induction Ω_n is defined for all integers n. In fact, by transfinite induction, Ω_n is defined for all finite and transfinite ordinals.

The Harsanyi–Aumann question can now be put rigorously as follows. Is there any n, finite or infinite, such that the state space Ω^{n^*} defines the partitions of itself tautologically, i.e., such that Ω_{ni} contains a unique element P_{ni} for each $i \in I$?

The most likely candidate would seem to be $n = \alpha$, where α is the smallest infinite ordinal. In that case $\Omega^{\alpha^*} = \Omega_0 \times \Omega_1 \times \Omega_2 \times \cdots$. However, as shown in Fagin et al. (1992), following the previous work in Fagin et al. (1991), the cardinality of Ω_{ni} is not only greater than one, it is infinite for all infinite ordinals n, including $n = \alpha$. This shows that the Harsanyi–Aumann doctrine is false. Properly expanded, the state space does not tautologically define the partitions.

To see why, reconsider our simple example with two payoff relevant states. Since the cardinality of Ω_0 is 2, the number of partitions of Ω_0 is also 2, and so the cardinality of Ω^1 is $2 \times (2 \times 2) = 8$. Taking into account the restrictions imposed by player 1's own rationality, the number of possible partitions player 1 could have of Ω^1 is equal to the number of partitions of the four elements $\{a, b, y_2, n_2\}$, namely 15. Hence the cardinality of Ω^2 is $8 \times (15 \times 15) = 1800$.

As we go up the hierarchy, the restrictions from individual rationality become more biting, but the cardinality of the base of states grows larger. Indeed it is evident from the analysis just given that if the cardinality of Ω_{ki} is at least two, then the cardinality of $\Omega_{k+1,i}$ is at least two. It follows that the cardinality of Ω^{k+1} must be at least 2^I times the cardinality of Ω^k, for all finite $k \geqslant 0$, if $\#I \geqslant 2$. It would be astonishing if there were only one partition of Ω^{α^*} consistent with player i's rationality.

The fact that there may be at least two partitions $P_i \neq Q_i$ in $\Omega_{\alpha i}$, that is partitions of Ω^{α^*} that are consistent with the rationality of agent i, raises an important question: how different can P_i and Q_i be? To answer this question we introduce a topology on Ω^{α^*}. Note that for each finite k, Ω_k is a finite set, hence it is natural to think of using the discrete topology on Ω_k. Since $\Omega^{\alpha^*} = \times_{k=1}^{\infty} \Omega_k$, it is also natural to take the product topology on Ω^{α^*}. With this topology we can state the following theorem adapted from Fagin et al. (1992).

Theorem. *Let Ω_0 be finite. Then the Harsanyi–Aumann expanded state space Ω^{α^*} allows for each agent $i \in I$ one and only one partition $P_i \in \Omega_{\alpha i}$ of Ω^{α^*} such that every partition cell $P_i(\omega)$, $\omega \in \Omega^{\alpha^*}$, is a closed subset of Ω^{α^*}.*

If we are willing to restrict our attention to partitions with closed cells, then this theorem can be considered a vindication of the Harsanyi–Aumann doctrine. The proof of the theorem is not difficult. Let P_i and Q_i be in $\Omega_{\alpha i}$, and suppose $P_i(\omega)$ and $Q_i(\omega)$ are closed subsets of Ω^{α^*}. From conditions (1) and (2), we know that for each finite k,

$$\text{Proj}_{\Omega^k} P_i(\omega) = \omega_{k+1,i}(\omega_1, \ldots, \omega_k) = \text{Proj}_{\Omega^k} Q_i(\omega).$$

But since $P_i(\omega)$ and $Q_i(\omega)$ are closed in the product topology, this implies that $P_i(\omega) = Q_i(\omega)$, and the theorem follows.

We can state an analogous theorem [from Fagin et al. (1992)] that may also give the reader a sense of how close to true one might consider the Harsanyi–Aumann doctrine.

Theorem. *Let $(P_1, \ldots, P_I)$ and $(Q_1, \ldots, Q_I)$ be in Ω_α, that is let P_i and Q_i be partitions of the expanded state space Ω^{α^*} that are consistent with i's rationality, for each $i \in I$. Let $\omega \in \Omega^{\alpha^*}$, and let $E \subset \Omega^{\alpha^*}$ be closed. Then i knows E at ω with respect to P_i, $P_i(\omega) \subset E$, if and only if i knows E at ω with respect to Q_i, $Q_i(\omega) \subset E$. Furthermore, E is common knowledge at ω with respect to the partitions $(P_1, \ldots, P_I)$ if and only if E is common knowledge at ω with respect to the partitions $(Q_1, \ldots, Q_I)$.*

Note that any event E which depends only on a finite number of levels of the hierarchy is necessarily closed. These elementary events are probably of the most interest to noncooperative game theory. They include for example any description

of the payoff relevant states, or what the players know about the payoff relevant states, and so on.

Here is an example of an event E that is not necessarily closed. Let A be a description of some payoff relevant states. Then E is defined as the set of ω at which i knows that it is not common knowledge between j and k that A happens.

In conclusion, we can say that the Harsanyi–Aumann doctrine can be partially vindicated by a rigorous construction of a knowledge hierarchy. If the partitions of the expanded state space Ω^{α^*} are restricted to have closed cells, then the state space Ω^{α^*} tautologically (uniquely) defines each agent's partition. A similar (positive) result was obtained by Mertens and Zamir (1985). If the state space is sufficiently enlarged, and if attention is restricted to *countably* additive *Borel* probabilities, then each state uniquely defines a conditional probability for each player. However, if more general (finitely additive) probabilities were allowed, then there would be many conditional probabilities consistent with a player's rationality.

Using our restrictions on potential partitions and probabilities, the knowledge of the players can always be described as in the first sections of this paper $(\Omega^*, P_1, \ldots, P_I)$, in which each player's knowledge pertains only to the state space Ω^* (and not to each other), and the partitions P_i are "common knowledge." As before, the universal state space Ω^* is the disjoint union of common knowledge components. In some of these there are a finite number of states, in others an infinite (uncountable) number. In some common knowledge components the players' conditional beliefs can all be explained as coming from a common prior; in others they cannot.

The restrictions to common priors, and finite Ω are nontrivial.[7] The "Harsanyi doctrine" asserts that it is reasonable that all agents should have the same prior, and many would agree. But the hierarchical argument we have just given does not provide any justification for this second doctrine.

16. Bounded rationality: Irrationality at some level

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; and most surprising of all, it banishes speculation. (Here speculation is distinguished from betting because it may not be common knowledge that the deal is agreed, as for example, the moment at which a stock market investor places a buy order.) Yet casual empiricism suggests that all of these are frequently observed phenomena. This section explores the possibility that it is not really common knowledge that agents optimize, though in fact they do.

[7] Mertens and Zamir (1985) show that any common knowledge component of Ω^* that is infinite can be "approximated" by a finite common knowledge component.

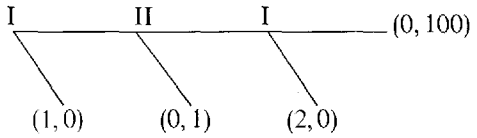
We have already seen that if actions are iterated knowledge to some large but finite level N, then we may observe behavior which is very different from that which could be seen if the actions were common knowledge. In particular, agents could disagree. We hesitate to say that agents would bet, since in agreeing to the wager it might become common knowledge that they are betting. The significance of common knowledge, however, is lost on all but the most sophisticated reasoner, who would have to calculate that "he wants to bet, he wants to bet knowing that I am betting against him, he wants to bet knowing that I want to bet knowing that he wants to bet against me knowing that I want to bet etc." Most agents do not have the computing power, or logical powers, to make this calculation. To allow for this limitation, we suppose that somewhere in the calculation agents no longer can deduce any significance from the fact the other fellow is betting, which is to say that they no longer restrict attention to cases where the other fellow is optimizing. Suppose that at the actual state of the world ω, agents optimize, and know that they optimize, and know that they know that they optimize, but only a finite number of times rather than the infinity required by common knowledge. That is, suppose that it is common knowledge what the agents are doing, but only iterated knowledge to level N that they are optimizing. Then what the agents wish to do at ω might not be the same as when the actions and optimality are common knowledge.

As an example, reconsider the first version of the envelopes puzzle. Imagine that the two sons have \$10 000 and \$1 000 in their envelopes respectively. Suppose it were common knowledge that had son 2 seen \$10 000 000 in his envelope, then he would bet (even though he could only lose in that case). Then the sons would be willing to bet against each other at state (4, 3), and in every other state (with m even and n odd). At state (4, 3), it is common knowledge that they are betting. Both sons are acting optimally given their information, both sons know that they are acting optimally, and they each know that the other knows that each is acting optimally. Of course it is not common knowledge that they are optimizing, since the state (6, 7) is reachable from (4, 3), and there the second son is not optimizing. The ex ante probability of nonoptimization here is only 1/12, and by extending the maximum amount in the envelopes we can make the probability of nonoptimal behavior arbitrarily small, and still guarantee that the sons bet against each other at (4, 3). (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.) The same logic, of course, applies to the question of agreeing to disagree. [For more on this, see Aumann (1992).]

The possibility of nonoptimal behavior can also have dramatic consequences for dynamic Bayesian games. We have already seen in the N-repeated Prisoner's Dilemma that even when both players are optimizing, the possibility that the other is not can induce two optimizing agents to cooperate in the early periods, even though they never would if it were common knowledge that they were optimizing.

Simply by letting the time horizon N be uncertain, Neyman (in unpublished work) has shown that the two agents can each be optimizing, can each know that they are optimizing, and so on up to $m < N$ times, yet still cooperate in the first period. Of course this is analogous to the envelopes example just discussed.

Games in extensive form sometimes give rise to a backward induction paradox noted by Binmore (1987), Reny (1992), and Bicchieri (1988), among others. Consider the following extensive form game:



In the unique dynamic Bayesian equilibrium, I plays down immediately. We usually explain this by suggesting that I figures that if he played across instead, then II would play down in order to avoid putting I on the move again. But if I is "irrational" enough to play across on his first move, why should not II deduce that I is irrational enough to play across on his second move? It would appear, according to these authors, that to interpret fully a dynamic Bayesian game one needs a theory of "irrationality," or counterfactual reasoning. The beginning of such a theory is provided by Selten's (1975) notion of the trembling hand, which is discussed at length in other chapters of this volume.

17. Bounded rationality: 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., it seems plausible to suppose that it is also common knowledge they are optimizing. In the last part of this section we return to the assumption that it is common knowledge that agents optimize, but we continue to examine 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, betting, and agreements to disagree emerge in equilibrium?[8]

There are a number of errors that are typically made by decision makers that suggest that we go beyond the orthodox Bayesian paradigm. Agents often forget, or ignore unpleasant information, or grasp only the superficial content of signals.

[8]Much of this section is taken from Geanakoplos (1989), which offers a fuller description of possible types of irrationality and derives a number of theorems about how they will affect behavior in a number of games.

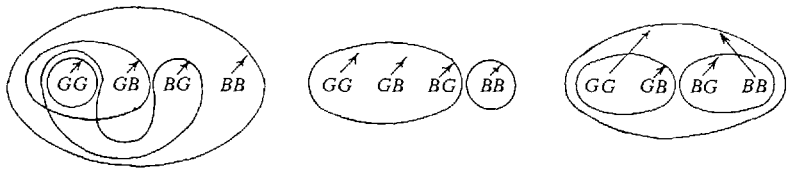
Many of these mistakes turn on the idea that agents often do not know that they do not know. For example, in the story of Silver Blaze, Sherlock Holmes draws the attention of the inspector to "the curious incident of the dog in the night-time." "The dog did nothing in the night-time," protested the inspector, to which Holmes replied "That was the curious incident." Indeed, the puzzle of the hats was surprising because it relied on the girls being so rational that they learned from each other's ignorance, which we do not normally expect. 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 gumma 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 $\{a, b\}$ 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.

We can describe the knowledge operator generated by the possibility correspondece P just as we did for partitions: $K(E) = \{\omega \in \Omega: P(\omega) \subset E\}$ for all events E. In the ozone example, $K(\{a\}) = \{a\}$, $K(\{b\}) = \phi$, and $K(\{a, b\}) = \{a, b\}$. The reader can verify that K satisfies the first four axioms of S5 described earlier, but it fails the fifth: $\sim K(\{a\}) = \{b\} \neq \phi = K \sim K(\{a\})$. Non-partitional information has been discussed by Shin (1993), Samet (1990), Geanakoplos (1989), and Brandenburger et al. (1992).

The definitions of Bayesian game and Bayesian Nash equilibrium do not rely on partitions. If we substitute possibility correspondences for the partitions, we can retain the definitions word for word. At a Bayesian Nash equilibrium, for each state $\omega \in \Omega$ agents use their information $P_i(\omega)$ to update their priors, and then they take what appears to them to be the optimal action. Nothing in this definition formally requires that there be a relationship between $P_i(\omega)$ and $P_i(\omega')$ for different ω and ω'. Moreover, the definitions of self-evident and public also do not rely on any properties of the correspondences P_i. We can therefore investigate our agreement theorems and nonspeculation theorems when agents have nonpartitional information.

Imagine a doctor Q who initially assigns equal probability to all the four states



Doctor Q | Doctor R | Doctor S

Diagram 2a | **Diagram 2b** | **Diagram 2c**

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 he is contemplating will succeed and the payoff will be 3. But if either is missing, i.e., if the state is any of GB, BG, or BB, the operation will fail and he will lose 2. Suppose that in his laboratories his 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 possibility correspondence of doctor Q is therefore described by Diagram 2a. 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, he 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 him to go ahead with the operation. (If the actual state of the world is BB, doctor Q will get no information from his lab and will decide not to do the operation). Yet a doctor R^0 who had no lab and knew nothing about which state was going to occur, would never choose to do the operation. Nor would doctor R, who can recognize whether or not both antibodies are missing. Doctor R's information is given by the partition in Diagram 2b. From an ex ante perspective, both doctors R^0 and R do better than doctor Q. We see from this example that when agents do not process information completely rationally, more knowledge may be harmful. Furthermore, Q does not satisfy the sure-thing principle. At each ω in the self-evident set $M = \{GG, GB, BG\}$, doctor Q would choose to operate, given his information $Q(\omega)$. Yet if told only $\omega \in M$, he would choose not to operate.

Doctor Q does not know what he does not know; the knowledge operator K_Q derived from his possibility correspondence Q does not satisfy $\sim K_Q = K_Q \sim K_Q$. But like the ozone scientist P, doctor Q does satisfy the first four S5 axioms. In particular, doctor Q knows what he knows. He can recognize at GB that the first

antibody is present, and whenever that condition obtains, he recognizes it. Formally, the possibility correspondence P_i gives rise to a knowledge operator K_i satisfying the "know that you know" (KTYK) axiom iff for all ω, $\omega' \in \Omega$, $\omega' \in P_i(\omega)$ implies $P_i(\omega') \subset P_i(\omega)$. Doctor Q's error is that he sometimes overlooks the condition of the first antibody while recognizing the condition of the second antibody, and at other states he does the reverse. If he paid attention to the antibodies in the same order, his knowledge would satisfy the "nested" property. Formally, a possibility correspondence satisfies the memory property called nested if for all ω and ω' with $P_i(\omega) \cap P_i(\omega') \neq \phi$, either $P_i(\omega) \subset P_i(\omega')$ or $P_i(\omega') \subset P_i(\omega)$.

Theorem. *Consider the one-person Bayesian games* $(\Omega, (P_i, A_i, \pi_i, u_i))$ *and* $(\Omega, (Q_i, A_i, \pi_i, u_i))$ *where* Q_i *is a partition and* P_i *is finer than* Q_i*, but is not necessarily a partition. Let* f *be a BNE in the first game, and* g *in the second. Then if* P_i *is nondeluded* ($\omega \in P_i(\omega)$ *for all* ω)*, and satisfies knowing that you know and nested, then* $\sum_{\omega \in \Omega} u_i\ (f(\omega), \omega)\pi_i(\omega) \geq \sum_{\omega \in \Omega} u_i(g(\omega), \omega)\pi_i(\omega)$*. Conversely, if* P_i *fails any one of these properties, then there exist* A_i, u_i, π_i, *and BNE* f *and* g *such that the above inequality fails.*

We could also give nccessary and sufficient conditions (nondeluded plus a property called positively balanced, which is implied by nested) for a nonpartitional decisionmaker to satisfy the sure-thing principle. Doctor Q's information processing leads to decisions that violate both the principle that more information is better and the sure-thing principle. By contrast, ozone scientist P will satisfy both the sure-thing principle and the principle that more information is better, despite not being perfectly rational. It turns out that more rationality is required for the principle that information is good than for the sure-thing principle.

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 or when both are absent, in which cases the experiment fails and reveals nothing at all. If doctor S takes his laboratory results at face value, then his information is described by Diagram 2c. 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.

Doctor S is thus worse off ex ante with more information. Note that S fails KTYK. But S satisfies nested and therefore positively balanced, which implies that S's behavior will satisfy the sure-thing principle.

Since Bayesian games are formally well-defined with nonpartitional information, we can put the doctors together and see whether they would speculate, or bet, or agree to disagree. In general, as we saw in the last section, we would have to expand the state space to take into account each doctor's uncertainty about the other doctor's information. But for illustration, let us suppose that the state space

Ω of four states already takes this into account. Again, this formally makes sense because in BNE each player thinks about the states and actions, not directly about the other players. Once we assign actions for all players to each state, we are back in that framework.

Furthermore, it might appear that it would make no sense to ask when an event $E \subset \Omega$ is common knowledge when agents have nonpartitional information. Since the agents do not fully understand their own information precisely, how can they think about what the others think about their information processing? The answer is that common knowledge can, as we saw, be understood in a way that depends only on self-evident events. Each of the doctors Q, R, S can be perfectly aware of which events E are self-evident, i.e., satisfy $P_i(\omega) \subset E$ for all $\omega \in E$. This degree of mutual awareness would not induce them to refine their possibility correspondences, except at $\omega = BB$.

After getting their information at any $\omega \in M = \{GG, GB, GB\}$, doctors Q and R 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 in M doctor Q will decide to perform the operation, and therefore the bet will come off. Moreover, the event M is public, so we can say that the bet is common knowledge at each $\omega \in M$. 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.

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 doctor R will come out ahead, despite having less information. In this BNE it is not known by doctor R that the bet is going to come off when doctors S 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 (with partition information, or even one who makes the same kinds of errors as doctor S) who doctor S would bet with and with whom it would be common knowledge that the bet was going to come off.

It can also be shown that ozone scientist P 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 degree of rationality of the agents (i.e., for the kinds of 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 (satisfying nondeluded, KTYK, and nested), and still not speculate, bet, or agree to disagree. But if agents are a little more irrational (satisfying nondeluded and positively balanced), they will speculate, but not bet or agree to disagree. If they get still more irrational (satisfying nondeluded and balanced), they

will speculate and bet, but not agree to disagree about the probability of an event.[9] Finally, with still more irrationality, they will speculate, bet, and agree to disagree.

References

Armbruster, W. and W. Boge (1979) 'Bayesian game theory', in: O. Moeschlin and D. Pallaschke, eds., *Game theory and related topics*. Amsterdam: North Holland.

Aumann, R. (1974) 'Subjectivity and correlation in randomized strategies', *Journal of Mathematical Economics*, **1**: 67–96.

Aumann, R. (1976) 'Agreeing to disagree', *The Annals of Statistics*, **4**: 1236–1239.

Aumann, R. (1987) 'Correlated equilibrium as an expression of Bayesian rationality', *Econometrica*, **55**: 1–18.

Aumann, R. (1992) 'Irrationality in game theory', in P. Dasgupta, D. Gale, O. Hart and E. Maskin, eds., *Economic analysis of markets and games*, pp. 214–227.

Aumann, R. (1989) 'Notes on interactive epistemology', mimeo.

Aumann, R. and A. Brandenburger (1991) 'Epistemic conditions for Nash equilibrium', Working Paper 91-042, Harvard Business School.

Aumann, R. and M. Maschler (1966) 'Game theoretic aspects of gradual disarmament', Chapter V in Report to the US Arms Central and Disarmament Agency, Washington, DC.

Bacharach, M. (1985) 'Some extensions of a claim of Aumann in an axiomatic model of knowledge', *Journal of Economic Theory*, **37**: 167–190.

Bicchieri, C. (1988) 'Common knowledge and backward induction: a solution to the paradox', in: M. Vardi, ed., *Proceedings of the Second Conference on Reasoning about Knowledge*. Los Altos: Morgan Kaufmann Publishers, pp. 381–394.

Binmore, K. (1987–88) 'Modeling rational players', *Economics and Philosophy*, **3**: 179–214, **4**: 9–55.

Boge, W. and Th. Eisele (1979) 'On solutions of Bayesian games', *International Journal of Game Theory*, **8**(4): 193–215.

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

Brandenburger, A. and E. Dekel (1987) 'Common knowledge with probability 1', *Journal of Mathematical Economics*, **16**(3): 237–245.

Brandenburger, A. and E. Dekel (1993) 'Hierarchies of belief and common knowledge', *Journal of Economic Theory*, **59**(1): 189–198.

Brandenburger, A., E. Dekel and J. Geanakoplos (1992) 'Correlated equilibrium with generalized information structures', *Games and Economic Behavior*, **4**(2): 182–201.

Cave, J. (1983) 'Learning to agree', *Economics Letters*, **12**: 147–152.

Chou, C. and J. Geanakoplos (1988) 'The power of commitment', Cowles Foundation Discussion Paper No. 885.

Dubey, P., J. Geanakoplos and M. Shubik (1987) 'The revelation of information in strategic market games: a critique of rational expectations equilibrium', *Journal of Mathematical Economics*, **16**(2): 105–138.

Fagin, R., J. Geanakoplos, J. Halpern and M. Vardi, 'The expressive power of the hierarchical approach to modeling knowledge and common knowledge', in M. Vardi, ed., *Fourth Symposium on Theoretical Aspects of Reasoning about Knowledge*. Los Altos: Morgan Kaufmann Publishers, 1992.

Fagin, R., J. Halpern and M. Vardi (1991) 'A model-theoretic analysis of knowledge', *Journal of the ACM*, **91**(2): 382–428.

Gamow, G. and M. Stern (1958) 'Forty unfaithful wives', in: *Puzzle math*. New York: The Viking Press, pp. 20–23.

Gardner, M. (1984) *Puzzles from other worlds*. Vintage.

[9] Samet (1990) had previously shown that non-deduced and KTYK are sufficient conditions to rule out agreeing to disagree. Non-deluded and balanced are necessary as well as sufficient conditions to rule out agreeing to disagree.

Geanakoplos, J. (1989) 'Game theory without partitions and applications to speculation and consensus', Cowles Foundation Discussion Paper No. 914, Yale University, forthcoming *Journal of Economic Theory*.
Geanakoplos, J. (1992) 'Common knowledge', *Journal of Economic Perspectives*, **6**(4): 58–82.
Geanakoplos, J. and H. Polemarchakis (1982) 'We can't disagree forever', *Journal of Economic Theory*, **28**: 192–200.
Gilboa, I. (1988) 'Information and meta-information', in M. Vardi, ed., *Theoretical aspects of reasoning about knowledge*. Los Altos: Morgan Kaufmann Publishers, pp. 1–18.
Green, J. and J. Laffont (1987) 'Posterior implementability in a two-person decision problem', *Econometrica*, **55**: 69–94.
Halpern, J.Y. (1986) 'Reasoning about knowledge: an overview', IBM Research Report RJ-5001.
Halpern, J. and Y. Moses (1984) 'Knowledge and common knowledge in a distributed environment', in: *Proceedings of the 3rd ACM Conference on Principles of Distributed Computing*, pp. 50–61.
Harsanyi, J. (1967, 1968) 'Games with incomplete information played by "Bayesian" players', Parts I–III, *Management Science*, **14**(3): 159–183; **14**(5): 320–334; **14**(7): 486–502.
Kaneko, M. (1987) 'Structural common knowledge and factual common knowledge', RUEE Working Paper 87–27, Department of Economics, Hitosubashi University.
Kreps, D. (1977) 'A note on fulfilled expectations equilibrium', *Journal of Economic Theory*, 32–43.
Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982) 'Rational cooperation in the finitely repeated Prisoner's Dilemma', *Journal of Economic Theory*, **27**: 245–23.
Kripke, S. (1963) 'Semantical analysis of model logic', *Z. Math Logik Grundlag der Math*, **9**: 67–96.
Lewis, D. (1969) *Convention: a philosophical study*. Cambridge: Harvard University Press.
McKelvey, R. and T. Page (1986) 'Common knowledge, consensus and aggregate information', *Econometrica*, **54**: 109–127.
Mertens, J.F. and S. Zamir (1985) 'Formulation of Bayesian analysis for games with incomplete information', *International Journal of Game Theory*, **14**: 17–24.
Milgrom, P. (1981) 'An axiomatic characterization of common knowledge', *Econometrica*, **49**: 219–222.
Milgram, P. and N. Stokey (1982) 'Information, trade and common knowledge', *Journal of Economic Theory*, **26**: 17–27.
Monderer, D. and D. Samet (1989) 'Approximating common knowledge with common beliefs', *Games and Economic Behavior*, **1**: 170–190.
Morris, S. (1991) 'The role of beliefs in economic theory', PhD. Dissertation, Yale University.
Nielsen, L. (1984) 'Common knowledge, communication and convergence of beliefs', *Mathematical Social Sciences*, **8**: 1–14.
Nielsen, L., A. Brandenburger, J. Geanakoplos, R. McKelvey and T. Page (1990) 'Common knowledge of an aggregate of expectations', *Econometrica*, 1235–1239.
Parikh, R. and P. Krasucki (1990) 'Communication, consensus and knowledge', *Journal of Economic Theory*, **52**(1): 178–189.
Rubinstein, A. (1989) 'The electronic mail game: strategic behavior under "almost common knowledge"', *American Economic Review*, **79**(3): 385–391.
Rubinstein, A. and A. Wolinsky (1990) 'Remarks on the logic of "agreeing to disagree" type results', *Journal of Economic Theory*, **51**: 184–193.
Reny, P.J. (1992) 'Rationality in extensive form games', *Journal of Economic Perspectives*, **6**(4): 103–118.
Samet, D. (1990) 'Ignoring ignorance and agreeing to disagree', *Journal of Economic Theory*, **52**(1): 190–207.
Savage, L. (1954) *The foundations of statistics*. New York: Wiley.
Sebenius, J. and J. Geanakoplos (1983) 'Don't bet on it: contingent agreements with asymmetric information', *Journal of the American Statistical Association*, **78**: 424–426.
Selten, R. (1975) 'Reexamination of the perfectness concept for equilibrium points in extensive games', *International Journal of Game Theory*, **4**(1): 25–55.
Shin, H. (1993) 'Logical structure of common knowledge', *Journal of Economic Theory*, **60**(1): 1–13.
Tan, T. and S. Werlang (1985) 'The Bayesian foundations of solution concepts of games', *Journal of Economic Theory*, **45**: 379–391.
Tirole, J. (1982) 'On the possibility of speculation under rational expectations', *Econometrica*, **50**: 1163–1181.