Common Knowledge: Analysis & Applications

Peter Vanderschraaf

May 1998

Technical Report No. CMU-PHIL-85

Philosophy

Methodology

Logic

Carnegie Mellon

Pittsburgh, Pennsylvania 15213

> one can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another.
>
> Gottlob Frege, "On Sense and Reference"

> When a man loses his wife in a department store without any prior understanding on where to meet if they get separated, the chances are good that they will find each other. It is likely that each will think of some obvious place to meet, so obvious that each will be sure that it is "obvious" to both of them. One does not simply predict where the other will go, which is wherever the first predicts the second to predict the first to go, and so ad infinitum. Not "What would I do if I were she?" but "What would I do if I were she wondering what she would do if she were wondering what I would do if I were she . . . ?"
>
> Thomas Schelling, *The Strategy of Conflict*

Frege (1892) took it to be obvious that we convey knowledge successfully via language. But to achieve this, we evidently need some common understanding or common knowledge of the language in use, which leads to questions Frege did not address: *What does it mean for a group of people to have common knowledge? Can we really attain common knowledge, and if so, how does this happen?*

Common knowledge is a phenomenon which underwrites much more of social life than the successful transmission of knowledge via language. In order to communicate or otherwise coordinate their behavior successfully, individuals typically require mutual or common understandings or background knowledge. Indeed, if a particular interaction results in "failure", the usual explanation for this is that the agents involved did not have the common knowledge that would have resulted in success. In the department store problem Schelling describes in the quoted passage, the spouses stand a good chance of finding one another because their common knowledge of each others' tastes and experiences leads them each to look for the other in a part of the store both know that both would tend to frequent. Since the spouses both love cappuccino, each expects the other to go to the coffee bar, and they find one another. But in a less happy case, if a pedestrian causes a minor traffic jam by crossing against a red light, she explains her mistake as the result of her not noticing, and therefore not knowing, the status of the traffic signal that all the motorists knew. The spouses coordinate successfully given their common knowledge, while the pedestrian and the motorists miscoordinate as the result of a breakdown in common knowledge.

Given the importance of common knowledge in social interactions, it is remarkable that only quite recently have philosophers and social scientists attempted to analyze the concept. David Hume (1740) was perhaps the first to make explicit reference to the importance of *mutual knowledge*, a notion somewhat weaker than common knowledge, in social coordination. In his account of convention in *A Treatise of Human Nature*, Hume argued that a necessary condition for coordinated activity was that agents all know what behavior to expect from one another. Without the requisite mutual knowledge, Hume maintained, mutually beneficial social conventions would disappear. Much later, J. E. Littlewood (1953) presented some examples of common-knowledge-type reasoning, and Thomas Schelling (1960) and John Harsanyi (1967-1968) argued that something like common knowledge is needed to explain certain inferences people make about each other. Yet it was

David Lewis (1969) who first gave an explicit analysis of common knowledge in the monograph *Convention*. Stephen Schiffer (1972) and Robert Aumann independently gave alternate definitions of common knowledge which are in some contexts more convenient to use than Lewis' definition. Schiffer's definition of common knowledge as a *hierarchy* of epistemic claims has become standard in the philosophical and social science literature. The analysis of common knowledge as a *hierarchy* of epistemic claims that Lewis, Schiffer and Aumann all adopt has become standard in the philosophical and social science literature. More recently, Margaret Gilbert (1989) proposed a somewhat different account of common knowledge which she argues is preferable to the standard account. Others have developed accounts of mutual knowledge, *approximate common knowledge* and *common belief* which require less stringent assumptions than the standard account, and which serve as more plausible models of what agents know in cases where strict common knowledge seems impossible (Brandenburger and Dekel 1987, Stinchcombe 1988, Monderer and Samet 1989, Rubinstein 1992). The analysis and applications of common knowledge and related multi-agent knowledge concepts has become a lively field of research.

The purpose of this essay is to overview of some of the most important results stemming from this contemporary research. The topics reviewed in each section of this essay are as follows: §1: examples which illustrate a variety of ways in which the actions of agents depend crucially upon their having, or lacking, certain common knowledge, §2: several proposed analyses of common knowledge, and an analysis of the weaker *common belief* concept which result from weakening the assumptions of Lewis' account of common knowledge, §3: applications of common knowledge and the related multi-agent knowledge concepts, particularly to *game theory* (von Neumann and Morgenstern 1944), in which common knowledge assumptions have been found to have great importance in justifying *solution concepts* for mathematical games.

§1. Motivating Examples

Many of the examples in this section are familiar in the common knowledge literature, although some of the details and interpretations presented here are new. Readers may want to ask themselves what, if any, distinctive aspects of mutual and common knowledge reasoning each example illustrates.

Example 1.1. The Clumsy Waiter[1]

A waiter serving dinner slips, and spills gravy on a guest's white silk evening gown. The guest glares at the waiter, and the waiter declares "I'm sorry. It was my fault." Why did the waiter say that he was at fault? He knew that he was at fault, and he knew from the guest's angry expression that she knew he was at fault. However, the sorry waiter wanted assurance that the guest *knew that he knew* he was at fault. By saying openly that he was at fault, the waiter knew that the guest knew what he wanted her to know, namely, that he knew he was at fault. Note that the waiter's declaration established at least three levels of nested knowledge.

Certain assumptions are implicit in the preceding story. In particular, the waiter must know that the guest knows he has spoken the truth, and that she can draw the desired conclusion from what he says in this context. More fundamentally, the waiter must know that if he announces "It was my fault." to the guest, she will interpret his intended meaning

---

[1]Thanks to Alan Hajek for this example, the only example in this section which does not appear elsewhere in the literature.

correctly and will infer what his making this announcement ordinarily implies in this context. This in turn implies that the guest must know that if the waiter announces "It was my fault." in this context, then the waiter indeed knows he is at fault. Then on account of his announcement, the waiter knows that the guest knows that he knows he was at fault. So we have a special case of Frege's truism that knowledge is transmitted via language, and an unusual one, since the waiter's announcement was meant to generate *higher-order* levels of knowledge of a fact each already knew.

Just a slight strengthening of the stated assumptions results in even higher levels of nested knowledge. Suppose the waiter and the guest each know that the other can infer what he infers from the waiter's announcement. Can the guest now believe that the waiter does not know that she knows that he knows he is at fault? If the guest considers this question, she reasons that if the waiter falsely believes it is possible that she does not know that he knows he is at fault, then the waiter must believe it to be possible that she cannot infer that he knows he is at fault from his own declaration. Since she knows she *can* infer that the waiter knows he is at fault from his declaration, she knows that the waiter knows she can infer this, as well. Hence the waiter's announcement establishes the fourth-order knowledge claim: The guest knows that the waiter knows that she knows that he knows he is at fault. By similar, albeit lengthier, arguments, the agents can verify that corresponding knowledge claims of even higher-order must also obtain under these assumptions. □

Example 1.2. The Barbecue Problem

This is a variation of an example first published by Littlewood (1953), although he notes that his version of the example was already well-known at the time.[2] n individuals enjoy a picnic supper together which includes barbecued spareribs. At the end of the meal, $k \geq 1$ of these diners have barbecue sauce on their faces. No one wants to continue the evening with a messy face, but no one wants to wipe her face if it's not messy, for this would make her appear neurotic. And no one wants to take the risk of being thought rude by telling anyone else that he has barbecue sauce on his face. Since no one can see her own face, none of the messy diners makes a move to clean her face. Then the cook who served the spareribs returns with a carton of ice cream. Amused by what he sees, the cook rings the dinner bell and makes the following announcement: "At least one of you has barbecue sauce on her face. I will ring the dinner bell over and over, until anyone who is messy ones has wiped her face. Then I will serve dessert." For the first $k-1$ rings, no one does anything. Then, at the kth ring, each of the messy individuals suddenly reaches for a napkin, and soon afterwards, the diners are all enjoying their ice cream.

How did the messy diners finally realize that their faces needed cleaning? The $k = 1$ case is easy, since in this case, the lone messy individual will realize he is messy immediately, since he sees that everyone else is clean. Consider the $k = 2$ case next. At the first ring, messy individual i_1 knows that one other person, i_2, is messy, but does not yet know about himself. At the second ring, i_1 realizes that he must be messy, since had i_2 been the only messy one, i_2 would have known this after the first ring when the cook made his announcement, and would have cleaned her face then. By a symmetric argument, messy diner i_2 also concludes that she is messy at the second ring, and both pick up a napkin at that time.

---

[2]The version of the story Littlewood analyzes involves a group of cannibals, some of whom are married to unfaithful wives, and a missionary who visits this group and makes a public announcement of the fact.

Let's next consider $k = 3$. Again at the first ring, each of the messy diners i_1, i_2, and i_3 knows the status of the other diners, but not her own. The situation is apparently unchanged after the second ring. But on the third ring, i_1 realizes that she is messy. For if i_2 and i_3 were the only messy ones, then they would have discovered this after the second ring by the argument of the previous paragraph. Since i_1 can see that all of the diners other than i_2 and i_3 are clean, she concludes that she must be messy. i_2 and i_3 draw similar conclusions at the third ring, and all clean their faces at that time.

The general case follows by induction. Suppose that if $k = j$, then each of the j messy diners can determine that he is messy after j rings. Then if $k = j + 1$, then at the $j + 1$st ring, each of the $j + 1$ individuals will realize that he is messy. For if he were not messy, then the other j messy ones would have all realized their messiness at the jth ring and cleaned themselves then. Since no one cleaned herself after the jth ring, at the $j + 1$st ring each messy person will conclude that someone besides the other j messy people must also be messy, namely, himself.

The "paradox" of this argument is that for $k > 1$, like the case of the clumsy waiter of Example 1.1, the cook's announcement told the diners something that each already knew. Yet apparently the cook's announcement also gave the diners useful information. How could this be? By announcing a fact already known to every diner, the cook made this fact *common knowledge* among them, enabling each of them to eventually deduce the condition of his own face after sufficiently many rings of the bell. Note that the inductive argument the agents run through depends upon the conclusions they each draw from several *counterfactual conditionals*. In general, the consequences of agents' common knowledge are intimately related to how they evaluate subjunctive and counterfactual conditionals.[3] □

Example 1.3. Backwards Induction

Does acting cooperatively with others serve one's self-interest? Plato and his successors recognized that in certain cases, the answer seems to be "No.". Hobbes (1651, pp. 101-102), for instance, considers the challenge of a "Foole", who claims that it is irrational to honor an agreement made with another who has already fulfilled his part of the agreement. Noting that in this situation one has gained all the benefit of the other's compliance, the Foole contends that it would now be best for him to break the agreement, thereby saving himself the costs of compliance. Of course, if the Foole's analysis of the situation is correct, then would the other party to the agreement not anticipate the Foole's response to agreements honored, and act accordingly?

Hume (1740, pp. 520-521) takes up this question, using an example: Two neighboring farmers each expect a bumper crop of corn. Each will require his neighbor's help in harvesting his corn when it ripens, or else a substantial portion will rot in the field. Since their corn will ripen at different times, the two farmers can ensure full harvests for themselves by helping each other when their crops ripen, and both know this. Yet the farmers do not help each other. For the farmer whose corn ripens later reasons that if she were to help the other farmer, then when her corn ripens he would be in the position of Hobbes' Foole, having already benefited

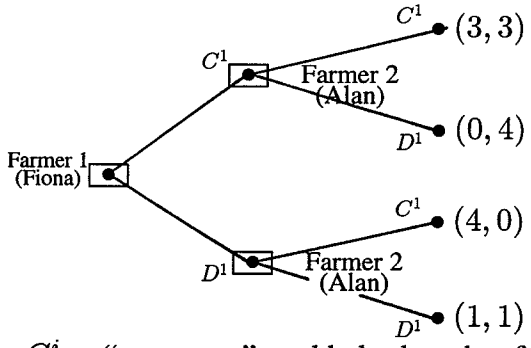
---

[3]Robert Vanderschraaf reminded me in conversation that a crucial assumption in this problem is that the cook is telling the diners the truth, that is, the cook's announcement generates common knowledge and not merely *common belief* that there is at least one messy individual. For if the agents believe the cook's announcements even if the cook does not reliably tell the truth, then should the cook mischievously announce that there is at least one messy individual when in fact all are clean, all will wipe their faces at once.

from her help. He would no longer have anything to gain from her, so he would not help her, sparing himself the hard labor of a second harvest. Since she cannot expect the other farmer to return her aid when the time comes, she will not help when his corn ripens first, and of course the other farmer does not help her when her corn ripens later.

The structure of Hume's *Farmers' Dilemma* problem can be summarized using the tree diagram of Figure 1.1.*a*.

Figure 1.1.*a*. The Farmers' Dilemma

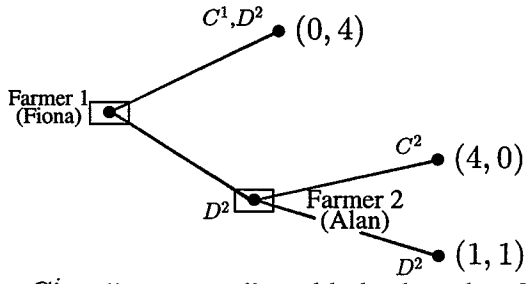


C^i = "cooperate", and help the other farmer
D^i = "defect", and leave the other farmer to work alone

This tree is an example of a *game in the extensive form*. At each stage i, the agent who moves can either choose C^i, which corresponds to helping or *cooperating*, or D^i, which corresponds to not helping or *defecting*. The relative preferences of the two agents over the various outcomes are reflected by the ordered pairs of *payoffs* each receives at any particular outcome. If, for instance, Fiona chooses C^1 and Alan chooses D^1, then Fiona's payoff is 0, her worst payoff, and Alan's is 4, his best payoff. In a game such as the Figure 1.1.*a* game, agents are *(Bayesian) rational* if each chooses an act that maximizes her expected payoff, given what she knows.

In the Farmers' Dilemma game, following the C^1, C^2-path is strictly better for both farmers than following the D^1, D^2-path. However, Fiona chooses D^1, as the result of the following simple argument: "If I were to choose C^1, then Alan, who is rational and who knows the payoff structure of the game, would choose D^2. I am also rational and know the payoff structure of the game. So I should choose D^1." Since Fiona knows that Alan is rational and knows the game's payoffs, she concludes that she need only analyze the *reduced* game of Figure 1.1.*b*.

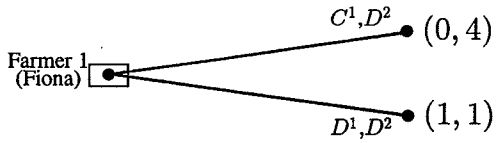
Figure 1.1.*b*.



C^i = "cooperate", and help the other farmer
D^i = "defect", and leave the other farmer to work alone

In this reduced game, Fiona is certain to gain a strictly higher payoff by choosing D^1 than if she chooses C^1, so D^1 is her unique best choice. Of course, when Fiona chooses D^1, Alan, being rational, responds by choosing D^2. If Fiona and Alan know: (i) that they are both rational, (ii) that they both know the payoff structure of the game, and (iii) that they both know (i) and (ii), then they both can predict what the other will do at every node of the Figure 1.1.a game, and conclude that they can rule out the D^1, C^2-branch of the Figure 1.1.b game and analyze just the reduced game of Figure 1.1.c.

Figure 1.1.*c*.



C = "cooperate", and help the other farmer
D = "defect", and leave the other farmer to work alone

On account of this *mutual knowledge*, both know that Fiona will choose D^1, and that Alan will respond with D^2. Hence, the D^1,D^2-outcome results if the Farmers' Dilemma game is played by agents having this mutual knowledge, though it is suboptimal since both agents would fare better at the C^1,C^2-branch.[4] This argument, which in its essentials is Hume's argument, is an example of a standard technique for solving sequential games known as *backwards induction*.[5] The basic idea behind backwards induction is that the agents engaged in a sequential game deduce how each will act throughout the entire game by ruling out the acts that are not payoff-maximizing for the agents who would move last, then ruling out the acts that are not payoff-maximizing for the agents who would move next-to-last, and so on. Clearly, backwards

---

[4]The mutual knowledge characterized by (i), (ii) and (iii) is sufficient both to account for the agents' following the D^1, D^2-outcome, and for their being able to predict each others' moves. However, weaker knowledge assumptions imply that the agents will play D^1, D^2, even if they might not both be able to predict this outcome before the start of play. As Fiona's quoted argument implies, if both are rational, both know the game, and Fiona knows that Alan is rational and knows the game, then the D^1, D^2-outcome is the result, even if Alan does not know that Fiona is rational or knows the game.

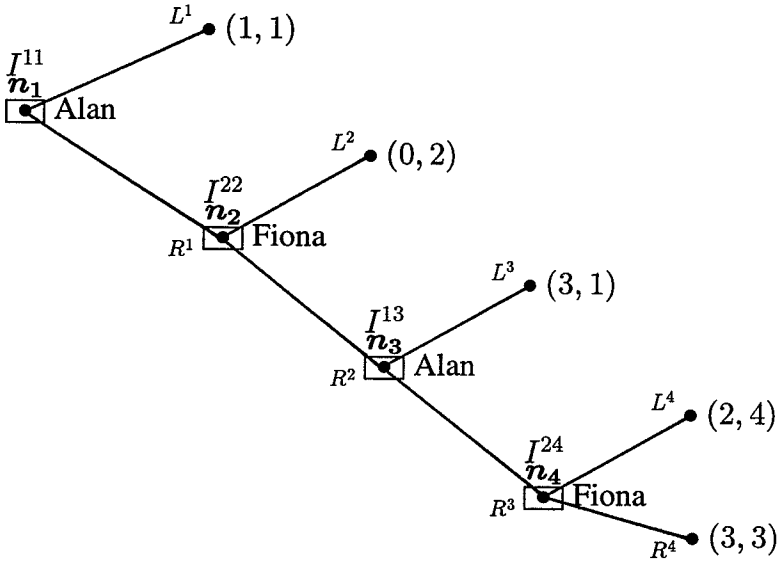
[5]Hume's analysis of the Farmers' Dilemma is perhaps the earliest example of a backwards induction argument applied to a sequential decision problem. See Skyrms (1996) and Vanderschraaf (1996) for more extended discussions of this argument.

induction arguments rely crucially upon what, if any, mutual knowledge the agents have regarding their situation, and they typically require the agents to evaluate the truth values of certain subjunctive conditionals, such as

If I (Fiona) were to choose C^1, then Alan would choose D^2.

The mutual knowledge assumptions required to construct a backwards induction solution to a game can become more complex as the number of stages in the game increases. To see this, consider the sequential game depicted in Figure 1.2. At each stage i, the agent who moves can either choose R^i, which in the first three stages gives the other agent an opportunity to move, or L^i, which ends the game.

Figure 1.2.



Like the Farmers' Dilemma, this game is a commitment problem for the agents. If each agent could trust the other to choose R^i at each stage, then they would each expect to receive a payoff of 3. However, Alan chooses L^1, leaving each with a payoff of only 1, as the result of the following backwards induction argument: "If node n_4 were to be reached, then Fiona, (being rational) would choose L^4. I, knowing this, would (being rational) choose L^3 if node n_3 were to be reached. Fiona, knowing *this*, would (being rational) choose L^2 if node n_2 were to be reached. Hence, I (being rational) should choose L^1." To carry out this backwards induction argument, Alan implicitly assumes that: (i) he knows that Fiona knows he is rational, and (ii) he knows that Fiona knows that he knows she is rational. Put another way, for Alan to carry out the backwards induction argument, at node n_1 he must know what Fiona must know at node n_2 to make L^2 her best response should n_2 be reached. While in the Farmer's Dilemma Fiona needed only *first-order* knowledge of Alan's rationality and *second-order* knowledge of Alan's knowledge of the game to derive the backwards induction solution, in the Figure 1.2 game, for Alan to be able to derive the backwards induction solution, the agents must have *third-order mutual knowledge* of the game and *second-order mutual knowledge* of rationality, and Alan must have *fourth-order* knowledge of this mutual knowledge of the game and *third-order* knowledge of their mutual knowledge of rationality.

This argument also involves several counterfactuals, since to construct it the agents must be able to evaluate conditionals of the form

If node n_i were to be reached, Alan (Fiona) would choose L^i (R^i).

which for $i > 1$ are counterfactual since third-order mutual knowledge of rationality implies that nodes n_2, n_3 and n_4 are never reached.

The method of backwards induction can be applied to any sequential game of *perfect information*, in which the agents can observe each others' moves in turn and can recall the entire history of play. However, as the number of potential stages of play increases, the backwards induction argument evidently becomes harder to construct. This raises certain questions below: (1) What precisely are the mutual or common knowledge assumptions that are required to justify the backwards induction argument for a particular sequential game? (2) As a sequential game increases in complexity, would we expect the mutual knowledge that is required for backwards induction to start to fail? □

Example 1.4. Coordination

The department store problem Schelling analyzes is an example of a *pure coordination problem*, that is, an interaction problem in which the interests of the agents coincide perfectly. Schelling (1960) and Lewis (1969), who were the first to make explicit the role common knowledge plays in social coordination, were also among the first to argue that coordination problems can be modeled using the analytic vocabulary of game theory. A very simple example of such a coordination problem is given in Figure 1.3.

Figure 1.3. The Department Store Problem

		Harold			
		s_1	s_2	s_3	s_4
Torrie	s_1	(1, 1)	(0, 0)	(0, 0)	(0, 0)
	s_2	(0, 0)	(1, 1)	(0, 0)	(0, 0)
	s_3	(0, 0)	(0, 0)	(1, 1)	(0, 0)
	s_4	(0, 0)	(0, 0)	(0, 0)	(1, 1)

s_i = search on floor i, $1 \leq i \leq 4$

The matrix of Figure 1.3 is an example of a *game in normal* or *strategic form*. At each outcome of the game, which corresponds to a cell in the matrix, the row (column) agent receives as payoff the first (second) element of the ordered pair in the corresponding cell. However, in strategic form games, each agent chooses without first being able to observe the choices of any other agent, so that all must choose as if they were choosing simultaneously. The Figure 1.3 game is a game of *pure coordination* (Lewis 1969), that is, a game in which at each outcome, each agent receives exactly the same payoff. One interpretation of this game is that Schelling's spouses, Torrie and Harold, are searching for each other in the department store, and there are four locations at which they might meet. Four outcomes at which the

spouses coordinate correspond to the strategy profiles (s_j, s_j), $1 \leq j \leq 4$, of the Figure 1.3 game. These four profiles are strict *Nash equilibria* (Nash 1950, 1951) of the game, that is, each agent has a decisive reason to follow her end of one of these strategy profiles provided that the other also follows this profile.[6] The difficulty the agents face is trying to select an equilibrium to follow. For suppose that Harold hopes to coordinate with Torrie on a particular equilibrium of the game, say (s_2, s_2). Harold reasons as follows: "Since there are several strict equilibria we might follow, I should follow my end of (s_2, s_2) if, and only if, I have sufficiently high expectations that Torrie will follow her end of (s_2, s_2). But I can only have sufficiently high expectations that Torrie will follow (s_2, s_2) if she has sufficiently high expectations that I will follow (s_2, s_2). For her to have such expectations, Torrie must have sufficiently high (second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2), for if Torrie doesn't have these (second-order) expectations, then she will believe I don't have sufficient reason to follow (s_2, s_2) and may therefore deviate from (s_2, s_2) herself. So I need to have sufficiently high (third-order) expectations that Torrie has sufficiently high (second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2). But this implies that Torrie must have sufficiently high (fourth-order) expectations that I have sufficiently high (third-order) expectations that Torrie has sufficiently high (second-order) expectations that I have sufficiently high expectations that she will follow (s_2, s_2), for if she doesn't, then she will believe I don't have sufficient reason to follow (s_2, s_2), and then she won't, either. Which involves me in fifth-order expectations regarding Torrie, which involves her in sixth-order expectations regarding me, and so on." What would suffice for Harold, and Torrie, to have decisive reason to follow (s_2, s_2) is that they each *know* that the other *knows* that . . . that the other will follow (s_2, s_2) for any number of levels of knowledge, which is to say that between Torrie and Harold it is *common knowledge*, as Lewis, Schiffer and Aumann define it, that they will follow (s_2, s_2). If agents follow a strict equilibrium in a pure coordination game as a consequence of their having common knowledge of the game, their rationality and their intentions to follow this equilibrium, and no other, then the agents are said to be following a *Lewis-convention* (Lewis 1969).

Lewis' theory of convention applies to a more general class of games than pure coordination games, but pure coordination games already model a variety of important social interactions. In particular, Lewis models conventions of language as equilibrium points of a pure coordination game. The role common knowledge plays in games of pure coordination sketched above of course raises further questions: (1) Can people ever attain the common knowledge which characterizes a Lewis-convention? (2) Would less stringent epistemic assumptions suffice to justify Nash equilibrium behavior in a coordination problem? □

Example 1.5. Coordination via E-mail

This example, due to Rubinstein (1987, 1992)[7], shows that a seemingly slight departure from common knowledge can dramatically change agents' prospects for successful coordination. Diane and Greta are faced with the coordination problem summarized by Figure 1.4. Note that their payoffs are dependent upon a pair of possible worlds. World ω_1 occurs with probability $\mu(\omega_1) = .51$, while ω_2 occurs with probability $\mu(\omega_2) = .49$. Hence, they coordinate with complete success by both choosing A (B) only if the state of the world is ω_1 (ω_2).

---

[6]See §3 for a formal definition of the Nash equilibrium concept.

[7]The version of the example Rubenstein presents is more general than the version presented here. Rubenstein notes that this game is closely related to the *coordinated attack problem* analyzed in Halpern (1986).

Figure 1.4. The E-mail Game

ω_1, $\mu(\omega_1) = .51$

Diane \ Greta	A	B
A	$(2, 2)$	$(0, -4)$
B	$(-4, 0)$	$(0, 0)$

ω_2, $\mu(\omega_2) = .49$

Diane \ Greta	A	B
A	$(0, 0)$	$(0, -4)$
B	$(-4, 0)$	$(2, 2)$

Suppose that Diane can observe the state of the world, but Greta cannot. We can interpret this game as follows: Greta and Diane would like to have a dinner together prepared by Aldo, their favorite chef. Aldo alternates between A and B, the two branches of Sorriso, their favorite restaurant. State ω_i is Aldo's location that day. At state ω_1 (ω_2), Aldo is at A (B). Diane, who is on Sorriso's special mailing list, receives notice of ω_i. Diane's and Greta's best outcome occurs when they meet where Aldo is working, so they can have their planned dinner. If they meet but miss Aldo, they are disappointed and do not have dinner after all. If either goes to A and finds herself alone, then she is again disappointed and does not have dinner. But what each really wants to avoid is going to B if the other goes to A. If either of them arrives at B alone, she not only misses dinner but must pay the exorbitant parking fee of the hotel which houses B, since the headwaiter of B refuses to validate the parking ticket of anyone who asks for a table for two and then sits alone. This is what Harsanyi (1967) terms a game of *incomplete information*, since the game's payoffs depend upon states which not all the agents know.

A is a "play-it-safe" strategy for both Greta and Diane.[8] By choosing A whatever the state of the world happens to be, the agents run the risk that they will fail to get the positive payoff of meeting where Aldo is, but each is also sure to avoid the really bad consequence of choosing B if the other chooses A. And since only Diane knows the state of the world, neither can use information regarding the state of the world to improve their prospects for coordination. For Greta has no such information, and since Diane knows this, she knows that Greta has to choose accordingly, so Diane must choose her best response to the move she anticipates Greta to make regardless of the state of the world Diane observes. Apparently Diane and Greta cannot achieve expected payoffs greater than 1.02 for each, their expected payoffs if they choose (A, A) at either state of the world.

If the state ω were common knowledge, then the conditional strategy profile (A, A) if $\omega = \omega_1$ and (B,B) if $\omega = \omega_2$ would be a strict Nash equilibrium at which each would achieve a payoff of 2. So the obvious remedy to their predicament would be for Diane to tell Greta Aldo's location in a face-to-face or telephone conversation and for them to agree to go where Aldo is, which would make the state ω and their intentions to coordinate on the best outcome given ω common knowledge between them. Suppose for some reason they cannot talk to each other, but they prearrange that Diane will send Greta an e-mail message if, and only if, ω_2 occurs. Suppose further that Greta's and Diane's e-mail systems are set up to send a reply

[8] In the terminology of decision theory, A is each agents' *maximin* strategy.

message automatically to the sender of any message received and viewed, and that due to technical problems there is a small probability, $\epsilon > 0$, that any message can fail to arrive at its destination. Then if Diane sends Greta a message, and receives an automatic confirmation, then Diane knows that Greta knows that ω_2 has occurred. If Greta receives an automatic confirmation of Diane's automatic confirmation, then Greta knows that Diane knows that Greta knows that ω_2 occurred, and so on. That ω_2 has occurred would become common knowledge if each agent received infinitely many automatic confirmations, assuming that all the confirmations could be sent and received in a finite amount of time.[9] However, because of the probability ϵ of transmission failure at every stage of communication, the sequence of confirmations stops after finitely many stages with probability one. With probability one, therefore, the agents fail to achieve full common knowledge. But they do at least achieve something "close" to common knowledge. Does this imply that they have good prospects of settling upon (B,B)?

Rubinstein shows that if the number of automatically exchanged confirmation messages is finite, then A is the only choice that maximizes expected utility for each agent, given what she knows about what they both know. Let T_2 denote the number of messages that Greta's e-mail system sends, and T_1 denote the number of messages that Diane's e-mail system sends. We might suppose that T_i appears on each agent's computer screen. If $T_1 = 0$, then Diane sends no message, that is, ω_1 has occurred, in which case Diane's unique best response is to choose A. If $T_2 = 0$, then Greta did not receive a message. She knows that in this case, either ω_1 has occurred and Diane did not send her a message, which occurs with probability .51, or ω_2 has occurred and Diane sent her a message which did not arrive, which occurs with probability $.49\epsilon$. If ω_1 has occurred, then Diane is sure to choose A, so Greta knows that whatever Diane might do at ω_2,

$$E(u_2(A) \mid T_2 = 0) \geq \frac{2(.51)+0(.49)\epsilon}{.51+.49\epsilon} > \frac{-4(.51)+2(.49)\epsilon}{.51+.49\epsilon} \geq E(u_2(B) \mid T_2 = 0)$$

so Greta is strictly better off choosing A no matter what Diane does at either state of the world. Suppose next that for all $T_i < t$, each agents' unique best response given her expectations regarding the other agent is A, so that the unique Nash equilibrium of the game is (A, A). Assume that $T_1 = t$. Diane is uncertain whether $T_2 = t$, which is the case if Greta received Diane's tth automatic confirmation and Greta's tth confirmation was lost, or if $T_2 = t - 1$, which is the case if Diane's tth confirmation was lost. Then $\mu_1(T_2 = t - 1 \mid T_1 = t) = z = \frac{\epsilon}{\epsilon+(1-\epsilon)\epsilon} > \frac{1}{2}$.[10] Thus it is more likely that Diane's last confirmation did not arrive than that Greta did receive this message. By the inductive assumption, Diane assesses that Greta will choose A if $T_2 = t - 1$. So

---

[9]This could be achieved if the e-mail systems were constructed so that each nth confirmation is sent 2^{-n} seconds after receipt of the nth message.

[10]If this does not look immediately obvious (and it did not to me!), consider that either

$E = [T_2 = t]$ = my tth confirmation was lost, or

$F = [T_2 = t]$ = my tth confirmation was received, and Greta's tth confirmation was lost

must occur, and that $p_1(T_1 = t \mid E) = p_1(T_1 = t \mid F) = 1$ because Diane can see her own computer screen, so so we can apply Bayes' Theorem as follows:

$$p_1(E \mid T_1 = t) = \frac{p_1(T_1=t|E)p_1(E)}{p_1(T_1=t|E)p_1(E)+p_1(T_1=t|F)p_1(F)} = \frac{p_1(E)}{p_1(E)+p_1(F)} = \frac{\epsilon}{\epsilon+(1-\epsilon)\epsilon}.$$

$$E(u_1(B) \mid T_1 = t) \leq -4z + 2(1-z) = -6z + 2 < -3 + 2 = -1\text{, and}$$
$$E(u_1(A) \mid T_1 = t) = 0 \text{ (since Diane knows that } \omega_2 \text{ is the case)}$$

so Diane's unique best action is A. Similarly, one can show that A is Greta's best reply if $T_2 = t$. So by induction, (A, A) is the unique Nash equilibrium of the game for every $t \geq 0$.

So even if agents have *"almost" common knowledge*, in the sense that the number of levels of knowledge in "Greta knows that Diane knows that . . . that Greta knows that ω_2 occurred." is very large, their behavior is quite different from their behavior given common knowledge that ω_2 has occurred. Indeed, as Rubinstein points out, given merely "almost" common knowledge, the agents choose as if no communication had occurred at all! Rubinstein also notes that this result violates our intuitions about what we would expect the agents to do in this case. If $T_i = 17$, wouldn't we expect agent i to choose B (Rubinstein 1992, p. 324)? Indeed, in many actual situations we might think it plausible that the agents would each expect the other to choose B even if $T_1 = T_2 = 2$, which is all that is needed for Diane to know that Greta has received her original message and for Greta to know that Diane knows this! □

§2. Alternative Accounts of Common Knowledge

Informally, a proposition A is *mutually known* among a set of agents if each agent knows that A. Mutual knowledge by itself implies nothing about what, if any, knowledge anyone attributes to anyone else. Suppose each student arrives for a class meeting knowing that the instructor will be late. That the instructor will be late is mutual knowledge, but each student might think only she knows the instructor will be late. However, if one of the students says openly "Peter told me he will be late again.", then the mutually known fact is now *commonly known*. Each student now knows that the instructor will be late, each student knows that each student knows that the instructor will be late, and so on, *ad infinitum*. The agents have common knowledge in the sense articulated informally by Schelling (1960), and more precisely by Lewis (1969) and Schiffer (1972). Schiffer uses the formal vocabulary of *epistemic logic* (Hintikka 1962) to state his definition of common knowledge. Schiffer's general approach was to augment a system of sentential logic with a set of knowledge operators corresponding to a set of agents, and then to define common knowledge as a hierarchy of propositions in the augmented system. Bacharach (1992) and Bicchieri (1993) adopt this approach, and develop logical theories of common knowledge which include soundness and completeness theorems. One can also develop alternate formal accounts of common knowledge in set-theoretic terms, which is the approach taken in this essay.[11]

The Hierarchical Account

Monderer and Samet (1988) and Binmore and Brandenburger (1989) give a particularly elegant set-theoretic definition of common knowledge. I will review this definition here, and then show that it is logically equivalent to the 'i knows that j knows that . . . k knows that A' hierarchy that Lewis (1969) and Schiffer (1972) argue characterizes common knowledge.[12]

---

[11]Aumann (1976) himself gives a set-theoretic account of common knowledge, which has been generalized in several articles in the literature, including Monderer and Samet (1988) and Binmore and Brandenburger (1989). Vanderschraaf (1997) gives the set-theoretic formulation of Lewis' account of common knowledge reviewed in this paper.

[12]This result appears in several articles in the literature, including Monderer's and Samet's and Binmore's and Brandenburger's articles on common knowledge.

Some preliminary notions must be stated first. Following C. I. Lewis (1943-1944) and Carnap (1947), propositions are formally subsets of a set Ω of *state descriptions* or *possible worlds*. One can think of the elements of Ω as representing Leibniz's possible worlds or Wittgenstein's possible states of affairs. Some results in the common knowledge literature presuppose that Ω is of finite cardinality. If this admittedly unrealistic assumption is needed in any context, this will be explicitly stated in this essay, and otherwise one may assume that Ω may be either a finite or an infinite set. A proposition $A \subseteq \Omega$ obtains (or is true) if the actual world $\omega \in \Omega$ is contained by A, that is, $\omega \in A$. Hence we say that A *obtains at* $\omega \in \Omega$ if $\omega \in A$. What an agent i knows about the possible worlds is stated formally in terms of a *knowledge operator* $\boldsymbol{K}_i$. Given a proposition $A \subseteq \Omega$, $\boldsymbol{K}_i(A)$ denotes a new proposition, corresponding to the set of possible worlds at which agent i knows that A obtains. $\boldsymbol{K}_i(A)$ is read as 'i knows (that) A (is the case)'. The knowledge operator $\boldsymbol{K}_i$ satisfies certain axioms, including:

$(K1)$ $\boldsymbol{K}_i(A) \subseteq A$

$(K2)$ $\Omega \subseteq \boldsymbol{K}_i(\Omega)$

$(K3)$ $\boldsymbol{K}_i(\bigcap_k A_k) = \bigcap_k \boldsymbol{K}_i(A_k)$

$(K4)$ $\boldsymbol{K}_i(A) \subseteq \boldsymbol{K}_i\boldsymbol{K}_i(A)$[13]

In words, $(K1)$ says that if i knows A, then A must be the case. $(K2)$ says that i knows that some possible world in Ω occurs no matter which possible world ω occurs. $(K3)$ says that i knows a conjunction if, and only if, i knows each conjunct. $(K4)$ is a *reflection axiom*, which says that if i knows A, then i knows that she knows A. Note that by $(K3)$, if $A \subseteq B$ then $\boldsymbol{K}_i(A) \subseteq \boldsymbol{K}_i(B)$, by $(K1)$ and $(K2)$, $\boldsymbol{K}_i(\Omega) = \Omega$, and by $(K1)$ and $(K4)$, $\boldsymbol{K}_i(A) = \boldsymbol{K}_i\boldsymbol{K}_i(A)$. Any system of knowledge satisfying $(K1)$-$(K4)$ corresponds to the modal system S_4 (Kripke 1963). If one were to relax the $(K1)$ axiom and retain the others, the resulting system would give a formal account of what an agent *believes*, but does not necessarily *know*.

A useful notion in the formal analysis of knowledge is that of a *possibility set*. An agent i's possibility set at a state of the world Ω is the smallest set of possible worlds that i thinks could be the case if ω is the actual world. More precisely,

Definition 2.1. Agent i's *possibility set* $P(\omega)$ at $\omega \in \Omega$ is defined as

$$\mathcal{H}_i(\omega) \equiv \bigcap\{E \mid \omega \in \boldsymbol{K}_i(E)\}$$

The collection of sets $\mathcal{H}_i = \bigcup_{\omega \in \Omega} \mathcal{H}_i(\omega)$ is i's *private information system.* □

Since in words, $\mathcal{H}_i(\omega)$ is the intersection of all propositions which i knows at ω, $\mathcal{H}_i(\omega)$ is the smallest proposition in Ω that i knows at ω. Put another way, $\mathcal{H}_i(\omega)$ is the most specific information that i has about the possible world ω. The intuition behind assigning agents private information systems is that while an agent i may not be able to perceive or comprehend every last detail of the world in which i lives, i does know certain facts about that world. The elements of i's information system represent what i knows immediately at a possible world. We also have the following

Proposition 2.2. $\boldsymbol{K}_i(A) = \{\omega \mid \mathcal{H}_i(\omega) \subseteq A\}$. □

---

[13] I abuse notation slightly, writing '$\boldsymbol{K}_i\boldsymbol{K}_j(A)$' for '$\boldsymbol{K}_i(\boldsymbol{K}_j(A))$'.

In many formal analyses of knowledge in the literature, possibility sets are taken as primitive and Proposition 2.2 is given as the definition of knowledge. If one adopts this viewpoint, then the axioms $(K1)$-$(K4)$ follow as consequences of the definition of knowledge. In many applications, the agents' possibility sets are assumed to *partition*[14] the set Ω, in which case $\mathcal{H}_i$ is called i's *private information partition.*

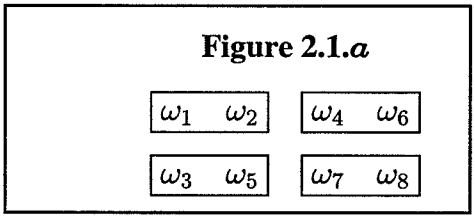
To illustrate the idea of possibility sets, let us return to the Barbecue Problem described in Example 1.2. Suppose there are three diners: Cathy, Jennifer and Mark. Then there are 8 relevant states of the world, summarized by Table 2.1.

Table 2.1

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8
Cathy	clean	messy	clean	clean	messy	messy	clean	messy
Jennifer	clean	clean	messy	clean	messy	clean	messy	messy
Mark	clean	clean	clean	messy	clean	messy	messy	messy

Each diner knows the condition of the other diners' faces, but not her own. Suppose the cook makes no announcement, after all. Then none of the diners knows the true state of the world whatever $\omega \in \Omega$ the actual world turns out to be, but they do know *a priori* that certain propositions are true at various states of the world. For instance, Cathy's information system before any announcement is made is depicted in Figure 2.1.*a*.

Figure 2.1.*a*



In this case, Cathy's information system is a partition $\mathcal{H}_1$ of Ω defined by

$$\mathcal{H}_1 = \{H_{CC}, H_{CM}, H_{MC}, H_{MM}\}$$

where

$H_{CC} = \{\omega_1, \omega_2\}$, that is, Jennifer and Mark are both clean.
$H_{CM} = \{\omega_4, \omega_6\}$, that is, Jennifer is clean and Mark is messy.
$H_{MC} = \{\omega_3, \omega_5\}$, that is, Jennifer is messy and Mark is clean.
$H_{MM} = \{\omega_7, \omega_8\}$, that is, Jennifer and Mark are both messy.

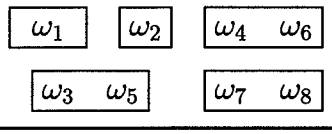
Cathy knows immediately which cell $\mathcal{H}_1(\omega)$ in her partition is the case at any state of the world, but does not know which is the true state at any $\omega \in \Omega$.

If we add in the assumption stated in Example 1.2 that if there is at least one messy diner, then the cook announces the fact, then Cathy's information partition is not depicted by Figure 2.1.*b*.

---

[14] A partition of a set Ω is a collection of sets $\mathcal{H} = \{H_1, H_2, \dots\}$ such that $H_i \cap H_j = \emptyset$ for $i \neq j$, and $\bigcup_i H_i = \Omega$.

Figure 2.1.*b*



In this case, Cathy's information system is a partition $\mathcal{H}_1$ of Ω defined by

$$\mathcal{H}_1 = \{H_{CCC}, H_{MCC}, H_{CM}, H_{MC}, H_{MM}\}$$

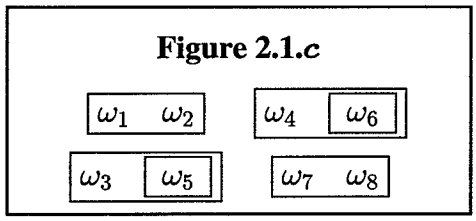
where

$H_{CCC} = \{\omega_1\}$, that is, Jennifer, Mark and I are all clean.
$H_{MCC} = \{\omega_2\}$, that is, Jennifer and Mark are clean and I am messy.
$H_{CM} = \{\omega_4, \omega_6\}$, that is, Jennifer is clean and Mark is messy.
$H_{MC} = \{\omega_3, \omega_5\}$, that is, Jennifer is messy and Mark is clean.
$H_{MM} = \{\omega_7, \omega_8\}$, that is, Jennifer and Mark are both messy.

In this case, Cathy's information partition is a *refinement* of the partition she has when there is no announcement, for in this case, then Cathy knows *a priori* that if ω_1 is the case there will be no announcement and will know immediately that she is clean, and Cathy knows *a priori* that if ω_2 is the case, then she will know immediately from the cook's announcement that she is messy.

A slightly more complex case occurs if we alter the Barbecue problem so that the cook makes an announcement only if case he sees at least two messy diners. Cathy's possibility set is now depicted by the diagram in Figure **2**.1.*c*.

Figure 2.1.*c*



This time, Cathy's information system does not partition Ω. For Cathy knows *a priori* that that at ω_5, the cook will make his announcement, and since at ω_5 Jennifer is messy and Mark is clean, Cathy will realize immediately that she is messy. However, Cathy also knows *a priori* that at ω_3, either ω_3 or ω_5 could be the case, since at ω_3 she does not know in advance whether or not the cook will make an announcement. Hence $\mathcal{H}_1(\omega_5) = \{\omega_5\}$, but $\mathcal{H}_1(\omega_3) = \{\omega_3, \omega_5\}$. Similarly, $\mathcal{H}_1(\omega_6) = \{\omega_6\}$, but $\mathcal{H}_1(\omega_4) = \{\omega_4, \omega_6\}$. Jennifer's and Mark's information systems given any of the above three scenarios are derived similarly to Cathy's information system, and the details of this are left as an exercise for the reader.

We can now define mutual and common knowledge as follows:

Definition 2.3. Let a set Ω of possible worlds together with a set of agents N be given.

(1) The proposition that A is (*first level* or *first order*) *mutual knowledge for the agents of* N, $\boldsymbol{K}_N^1(A)$, is the set defined by $\boldsymbol{K}_N^1(A) \equiv \bigcap_{i \in N} \boldsymbol{K}_i(A)$.

(2) The proposition that A is *mth level* (or *mth order*) *mutual knowledge among the agents of* N, $\boldsymbol{K}_N^m(A)$, is defined recursively as the set $\boldsymbol{K}_N^m(A) \equiv \bigcap_{i \in N} \boldsymbol{K}_i(\boldsymbol{K}_N^{m-1}(A))$.

(3) The proposition that A is *common knowledge* among the agents of N, $\boldsymbol{K}_N^*(A)$, is defined as the set $\boldsymbol{K}_N^*(A) \equiv \bigcap_{m=1}^{\infty} \boldsymbol{K}_N^m(A)$. □

As a consequence of Proposition 2.2, the agents' private information systems determine an *a priori* structure of propositions over the space of possible worlds regarding what they can know, including what mutual and common knowledge they potentially have. The world $\omega \in \Omega$ which obtains determines *a posteriori* what individual, mutual and common knowledge agents in fact have. Hence, one can read $\omega \in \boldsymbol{K}_i(A)$ as 'i knows A at (possible world) ω', $\omega \in \boldsymbol{K}_N^m(A)$ as 'A is mth level mutual knowledge for the agents of N at ω', and so on. If ω obtains, then one can conclude that i does know A, that A is mth level mutual knowledge, and so on. Common knowledge of a proposition E implies common knowledge of all that E implies, as is shown in the following:

Proposition 2.3. If $\omega \in \boldsymbol{K}_N^*(E)$ and $E \subseteq F$, then $\omega \in \boldsymbol{K}_N^*(F)$.

PROOF. If $E \subseteq F$, then as we observed earlier $\boldsymbol{K}_i(E) \subseteq \boldsymbol{K}_i(F)$, so

$$\boldsymbol{K}_N^1(E) = \bigcap_{i \in N} \boldsymbol{K}_i(E) \subseteq \bigcap_{i \in N} \boldsymbol{K}_i(F) = \boldsymbol{K}_N^1(F)\,.$$

If we not set $E' = \boldsymbol{K}_N^n(E)$ and $F' = \boldsymbol{K}_N^n(F)$, then by the argument just given we have

$$\boldsymbol{K}_N^{n+1}(E) = \boldsymbol{K}_N^1(E') \subseteq \boldsymbol{K}_N^1(F') = \boldsymbol{K}_N^{n+1}(F)$$

so we have mth level mutual knowledge for every $n \geq 1$. Hence if $\omega \in \bigcap_{n=1}^{\infty} \boldsymbol{K}_N^n(E)$ then $\omega \in \bigcap_{n=1}^{\infty} \boldsymbol{K}_N^n(F)$. □

Note that $(\boldsymbol{K}_N^m(E))_{m \geq 1}$ is a decreasing sequence of events, in the sense that $\boldsymbol{K}_N^{m+1}(E) \subseteq \boldsymbol{K}_N^m(E)$ for all $m \geq 1$. It is also easy to check that if everyone knows E, then E must be true, that is, $\boldsymbol{K}_N^1(E) \subseteq E$. If Ω is assumed to be finite, then if E is common knowledge at ω, this implies that there must be a finite m such that $\boldsymbol{K}_N^m(E) = \bigcap_{n=1}^{\infty} \boldsymbol{K}_N^m(E)$.

The following result relates the set-theoretic definition of common knowledge to the hierarchy of 'i knows that j knows that $\cdots k$ knows A' statements.

Theorem 2.4. $\omega \in \boldsymbol{K}_N^m(A)$ iff

(1) For all agents $i_1, i_2, \ldots, i_m \in N$, $\omega \in \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_m}(A)$.

Hence, $\omega \in \boldsymbol{K}_N^*(A)$ iff (1) is the case for each $m \geq 1$. □

PROOF. Note first that

$$(2) \qquad \bigcap_{i_1 \in N} \boldsymbol{K}_{i_1}\left(\bigcap_{i_2 \in N} \boldsymbol{K}_{i_2}\left(\cdots\left(\bigcap_{i_{m-1} \in N} \boldsymbol{K}_{i_{m-1}}\left(\bigcap_{i_m \in N} \boldsymbol{K}_{i_m}(A)\right)\right)\right)\right)$$

$$= \bigcap_{i_1 \in N} \boldsymbol{K}_{i_1}\left(\bigcap_{i_2 \in N} \boldsymbol{K}_{i_2}\left(\cdots\left(\bigcap_{i_{m-1} \in N} \boldsymbol{K}_{i_{m-1}}(\boldsymbol{K}_N^1(A))\right)\right)\right)$$

$$= \bigcap_{i_i \in N} \boldsymbol{K}_{i_1}\left(\bigcap_{i_2 \in N} \boldsymbol{K}_{i_2} \cdots \left(\bigcap_{i_{m-2} \in N} \boldsymbol{K}_{i_{m-2}}(\boldsymbol{K}_N^2(A))\right)\right)$$

$$= \cdots$$

$$= \bigcap_{i_i \in N} \boldsymbol{K}_{i_1}(\boldsymbol{K}_N^{m-1}(A)) = \boldsymbol{K}_N^m(A)$$

By (2), $\boldsymbol{K}_N^m(A) \subseteq \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_m}(A)$ for $i_1, i_2, \ldots, i_m \in N$, so if $\omega \in \boldsymbol{K}_N^m(A)$ then condition (1) is satisfied. Condition (1) is equivalent to

$$\omega \in \bigcap_{i_1 \in N} \boldsymbol{K}_{i_1}\left(\bigcap_{i_2 \in N} \boldsymbol{K}_{i_2}\left(\cdots\left(\bigcap_{i_2 \in N} \boldsymbol{K}_{i_{m-1}}\left(\bigcap_{i_m \in N} \boldsymbol{K}_{i_m}(A)\right)\right)\right)\right)$$

so by (2), if (1) is satisfied then $\omega \in K_N^m(A)$. □

The condition that $\omega \in \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_m}(A)$ for all $m \geq 1$ and all $i_1, i_2, \ldots, i_m \in N$ is Schiffer's definition of common knowledge, and is often used as the definition of common knowledge in the literature.

Lewis' Account

Lewis is credited with the idea of characterizing common knowledge as a hierarchy of 'i knows that j knows that ... k knows that A' propositions. However, it is far less well recognized that in *Convention*, Lewis also gives an algorithm which generates such a hierarchy from a finite set of assumptions regarding the agents' knowledge. These assumptions taken together constitute Lewis' official definition of common knowledge. Lewis' presentation of this definition and the algorithm is informal, and occasionally lacking in detail. It is probably for this reason that Aumann is often credited with presenting the first finitary method of generating the common knowledge hierarchy (Aumann 1976). A mathematically precise account of Lewis' analysis of common knowledge is given here, and it is shown that Lewis' analysis does result in the common knowledge hierarchy following from a finite set of axioms.

Lewis presents his account of common knowledge on pp. 52-57 of *Convention*. Lewis does not specify what account of knowledge is needed for common knowledge. As it turns out, Lewis' account is satisfactory for any formal account of knowledge in which the knowledge operators $\boldsymbol{K}_i$, $i \in N$, satisfy $(K1)$, $(K2)$ and $(K3)$. A crucial assumption in Lewis' analysis of common knowledge is that agents know they share the same "rationality, inductive standards and background information (Lewis 1969, p. 53)" with respect to a state of affairs A', that is, if an agent can draw any conclusion from A', she knows that all can do likewise. This idea is made precise in the following

Definition 2.5. Given a set of agents N and a proposition $A' \subseteq \Omega$, the agents of N are *symmetric reasoners with respect to* A' (or *A'-symmetric reasoners*) iff, for each $i, j \in N$ and for any proposition $E \subseteq \Omega$, if $\boldsymbol{K}_i(A') \subseteq \boldsymbol{K}_i(E)$ and $\boldsymbol{K}_i(A') \subseteq \boldsymbol{K}_i\boldsymbol{K}_j(A')$, then $\boldsymbol{K}_i(A') \subseteq \boldsymbol{K}_i\boldsymbol{K}_j(E)$. □[15]

The definiens says that for each agent i, if i can infer from A' that E is the case and that everyone knows that A' is the case, then i can also infer that everyone knows that E is the case.

---

[15]Thanks to Chris Miller and Jarah Evslin for suggesting the term 'symmetric reasoner' to describe the parity of reasoning powers that Lewis relies upon in his treatment of common knowledge. Lewis does not explicitly include the notion of A'-symmetric reasoning into his definition of common knowledge, but he makes use of the notion implicitly in his argument for how his definition of common knowledge generates the mutual knowledge hierarchy.

Definition 2.6. A proposition E is *Lewis-common knowledge at* $\omega \in \Omega$ among the agents of a set $N = \{1, \ldots, n\}$ iff there is a proposition A^* such that $\omega \in A^*$, the agents of N are A^*-symmetric reasoners, and for every $i \in N$,

$(L1)$ $\omega \in \boldsymbol{K}_i(A^*)$

$(L2)$ $\boldsymbol{K}_i(A^*) \subseteq \boldsymbol{K}_i(\bigcap_{j \in N} \boldsymbol{K}_j(A^*))$

$(L3)$ $\boldsymbol{K}_i(A^*) \subseteq \boldsymbol{K}_i(E)$

A^* is a *basis* for the agents' common knowledge. $\boldsymbol{L}^*_N(E)$ denotes the proposition defined by $(L1)$-$(L3)$ for a set N of A^*-symmetric reasoners, so we can say that E is Lewis-common knowledge for the agents of N iff $\omega \in \boldsymbol{L}^*_N(E)$. □

In words, $(L1)$ says that i knows A^* at ω. $(L2)$ says that if i knows that A^* obtains, then i knows that everyone knows that A^* obtains. This axiom is meant to capture the idea that common knowledge is based upon a proposition A^* that is *publicly known*, as is the case when agents hear a public announcement. If the agents' knowledge is represented by partitions, then a typical basis for the agents' common knowledge would be an element $\mathcal{M}(\omega)$ in the meet[16] of their partitions. $(L3)$ says that i can infer from A^* that E.

A human agent obviously cannot work her way mentally through an infinite mutual knowledge hierarchy. Lewis argues that this is not a problem for his analysis of common knowledge, since the mutual knowledge claims of a common knowledge hierarchy for a chain of logical consequences, not a series of steps in anyone's actual reasoning. Lewis uses an example to show how his definition of common knowledge generates the first few levels of mutual knowledge. In fact, Lewis' definition implies the entire common knowledge hierarchy, as is shown in the following result.

Proposition 2.7. $\boldsymbol{L}^*_N(E) \subseteq \boldsymbol{K}^*_N(E)$, that is, Lewis-common knowledge of E implies common knowledge of E.

PROOF. Suppose that $\omega \in \boldsymbol{L}^*_N(E)$. By definition, there is a basis proposition A^* such that $\omega \in A^*$. It suffices to show that for each $m \geq 1$ and for all agents $i_1, i_2, \ldots, i_m \in N$,

$$\omega \in \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_m}(E) .$$

We prove the result by induction on m. The $m = 1$ case follows at once from $(L1)$ and $(L3)$. Now if we assume that for $m = k$, $\omega \in \boldsymbol{L}^*_N(E)$ implies $\omega \in \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_k}(E)$, then $\boldsymbol{L}^*_N(E) \subseteq \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_k}(E)$ because ω is an arbitrary possible world, so $\boldsymbol{K}_{i_1}(A^*) \subseteq \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_k}(E)$ by $(L3)$. Since $(L2)$ is the case and the agents of N are A^*-symmetric reasoners, $\boldsymbol{K}_{i_1}(A^*) \subseteq \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_{k+1}}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_k}(E)$ for any $i_{k+1} \in N$, so $\omega \in \boldsymbol{K}_{i_1}\boldsymbol{K}_{i_{k+1}}\boldsymbol{K}_{i_2} \cdots \boldsymbol{K}_{i_k}(E)$ by $(L1)$, which completes the induction since $i_1, i_{k+1}, i_2, \ldots, i_k$ are $k + 1$ arbitrary agents of N. □

Aumann's Account

Aumann (1976) gives a different characterization of common knowledge which gives another simple algorithm for determining what information is commonly known. Aumann's original account assumes that the each agent's possibility set forms a private information partition of the space Ω of possible worlds. Aumann shows that a proposition C is common

---

[16]The *meet* $\mathcal{M}$ of a collection $\mathcal{H}_i$, $i \in N$, of partitions is the finest common coarsening of the partitions. More specifically, for any $\omega \in \Omega$, if $\mathcal{M}(\omega)$ is the element of $\mathcal{M}$ containing ω, then

(i) $\mathcal{H}_i(\omega) \subseteq \mathcal{M}(\omega)$ for all $i \in N$, and

(ii) For any other $\mathcal{M}'$ satisfying (i), $\mathcal{M}(\omega) \subseteq \mathcal{M}'(\omega)$.

knowledge if, and only if, C contains a cell of the meet of the agents' partitions. One way to compute the meet $\mathcal{M}$ of the partitions $\mathcal{H}_i, i \in N$ is to use the idea of "reachability".

Definition 2.8. A state $\omega' \in \Omega$ is *reachable* from $\omega \in \Omega$ iff there exists a sequence $\omega_0 \equiv \omega$, $\omega_1, \omega_2, \ldots, \omega_m \equiv \omega'$ such that for each $k \in \{0, 1, \ldots, m-1\}$, there exists an agent $i_k \in N$ such that $\mathcal{H}_{i_k}(\omega_k) = \mathcal{H}_{i_k}(\omega_{k+1})$. □

In words, ω' is reachable from ω if there exists a sequence or "chain" of states from ω to ω' such that two consecutive states are in the same cell of some agent's information partition. To illustrate the idea of reachability, let us return to the modified Barbecue Problem for in which Cathy, Jennifer and Mark receive no announcement. Their information partitions are all depicted in Figure **2.1.***d*.

Figure 2.1.*d*. Information Partitions in the Barbecue Problem

Cathy		Jennifer		Mark	
ω_1 ω_2	ω_4 ω_6	ω_1 ω_3	ω_2 ω_5	ω_1 ω_4	ω_2 ω_6
ω_3 ω_5	ω_7 ω_8	ω_4 ω_7	ω_6 ω_8	ω_3 ω_7	ω_4 ω_8

One can understand the importance of the notion of reachability in the following way: If ω' is reachable from ω, then if ω obtains then some agent can reason that some other agent thinks that ω' is possible. Looking at the diagram, if $\omega = \omega_1$ occurs, then Cathy (who knows only that $\{\omega_1, \omega_2\}$ has occurred) knows that Jennifer thinks that ω_5 might have occurred (even though Cathy knows that ω_5 did not occur). So Cathy cannot rule out the possibility that Jennifer thinks that Mark thinks that that ω_8 might have occurred. And Cathy cannot rule out the possibility that Jennifer thinks that Mark thinks that Cathy believes that ω_7 is possible. In this sense, ω_7 is reachable from ω_1. Note that one can show similarly in this example that any state is reachable from any other state.

Lemma 2.9. $\omega' \in \mathcal{M}(\omega)$ iff ω' is reachable from ω.

PROOF. Exercise. □

Lemma 2.10. $\mathcal{M}(\omega)$ is common knowledge for the agents of N at ω.

PROOF. Since $\mathcal{M}$ is a coarsening of $\mathcal{H}_i$ for each $i \in N$, $\boldsymbol{K}_i(\mathcal{M}(\omega))$. Hence, $\boldsymbol{K}_N^1(\mathcal{M}(\omega))$, and since by definition $\boldsymbol{K}_i(\mathcal{M}(\omega)) = \{\omega \mid \mathcal{H}_i(\omega) \subseteq \mathcal{M}(\omega)\} = \mathcal{M}(\omega)$,

$$\boldsymbol{K}_N^1(\mathcal{M}(\omega)) = \bigcap_{i \in N} \boldsymbol{K}_i(\mathcal{M}(\omega)) = \mathcal{M}(\omega) .$$

Applying the recursive definition of mutual knowledge, for any $m \geq 1$,

$$\boldsymbol{K}_N^m(\mathcal{M}(\omega)) = \bigcap_{i \in N} \boldsymbol{K}_i(\boldsymbol{K}_N^{m-1}(\mathcal{M}(\omega)) = \bigcap_{i \in N} \boldsymbol{K}_i(\mathcal{M}(\omega)) = \mathcal{M}(\omega)$$

so, since $\omega \in \mathcal{M}(\omega)$, by definition we have $\omega \in \boldsymbol{K}_N^*(\mathcal{M}(\omega))$. □

Theorem 2.11 (Aumann 1976). Let $\mathcal{M}$ be the meet of the agents' partitions $\mathcal{H}_i, i \in N$. A proposition $E \subseteq \Omega$ is common knowledge for the agents of N at ω iff $\mathcal{M}(\omega) \subseteq E$.

In Aumann (1976), E is *defined* to be common knowledge at ω iff $\mathcal{M}(\omega) \subseteq E$.

PROOF. ($\Leftarrow$) By Lemma 4, $\mathcal{M}(\omega)$ is common knowledge at ω, so E is common knowledge at ω by Proposition 2.

($\Rightarrow$) We must show that $\boldsymbol{K}_N^*(E)$ implies that $\mathcal{M}(\omega) \subseteq E$. Suppose that there exists $\omega' \in \mathcal{M}(\omega)$ such that $\omega' \notin E$. Since $\omega' \in \mathcal{M}(\omega)$, ω' is reachable from ω, so there exists a sequence $0, 1, \ldots, m-1$ with associated states $\omega_1, \omega_2, \ldots, \omega_m$ and information sets $\mathcal{H}_{i_k}(\omega_k)$ such that $\omega_0 = \omega$, $\omega_m = \omega'$ and $\omega_k \in \mathcal{H}_{i_k}(\omega_{k+1})$. But at information set $\mathcal{H}_{i_k}(\omega_m)$, agent i_k does not know event E. Working backwards on k, we see that event E cannot be common knowledge, that is, agent i_1 cannot rule out the possibility that agent i_2 thinks that $\cdots$ that agent i_{m-1} thinks that agent i_m does not know E. □

Note that the Proof of Theorem 2 required the use of only $(K1)$-$(K3)$. If $E = \boldsymbol{K}_N^1(E)$, then E is a *public event* (Milgrom 1981) or a *common truism* (Binmore and Brandenburger 1989). Clearly, a common truism is common knowledge whenever it occurs, since in this case $E = \boldsymbol{K}_N^1(E) = \boldsymbol{K}_N^2(E) = \cdots$, so $E = \boldsymbol{K}_N^*(E)$. The proof of Theorem 5 shows that the common truisms are precisely the elements of $\mathcal{M}$ and unions of elements of $\mathcal{M}$, so any commonly known event is the consequence of a common truism.

Gilbert's Account

Gilbert (1989, Chapter 3) presents an alternative account of common knowledge, which is meant to be more intuitively plausible than Lewis' and Aumann's accounts. Gilbert gives a highly detailed description of the circumstances under which agents have common knowledge.

Definition 2.13. A set of agents N are in a *common knowledge situation* $\mathcal{S}(A)$ with respect to a proposition A if, and only if, $\omega \in A$ and for each $i \in N$,

G_1 : i is *epistemically normal*, in the sense that i has normal perceptual organs which are functioning normally and has normal reasoning capacity.[17]

G_2 : i has the concepts needed to fulfill the other conditions.

G_3 : i perceives the other agents of N.

G_4 : i perceives that G_1 and G_2 are the case.

G_5 : i perceives that the state of affairs described by A is the case.

G_6 : i perceives that all the agents of N perceive that A is the case. □

There may appear to be some redundancy in Gilbert's definition, since presumably an agent would not perceive A unless A is the case. Gilbert is evidently trying to give a more explicit account of single agent knowledge than Lewis and Aumann give. For Gilbert, agent i knows that a proposition E is the case if, and only if, $\omega \in E$, that is, E is true, and either i perceives that the state of affairs E describes obtains or i can infer E as a consequence of other propositions i knows, given sufficient inferential capacity.

Like Lewis, Gilbert recognizes that human agents do not in fact have unlimited inferential capacity. To generate the infinite hierarchy of mutual knowledge, Gilbert introduces the device of an agent's *smooth-reasoner counterpart.* The smooth-reasoner counterpart i' of an agent i is an agent that draws every logical conclusion from every fact that i knows. Gilbert stipulates that i' does not have any of the constrains on time, memory, or reasoning ability that i might have, so i' can literally think through the infinitely many levels of a common knowledge hierarchy.

Definition 2.14. If a set of agents N are in a common knowledge situation $\mathcal{S}_N(A)$ with respect to A if, then the corresponding set N' of their smooth-reasoner counterparts is in a *parallel situation* $\mathcal{S}'_{N'}(A)$ if, and only if, for each $i' \in N$,

---

[17] Gilbert does not elaborate further on what counts as epistemic normality.

G_1' : i' can perceive anything that the counterpart i can perceive.
G_2' : G_2-G_6 obtain for i' with respect to A and N', same as for the counterpart i with respect to A and N.
G_3' : i' perceives that all the agents of N' are smooth-reasoners. □
From this definition we get the following immediate consequence:
Proposition 2.15. If a set N' of smooth-reasoner counterparts to a set N of agents are in a situation $\mathcal{S}'_{N'}(A)$ parallel to a common knowledge situation $\mathcal{S}_N(A)$ of N, then

$(*)$ For all $m \in \mathbb{N}$ and for any $i'_1, \ldots, i'_m$, $\boldsymbol{K}_{i'_1}\boldsymbol{K}_{i'_2} \cdots \boldsymbol{K}_{i'_m}(A)$,

so consequently, $\boldsymbol{K}^m_{N'}(A)$ for any $m \in \mathbb{N}$. □
Gilbert argues that, given $\mathcal{S}'_{N'}(A)$, the smooth-reasoner counterparts of the agents of N actually satisfy a much stronger condition, namely mutual knowledge $\boldsymbol{K}^\alpha_{N'}(A)$ to the level of any ordinal number α, finite or infinite. When this stronger condition is satisfied, the proposition A is said to be *open* to the agents of* N. With the concept of open*-ness, Gilbert gives her definition of common knowledge.
Definition 2.16. A proposition $E \subseteq \Omega$ is *Gilbert-common knowledge* among the agents of a set $N = \{1, \ldots, n\}$ if, and only if,
G_1^* : E is open* to the agents of N.
G_2^* : For every $i \in N$, $\boldsymbol{K}_i(G_1^*)$.
$\boldsymbol{G}^*_N(E)$ denotes the proposition defined by for a set N of A^*-symmetric reasoners, so we can say that E is Lewis-common knowledge for the agents of N iff $\omega \in \boldsymbol{G}^*_N(E)$. □

One might think that an immediate corollary to Gilbert's definition is that Gilbert-common knowledge implies the hierarchical common knowledge of Definition 2.3. However, this claim follows only on the assumption that an agent knows all of the propositions that her smooth-reasoner counterpart reasons through. Gilbert does not explicitly endorse this position, although she correctly observes that Lewis and Aumann are committed to something like it.[18] Gilbert maintains that her account of common knowledge expresses our intuitions with respect to common knowledge better than Lewis' and Aumann's accounts, since the notion of open*-ness presumably makes explicit that when a proposition is common knowledge, it is "out in the open", so to speak.

Common p-Belief

In certain contexts, agents might not be able to achieve common knowledge. Might they achieve something "close"? One weakening of common knowledge is of course mth level mutual knowledge. For a high value of m, $\boldsymbol{K}^m_N(A)$ might seem a good approximation of $\boldsymbol{K}^*_N(A)$. However, as the e-mail game of Example 1.5 shows, simply truncating the common knowledge hierarchy at any finite level can lead agents to behave as if they had no mutual knowledge at all. Brandenburger and Dekel (1987), Stinchcombe (1988) and Monderer and Samet (1989) explore another option, which is to weaken the properties of the $\boldsymbol{K}^*_N$ operator. Monderer and Samet motivate this approach by noting that even if a mutual knowledge hierarchy stops at a certain level, agents might still have higher level mutual *beliefs* about the

[18] Gilbert (1989, p. 193) also maintains that her account of common knowledge has the advantage of not requiring that the agents reason through an infinite hierarchy of propositions. On her account, the agents' smooth-reasoner counterparts do all the necessary reasoning for them. However, Gilbert fails to note that Aumann's and Lewis' accounts of common knowledge do not imply that agents must reason through the infinite mutual knowledge hierarchy, either.

proposition in question. So they replace the knowledge operator $\boldsymbol{K}_i$ with a *belief operator* $\boldsymbol{B}_i^p$:

Definition 2.17. If $\mu(\cdot)$ is agent i's probability distribution over Ω, then

$$\boldsymbol{B}_i^p(A) = \{\omega \mid \mu(A|\mathcal{H}_i(\omega)) \geq p\} \ . \ \square$$

$\boldsymbol{B}_i^p(A)$ is to be read 'i believes A (given i's private information) with probability at least p at ω', or 'i p-believes A'. The belief operator $\boldsymbol{B}_i^p$ satisfies axioms $(K2)$, $(K3)$, and $(K4)$ of the knowledge operator. $\boldsymbol{B}_i^p$ does not satisfy $(K1)$. but does satisfy the weaker property

$$(\boldsymbol{B}_i^p 1)\mu(A|\boldsymbol{B}_i^p(A)) \geq p$$

that is, if one believes A with probability at least p, then the probability of A is indeed at least p.

One can define *mutual* and *common p-beliefs* recursively in a manner similar to the definition of mutual and common knowledge:

Definition 2.18. Let a set Ω of possible worlds together with a set of agents N be given.

(1) The proposition that A is (*first level* or *first order*) *mutual p-belief for the agents of* N, $\boldsymbol{B}_N^{p\,1}(A)$, is the set defined by $\boldsymbol{B}_N^{p\,1}(A) \equiv \bigcap_{i \in N} \boldsymbol{B}_i^p(A)$.

(2) The proposition that A is *mth level* (or *mth order*) *mutual p-belief among the agents of* N, $\boldsymbol{B}_N^{p\,m}(A)$, is defined recursively as the set $\boldsymbol{B}_N^{p\,m}(A) \equiv \bigcap_{i \in N} \boldsymbol{B}_i^p(\boldsymbol{B}_N^{p\,m-1}(A))$.

(3) The proposition that A is *common p-belief* among the agents of N, $\boldsymbol{B}_N^{p\,*}(A)$, is defined as the set $\boldsymbol{B}_N^{p\,*}(A) \equiv \bigcap_{m=1}^{\infty} \boldsymbol{B}_N^{p\,m}(A)$. $\square$

If A is common (or mth level mutual) knowledge at world ω, then A is common (mth level) p-belief at ω for every value of p. So mutual and common p-beliefs formally generalize the mutual and common knowledge concepts. However, note that $\boldsymbol{B}_N^{1\,*}(A)$ is not necessarily the same proposition as $\boldsymbol{K}_N^*(A)$, that is, even if A is common 1-belief, A can fail to be common knowledge.

Common p-belief forms a hierarchy similar to a common knowledge hierarchy:

Theorem 2.19. $\omega \in \boldsymbol{B}_N^{p\,m}(A)$ iff

(1) For all agents $i_1, i_2, \ldots, i_m \in N$, $\omega \in \boldsymbol{B}_{i_1}^p \boldsymbol{B}_{i_2}^p \cdots \boldsymbol{B}_{i_m}^p(A)$.

Hence, $\omega \in \boldsymbol{B}_N^{p\,*}(A)$ iff (1) is the case for each $m \geq 1$.

PROOF. Similar to the proof of Theorem 2.μ. $\square$

§3. Applications of Multi-Agent Knowledge Concepts

Convention

Schelling's Department Store problem of Example 1.4 is a very simple example in which the agents "solve" their coordination problem appropriately by establishing a *convention*. Using the vocabulary of game theory, Lewis (1969) defines a convention as a *strict coordination equilibrium* of a game which agents follow on account of their common knowledge that they all prefer to follow this coordination equilibrium. A coordination equilibrium of a game is a strategy combination such that no agent is better off if any agent unilaterally deviates from this combination. As with equilibria in general, a coordination equilibrium is *strict* if any agent who deviates unilaterally from the equilibrium is strictly

worse off. The strategic form game of Figure 1.3 summarizes Torrie's and Harold's situation. The Department Store game has four Nash equilibrium outcomes in pure strategies: (s_1, s_1), (s_2, s_2), (s_3, s_3) and (s_4, s_4).[19] These four equilibria are all strict coordination equilibria. If the agents follow either of these equilibria, then they coordinate successfully. For agents to be following a Lewis-convention in this situation, they must follow one of the game's coordination equilibria. However, for Lewis to follow a coordination equilibrium is not a sufficient condition for agents to be following a convention. For suppose that Torrie and Harold fail to analyze their predicament properly at all, but Torrie chooses s_2 and Harold chooses s_2, so that they coordinate at (s_2, s_2) by sheer luck. Lewis does not count accidental coordination of this sort as a convention.

Suppose next that both agents are Bayesian rational, and that part of what each agent knows is the payoff structure of the Intersection game. If the agents expect each other to follow (s_2, s_2) and they consequently coordinate successfully, are they then following a convention? Not necessarily, contends Lewis, in a subtle argument on p. 59 of *Convention*. For while each knows the game and that she is rational, she might not attribute like knowledge to the other agent. If each agent believes that the other agent will follow her end of the (s_2, s_2) equilibrium mindlessly, then her best response is to follow her end of (s_2, s_2). But in this case the agents coordinated as the result of their each falsely believing that the other acts like an automaton, and Lewis thinks that any proper account of convention must require that agents have *correct* beliefs about one another. In particular, Lewis requires that each agent involved in a convention must have mutual expectations that each is acting with the aim of coordinating with the other, that is, that each knows that:

A_1: Both are rational. ,

A_2: Both know the payoff structure of the game. , *and*

A_3: Both intend to follow (s_2, s_2), and not some other strategy combination.

Suppose that the agents' beliefs are appropriately augmented so that each agent knows that A_1, A_2 and A_3 are the case. Again they coordinate on (s_2, s_2). Are they following a convention this time? Still not necessarily, says Lewis. For what if it turns out that Torrie thinks that Harold does not know that they are both rational? Then Torrie has a false belief about Harold. Beyond this, there are two other points which Lewis does not himself raise in this argument, but which clearly support his view. First, it would be counterintuitive, at the very least, to suppose that any agent following a convention believes that he has reasoning abilities that the other agents lack. If Torrie has determined that A_1, A_2 and A_3 are the case, then if they are following a convention she should expect that Harold has arrived at the same conclusion. Second, what could explain Torrie's knowledge of A_3? The most natural explanation for Torrie's expectation that Harold will follow his end of (s_2, s_2) is that Torrie knows that Harold knows that A_1, A_2 and A_3 are the case. So convention evidently involves agents having at least *second-order* mutual knowledge of A_1, A_2 and A_3, that is, Harold (Torrie) must know that Torrie (Harold) knows that A_1, A_2 and A_3 are the case. But this raises the question: Can

---

[19] An agent's *pure strategies* in a noncooperative game are simply the alternative acts this agent might choose as defined by the game. An agent follows a *mixed strategy* by observing the outcome of a random experiment and then choosing a pure strategy according to the outcome of this experiment. A strategy is *completely mixed* if before the experiment is performed, each pure strategy has a positive probability of being played.

third-order mutual knowledge that A_1, A_2 and A_3 obtain fail? No, argues Lewis. For if Harold thought that Torrie did not know that Harold knew that A_1, A_2 and A_3 were the case, then Harold would have a false belief about Torrie. The additional supporting points also kick in again: If Harold has second-order mutual knowledge that A_1, A_2 and A_3 obtain, then he should conclude that Torrie also has this second-order mutual knowledge. To conclude otherwise would require Harold to assume, counterintuitively, that he has analyzed their deliberations in this situation in a way that Torrie cannot. And how did Harold get his second-order mutual knowledge of A_3? The most obvious way to account for Harold's second-order mutual knowledge would be to attribute to Harold the knowledge that Torrie has second-order mutual knowledge that A_1, A_2 and A_3 are the case. So convention requires third-order mutual knowledge that A_1, A_2 and A_3 are the case. And the argument can be continued for any higher level of mutual knowledge.

Lewis concludes that a necessary condition for agents to be following a convention is that their preferences to follow the corresponding coordination equilibrium be common knowledge. So on Lewis' account, a convention for a set of agents is a coordination equilibrium which the agents follow on account of their common knowledge of their rationality, the payoff structure of the relevant game and that each agent follows her part of the equilibrium.

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among the members of P,
> (1) everyone conforms to R;
> (2) everyone expects everyone else to conform to R;
> (3) everyone has approximately the same preferences regarding all possible combinations of actions;
> (4) everyone prefers that everyone conform to R, on condition that at least all but one conform to R;
> (5) everyone would prefer that everyone conform to R', on condition that at least all but one conform to R',
> where R' is some possible regularity in the behavior of members of P in S, such that no one in any instance of S among members of P could conform both to R' and to R. (Lewis 1969, p. 76)[20]

Lewis includes the requirement that there be an alternate coordination equilibrium R' besides the equilibrium R that all follow in order to capture the fundamental intuition that how the agents who follow a convention behave depends crucially upon how they expect the others to behave. In the Department Store game, the (s_2, s_2) equilibrium is a Lewis-convention when Torrie and Harold have common knowledge of A_1, A_2 and A_3. Had their expectations been different, so either had believed that the other would not follow (s_2, s_2), then the outcome might have been very different.

---

[20]Lewis, (1969), p. 76. Lewis gives a further definition of agents following a convention to a *certain degree* if only a certain percentage of the agents actually conform to the coordination equilibrium corresponding to the convention. See Lewis (1969, pp. 78-79).

Sugden (1986) and Vanderschraaf (1997) argue that it is not crucial to the notion of convention that the corresponding equilibrium be a coordination equilibrium. Lewis' key insight is that a convention is a pattern of mutually beneficial behavior which depends on the agents' common knowledge that all follow *this* pattern, and no other. Vanderschraaf gives a more general definition of convention as a *strict* equilibrium together with common knowledge that all will follow this equilibrium, together with common knowledge that all might have followed a different equilibrium had their beliefs about each other been different. An example of this more general kind of convention is given below in the discussion of the Figure 3.1 example.

The "No Disagreement" Theorem

Aumann (1976) originally used his definition of common knowledge to prove a celebrated result that says that in a certain sense, agents cannot "agree to disagree" about their beliefs, formalized as probability distributions, if they start with common prior beliefs. Since agents in a community often hold different opinions and know they do so, one might attribute such differences to the agents' having different private information. Aumann's surprising result is that even if agents condition their beliefs on private information, mere common knowledge of their conditioned beliefs together with the common prior assumption (CPA) implies that their beliefs cannot be different, after all!

Theorem 3.1. Let Ω be a finite set of states of the world. Suppose that

(i) Agents i and j have a common prior probability distribution $\mu(\,\cdot\,)$ over the events of Ω such that $\mu(\omega) > 0$ for each $\omega \in \Omega$, and

(ii) It is common knowledge at ω that i's posterior probability of event E is $q_i(E)$ and that j's posterior probability of E is $q_j(E)$.

Then $q_i(E) = q_j(E)$.

PROOF. Let $\mathcal{M}$ be the meet of all the agents' partitions, and let $\mathcal{M}(\omega)$ be the element of $\mathcal{M}$ containing ω. Since $\mathcal{M}(\omega)$ consists of cells common to every agents information partition, we can write $\mathcal{M}(\omega) = \bigcup_k H_{ik}$, where each $H_{ik} \in \mathcal{H}_i$. Since i's posterior probability of event E is common knowledge, it is constant on $\mathcal{M}(\omega)$, and so

$$q_i(E) = \mu(E \mid H_{ik}) \text{ for all } k\,.$$

Hence,

$$\mu(E \cap H_{ik}) = q_i(E)\mu(H_{ik})$$

and so

$$\begin{aligned}\mu(E \cap \mathcal{M}(\omega)) &= \mu(E \cap \bigcup_k H_{ik}) = \mu(\bigcup_k E \cap H_{ik}) \\ &= \sum_k \mu(E \cap H_{ik}) = \sum_k q_i(E)\mu(H_{ik}) \\ &= q_i(E)\sum_k \mu(H_{ik}) = q_i(E)\mu(\bigcup_k H_{ik}) = q_i(E)\mu(\mathcal{M}(\omega))\end{aligned}$$

Applying the same argument to j, we have

$$\mu(E \cap \mathcal{M}(\omega)) = q_j(E)\mu(\mathcal{M}(\omega))$$

so we must have $q_i(E) = q_j(E)$. □

In a later article, Aumann (1987) argues that the assumptions that Ω is finite and that $\mu(\omega) > 0$ for each $\omega \in \Omega$ reflect the idea that agents only regard as "really" possible a finite collection of salient worlds to which they assign positive probability, so that one can drop the

states with probability 0 from the description of the state space. Aumann also notes that this result implicitly assumes that the agents have common knowledge of their partitions, since a description of each possible world includes a description of the agents' possibility sets. And of course, this result depends crucially upon (*i*),the CPA.

Aumann's "no disagreement" theorem has been generalized in a number of ways in the literature (McKelvey and Page 1986, Monderer and Samet 1989, Geanakoplos 1994). However, all of these "no disagreement" results raise the same philosophical puzzle raised by Aumann's original result: How are we to explain differences in belief? Aumann's result leaves us with two options: (1) admit that at some level, common knowledge of the agents' beliefs or how they form their beliefs fails, or (2) deny the CPA. For instance, agents in the real world often do not express their opinions probabilistically. If one agent announces 'I believe that E is the case.' while another announces 'I doubt that E is the case.', then they might attribute their divergent opinions to a lack of common knowledge of each other's true posteriors for E. Even if agents do assign precise posterior probabilities to an event, Aumann shows that if they have merely first-order mutual knowledge of the posteriors, they can "agree to disagree". Suppose that $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, that $\mathcal{H}_1 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$ and $\mathcal{H}_2 = \{\{\omega_1, \omega_2, \omega_3\}, \{\omega_4\}\}$, and that $\mu(\omega_i) = \frac{1}{4}$. Then if $E = \{\omega_1, \omega_4\}$, then at ω_1,

$$q_1(E) = \mu(E \mid \{\omega_1, \omega_2\}) = \tfrac{1}{2} \text{ and } q_2(E) = \mu(E \mid \{\omega_1, \omega_2, \omega_3\}) = \tfrac{1}{3}\,.$$

Moreover, at $\omega = \omega_1$ agent 1 knows that $\mathcal{H}_2(\omega) = \{\omega_1, \omega_2, \omega_3\}$, so she knows that $q_2(E) = \frac{1}{3}$. Agent 2 knows at ω_1 that either $\mathcal{H}_1(\omega) = \{\omega_1, \omega_2\}$ or $\mathcal{H}_1(\omega) = \{\omega_3, \omega_4\}$, so either way he knows that $q_1(E) = \frac{1}{2}$. Hence the agents' posteriors are mutually known, and yet they are unequal. The reason for this is that the posteriors are not common knowledge. For agent 2 does not know what agent 1 thinks $q_2(E)$ is, since if $\omega = \omega_3$, which is consistent with what agent 2 knows, then agent 1 will believe that $q_2(E) = \frac{1}{3}$ with probability $\frac{1}{2}$ (if $\omega = \omega_3$) and $q_2(E) = 1$ with probability $\frac{1}{2}$ (if $\omega = \omega_4$). Aumann's result could fail if the agents' partitions are not common knowledge. For suppose in the example just given, the agents do not know each other's partitions. Then at $\omega = \omega_1$, if their posteriors are common knowledge, then agent 1, who knows that $\omega \in \{\omega_1, \omega_2\}$, can explain agent 2's posterior as the result of agent 2 having observed either $\{\omega_1, \omega_2, \omega_3\}$, $\{\omega_1, \omega_2, \omega_4\}$, $\{\omega_1, \omega_3, \omega_4\}$ or $\{\omega_2, \omega_3, \omega_4\}$. Still another way Aumann's result might fail is if agents do not have common knowledge that they update their beliefs by Bayesian conditionalization. Then clearly, agents can explain divergent opinions as the result of others having modified their beliefs in the "wrong" way. However, there are cases in which none of these explanations will seem convincing. For instance, odds makers sometimes publicly announce different probabilities for an event, such as a particular winner of a prize at a forthcoming Academy Awards presentation, and they will know that none of them have *any* private information regarding the event. In cases such as this, the agents have common knowledge that they all have the same information structure and common knowledge of their posteriors. And knowing that they are all competent odds makers, they have common knowledge that they update by Bayesian conditionalization. Still, the odds makers' beliefs violate the conclusion of Aumann's result. More generally, denying the requisite common knowledge seems a rather *ad hoc* move. For instance, to deny that agents have common knowledge of information structures is simply to deny that agents can all infer the same conclusions regarding possible worlds as Aumann defines them. To deny that agents have common knowledge that they update their beliefs by Bayesian conditionalization is to assert

that some believe that some might be updating their beliefs *incoherently*, in the sense that their belief updating leaves them open to a *Dutch book* (Skyrms 1984). As just noted, these failures of agents' beliefs in each others' competence do not fail in all cases. Why should one think that such failures of common knowledge provide a general explanation for divergent beliefs?

What of the second option, that is, denying the CPA?[21] The main argument put forward in favor of the CPA is that any differences in agents' probabilities should be the result of their having different information only, that is, there is no reason to think that the different beliefs that agents have regarding the same event are the result of anything other than their having different information. However, one can reply that this argument amounts simply to a restatement of the Harsanyi Doctrine.[22] And while defenders of the Harsanyi Doctrine may be right in thinking that there is apparently no compelling reason to think that agents' priors can be different, neither is there compelling reason to think they must be the same! In any event, while the controversy over the Harsanyi Doctrine remains unresolved, we can conclude that the "no disagreement" results have interesting implications for the viability of common knowledge and the very nature of probability. Defenders of the CPA take an *objectives* view of probability, and by virtue of the "no disagreement" results are evidently committed to the view that common knowledge of agents beliefs and how they are formed is a rare phenomenon in the world. Those who are prepared to deny the CPA allow for a genuinely *subjectivist* conception of probability. They take the view that common knowledge of agents' beliefs and how they come by them can be a commonplace phenomenon, and that differences in opinion can stem from differences in (subjective) prior probabilities.

Strategic Form Games

Lewis formulated the notion of common knowledge as part of his general account of conventions. In the years following the publication of *Convention*, game theorists have recognized that any explanation of a particular pattern of play in a game depends crucially on mutual and common knowledge assumptions. More specifically, *solution concepts* in game theory are both motivated and justified in large part by the mutual or common knowledge the agents in the game have regarding their situation. A modest starting point is to assume that the agents are sophisticated enough to have common knowledge of the full payoff structure of the game they are engaged in and that they are all rational. Suppose further that no other information is common knowledge. In other words, each agent knows that her opponents are expected utility maximizers, but does not in general know exactly which strategies they will choose or what their probabilities for her acts are. These common knowledge assumptions are the motivational basis for the solution concept for noncooperative games known as *rationalizability*, introduced independently by Bernheim (1984) and Pearce (1984). Roughly speaking, a *rationalizable strategy* is any strategy an agent may choose without violating common knowledge of Bayesian rationality. Bernheim and Pearce argue that when only the structure of the game and the agents' Bayesian rationality are common knowledge, the game should be considered "solved" if every agent plays a rationalizable strategy. For instance, in the "Chicken" game with payoff structure defined by Figure 3.1, if Kay and Amie have common knowledge of all of the payoffs at every strategy combination, and they have common

---

[21]Harsanyi (1968) is the most famous defender of the CPA. Indeed, Aumann (1974, 1987) calls the CPA the *Harsanyi Doctrine* in Harsanyi's honor.

[22]Alan Hajek first pointed this out to me in conversation.

knowledge that both are Bayesian rational, then any of the four pure strategy profiles is rationalizable.

Figure 3.1. Chicken

		Kay s_1	Kay s_2
Amie	s_1	(3, 3)	(2, 4)
	s_2	(4, 2)	(0, 0)

s_1 = cooperate, s_2 = defect

For if their beliefs about each other are defined by the probabilities

$$\alpha_1 = \mu_1(\text{Kay plays } s_1), \text{ and } \alpha_2 = \mu_2(\text{Amie plays } s_1)$$

then

$$E(u_i(s_1)) = 3\alpha_i + 2(1 - \alpha_i) = \alpha_i + 1, \text{ and } E(u_i(s_2)) = 4\alpha_i + 0(1 - \alpha_i) = 4\alpha_i,\ i = 1, 2$$

so each agent maximizes her expected utility by playing s_1 if $\alpha_i + 1 \geq 4\alpha_i$ or $\alpha_i \leq \frac{1}{3}$ and maximizes her expected utility by playing s_2 if $\alpha_i \geq \frac{1}{3}$. If it so happens that $\alpha_i > \frac{1}{3}$ for both agents, then both conform with Bayesian rationality by playing their respective ends of the strategy combination (s_2, s_2) *given their beliefs*, even though each would want to defect from this strategy combination were she to discover that the other is in fact going to play s_2. Note that the game's pure strategy Nash equilibria, (s_1, s_2) and (s_2, s_1), are rationalizable, since it is rational for Amie and Kay to conform with either equilibrium given appropriate distributions. In general, the set of a game's rationalizable strategy combinations contains the set of the game's pure strategy Nash equilibria, and this example shows that the containment can be proper.

To show that rationalizability is a nontrivial notion, consider the 2-agent game with payoff structure defined by Figure 3.2.*a*.

Figure 3.2.*a*

		Kay s_1	Kay s_2	Kay s_3
Amie	s_1	(4,3)	(1,2)	(3,4)
	s_2	(1,1)	(0,5)	(1,1)
	s_3	(3,4)	(1,3)	(4,3)

In this game, s_1 and s_3 strictly dominate s_2 for Amie, so Amie cannot play s_2 on pain of violating Bayesian rationality. Kay knows this, so Kay knows that the only pure strategy profiles which are possible outcomes of the game will be among the six profiles in which Amie does not choose s_2. In effect, the 3×3 game is reduced to the 2×3 game defined by Figure 3.2.*b*.

Figure 3.2.*b*

		Kay		
		s_1	s_2	s_3
Amie	s_1	(4,3)	(1,2)	(3,4)
	s_3	(3,4)	(1,3)	(4,3)

In this reduced game, s_2 is strictly dominated for Kay by s_1, and so Kay will rule out playing s_2. Amie knows this, and so she rules out strategy combinations in which Kay plays s_2. The rationalizable strategy profiles are the four profiles that remain after deleting all of the strategy combinations in which either Amie or Kay play s_2. In effect, common knowledge of Bayesian rationality reduces the 3×3 game of Figure 3.2.*a* to the 2×2 game defined by Figure 3.2.*c*, since Amie and Kay both know that the only possible outcomes of the game are (s_1, s_1), (s_1, s_3), (s_3, s_1) and (s_3, s_3).

Figure 3.2.*c*

		Kay	
		s_1	s_3
Amie	s_1	(4,3)	(3,4)
	s_3	(3,4)	(4,3)

Rationalizability can be defined formally in several ways. A variation of Bernheim's original (1984) definition is given here. I first give the usual definitions of a game in strategic form, expected utility and agents' distributions over their opponents' strategies to establish notation.

Definition 3.2. A *game* Γ is an ordered triple $(N,S,\boldsymbol{u})$ consisting of the following elements:

(*a*) A finite set $N = \{1, 2, \ldots, n\}$, called the *set of agents* or *players*.

(*b*) For each agent $k \in N$, there is a finite set $S_k = \{s_{k1}, s_{k2}, \ldots, s_{kn_k}\}$, called the *alternative pure strategies* for agent k. The Cartesian product $S = S_1 \times \cdots \times S_n$ is called the *pure strategy set* for the game Γ.

(*c*) A map $\boldsymbol{u}$: $S \to \mathbb{R}^n$, called the *utility* or *payoff function* on the pure strategy set. At each strategy combination $\boldsymbol{s} = (s_{1j_1}, \ldots, s_{nj_n}) \in S$, agent k's particular payoff or utility is given by the kth component of the value of $\boldsymbol{u}$, that is, agent k's utility u_k at $\boldsymbol{s}$ is determined by

$$u_k(\boldsymbol{s}) = I_k \circ \boldsymbol{u}(s_{1j_1}, \ldots, s_{nj_n}) ,$$

where $I_k(\boldsymbol{x})$ projects $\boldsymbol{x} \in \mathbb{R}^n$ onto its kth component. □

The subscript ' $-k$' indicates the result of removing the kth component of an n-tuple or an n-fold Cartesian product. For instance,

$$S_{-k} = S_1 \times \cdots \times S_{k-1} \times S_{k+1} \times \cdots \times S_n$$

denotes the pure strategy combinations that agent k's opponents may play.

Now let us formally introduce a system of the agents' beliefs into this framework. $\Delta_k(S_{-k})$ denotes the set of probability distributions over the measurable space $(S_{-k}, \mathfrak{F}_k)$, where $\mathfrak{F}_k$ denotes the Boolean algebra generated by the strategy combinations S_{-k}. Each agent k has a probability distribution $\mu_k \in \Delta_k(S_{-k})$, and this distribution determines the *(Savage) expected utilities* for each of k's possible acts:

$$E(u_k(s_{kj})) = \sum_{\boldsymbol{A}_{-k} \in S_{-k}} u_k(s_{kj}, \boldsymbol{s}_{-k}) \mu_k(\boldsymbol{s}_{-k}) \ , \ j = 1, 2, \ldots, n_k \ .$$

If i is an opponent of k, then i's individual strategy s_{ij} may be characterized as a union of strategy combinations $\bigcup\{\boldsymbol{s}_{-k} \mid s_{ij} \in \boldsymbol{s}_{-k}\} \in \mathfrak{F}_k$, and so k's marginal probability for i's strategy s_{ij} may be calculated as follows:

$$\mu_k(s_{ij}) = \sum_{\{s_{-k} | s_{ij} \in s_{-k}\}} \mu_k(s_{-k}) \ .$$

Definition 3.3. Given that each agent $k \in N$ has a probability distribution $\mu_k \in \Delta_k(S_{-k})$, the system of beliefs

$$\mu = (\mu_1, \ldots, \mu_n) \in \Delta_1(S_{-1}) \times \cdots \times \Delta_n(S_{-n})$$

is *Bayes concordant* if, and only if,

(3.i) For $i \neq k$, $\mu_i(s_{kj}) > 0 \Rightarrow s_{kj}$ maximizes k's expected utility for some $\sigma_k \in \Delta_k(S_{-k})$,

and (3.i) is common knowledge. A pure strategy combination $\boldsymbol{s} = (s_{1j_1}, \ldots, s_{nj_n}) \in S$ is *rationalizable* if, and only if, the agents have a Bayes concordant system μ of beliefs and, for each agent $k \in N$,

(3.ii) $$E(u_k(s_{kj_k})) \geq E(u_k(s_{ki_k})) \text{ for } i_k \neq j_k \ .\square$$[23]

---

[23]In their original papers, Bernheim (1984) and Pearce (1984) included in their definitions of rationalizability the requirement that the agents' probability distributions over their opponents satisfy *probabilistic independence*, that is, for each agent k and for each

$$\boldsymbol{s}_{-k} = (s_{1j_1}, \ldots, s_{k-1j_{k-1}}, s_{k+1j_{k+1}}, \ldots, s_{nj_n}) \in S_{-k}$$

k's joint probability must equal the product of k's marginal probabilities, that is,

$$p_k(\boldsymbol{s}_{-k}) = p_k(s_{1j_1}) \cdot \cdots p_k(s_{k-1j_{k-1}}) \cdot p_k(s_{k+1j_{k+1}}) \cdot \cdots p_k(s_{nj_n}) \ .$$

Brandenburger and Dekel (1987), Skyrms (1990) and Vanderschraaf (1995) all argue that the probabilistic independence requirement is not well-motivated, and do not include this requirement in their presentations of rationalizability.

Bernheim (1984) calls a Bayes concordant system of beliefs a "consistent" system of beliefs. Since the term "consistent beliefs" is used in this paper to describe probability distributions that agree with respect to a mutual opponent's strategies, I use the term "Bayes concordant system of beliefs" rather than Bernheim's "consistent system of beliefs".

The following result shows that the common knowledge restriction on the distributions in Definition 3.1 formalizes the assumption that the agents have common knowledge of Bayesian rationality.

Proposition 3.4. In a game Γ, common knowledge of Bayesian rationality is satisfied if, and only if, $(3.i)$ is common knowledge.

PROOF. Suppose first that common knowledge of Bayesian rationality is satisfied. Since it is common knowledge that agent i knows that agent k is Bayesian rational, it is also common knowledge that if $\mu_i(s_{kj}) > 0$, then s_{kj} must be optimal for k given some belief over S_{-k}, so $(3.i)$ is common knowledge.

Suppose now that $(3.i)$ is common knowledge. Then, by $(3.i)$, agent i knows that agent k is Bayesian rational. Since $(3.i)$ is common knowledge, all statements of the form 'For $i, j, \ldots, k \in N$, i knows that j knows that $\ldots$ k is Bayesian rational' follow by induction. □

When agents have common knowledge of the game and their Bayesian rationality only, one can predict that they will follow a rationalizable strategy profile. However, rationalizability becomes an unstable solution concept if the agents come to know more about one another. For instance, in the Chicken example above with $\alpha_i > \frac{1}{3}$, $i = 1, 2$, if either agent were to discover the other agent's beliefs about her, she would have good reason not to follow the (s_2, s_2) profile and to revise her own beliefs regarding the other agent. If, in the other hand, it so happens that $\alpha_1 = 1$ and $\alpha_2 = 0$, so that the agents maximize expected payoff by following the (s_2, s_1) profile, then should the agents discover their beliefs about each other, they will still follow (s_2, s_1). Indeed, if their beliefs are common knowledge, then one can predict with certainty that they will follow (s_2, s_1). The Nash equilibrium (s_2, s_1) is characterized by the belief distributions defined by $\alpha_1 = 1$ and $\alpha_2 = 0$.

The Nash equilibrium is a special case of *correlated equilibrium concepts*, which are defined in terms of the belief distributions of the agents in a game. In general, a correlated equilibrium-in-beliefs is a system of agents' probability distributions which remains stable given common knowledge of the game, rationality and the *beliefs themselves*. We will review two alternative correlated equilibrium concepts (Aumann 1974, 1987; Vanderschraaf 1995), and show how each generalizes the Nash equilibrium concept.

Definition 3.4. Given that each agent $k \in N$ has a probability distribution $\mu_k \in \Delta_k(S_{-k})$, the system of beliefs

$$\mu^* = (\mu_1^*, \ldots, \mu_n^*) \in \Delta_1(S_{-1}) \times \cdots \times \Delta_n(S_{-n}),$$

is an *endogenous correlated equilibrium* if, and only if,

$(3.iii)$ For $i \neq k$, $\mu_i^*(s_{kj}) > 0 \Rightarrow s_{kj}$ maximizes k's expected utility given μ_k^*. □

If μ^* is an endogenous correlated equilibrium, a pure strategy combination $\boldsymbol{s}^* = (s_1^*, \ldots, s_n^*) \in S$ is an *endogenous correlated equilibrium strategy combination given* μ^* if, and only if, for each agent $k \in N$,

$$(3.iv) \qquad E(u_k(s_k^*)) \geq E(u_k(s_{ki})) \text{ for } s_{ki} \neq s_k^*.$$

Hence, the endogenous correlated equilibrium μ^* restricts the set of strategies that the agents might follow, as do the Bayes concordant beliefs of rationalizability. However, the endogenous correlated equilibrium concept is a proper refinement of rationalizability, because the latter does not presuppose that condition $(3.iii)$ holds with respect to the beliefs one's opponents actually have. If exactly one pure strategy combination $\boldsymbol{s}^*$ satisfies $(3.iv)$ given μ^*,

then μ^* is a *strict equilibrium*, and in this case one can predict with certainty what the agents will do given common knowledge of the game, rationality and their beliefs.

Note that Definition 3.4 says nothing about whether or not the agents regard their opponents' strategy combinations as probabilistically independent. Also, this definition does not require that the agents' probabilities are *consistent*, in the sense that agents' probabilities for a mutual opponent's acts agree. A simple refinement of the endogenous correlated equilibrium concept characterizes the Nash equilibrium concept.

Definition 3.5. A system of agents' beliefs μ^* is a *Nash equilibrium* if, and only if,

(a) Condition $(3.iii)$ is satisfied,

(b) For each $k \in N$, μ_k^* satisfies probabilistic independence, and

(c) For each $s_{kj} \in S_k$, if $i, l \neq k$ then $\mu_i^*(s_{kj}) = \mu_l^*(s_{kj})$. □

In other words, an endogenous correlated equilibrium is a Nash equilibrium-in-beliefs when each agent regards the moves of his opponents as probabilistically independent and the agents' probabilities are consistent. Note that in the 2-agent case, conditions (b) and (c) of the Definition 3.5 are always satisfied, so for 2-agent games the endogenous correlated equilibrium concept reduces to the Nash equilibrium concept. Conditions (b) and (c) are traditionally assumed in game theory, but Skyrms (1991) and Vanderschraaf (1995) argue that there may be good reasons to relax these assumptions in games with 3 or more agents.

Brandenburger and Dekel (1988) show that in 2-agent games, if the beliefs of the agents are common knowledge, condition $(3.iii)$ characterizes a Nash equilibrium-in-beliefs. As they note, condition $(3.iii)$ characterizes a Nash equilibrium in beliefs for the n-agent case if the probability distributions are consistent and satisfy probabilistic independence. Proposition 3.6 extends Brandenburger and Dekel's result to the endogenous correlated equilibrium concept by relaxing the consistency and probabilistic independence assumptions.

Proposition 3.6. Assume that the probabilities

$$\mu = (\mu_1, \ldots, \mu_n) \in \Delta_1(S_{-1}) \times \cdots \times \Delta_n(S_{-n})$$

are common knowledge. Then common knowledge of Bayesian rationality is satisfied if, and only if, μ is an endogenous correlated equilibrium.

PROOF. Suppose first that common knowledge of Bayesian rationality is satisfied. Then, by Proposition 3.4, for a given agent $k \in N$, if $\mu_i(s_{kj}) > 0$ for each agent $i \neq k$, then s_{kj} must be optimal for k given some distribution $\sigma_k \in \Delta_k(S_{-k})$. Since the agents' distributions are common knowledge, this distribution is precisely μ_k, so $(3.iii)$ is satisfied for k. $(3.iii)$ is similarly established for each other agent $i \neq k$, so μ is an endogenous correlated equilibrium.

Now suppose that μ is an endogenous correlated equilibrium. Then, since the distributions are common knowledge, $(3.i)$ is common knowledge, so common knowledge of Bayesian rationality is satisfied by Proposition 3.4. □

Corollary 3.7. (Brandenburger and Dekel, 1988). Assume in a 2-agent game that the probabilities

$$\mu = (\mu_1, \mu_2) \in \Delta_1(S_{-1}) \times \Delta_2(S_{-2})$$

are common knowledge. Then common knowledge of Bayesian rationality is satisfied if, and only if, μ is a Nash equilibrium.

PROOF. The endogenous correlated equilibrium concept reduces to the Nash equilibrium concept in the 2-agent case, so the corollary follows by Proposition 3.6. □

If μ^* is a strict equilibrium, then one can predict which pure strategy profile the agents in a game will follow given common knowledge of the game, rationality and μ^*. But if μ^* is such that several distinct pure strategy profiles satisfy $(3.iv)$ with respect to μ^*, then one can no longer predict with certainty what the agents will do. For instance, in the Chicken game of Figure 3.1, the belief distributions defined by $\alpha_1 = \alpha_2 = \frac{1}{3}$ together are a Nash equilibrium-in-beliefs. Given common knowledge of this equilibrium, either pure strategy is a best reply for each agent, in the sense that either pure strategy maximizes expected utility. Indeed, if agents can also adopt randomized or *mixed* strategies at which they follow one of several pure strategies according to the outcome of a chance experiment, then any of the infinitely mixed strategies an agent might adopt in Chicken is a best reply given μ^*.[24] So the endogenous correlated equilibrium concept does not determine the exact outcome of a game in all cases, even if one assumes probabilistic consistency and independence so that the equilibrium is a Nash equilibrium.

An alternate correlated equilibrium concept formalized by Aumann (1974, 1987) does give a determinate prediction of what agents will do in a game given appropriate common knowledge. To illustrate Aumann's correlated equilibrium concept, let us consider the Figure 3.1 game once more. If Kay and Amie can tie their strategies to their knowledge of the possible worlds in a certain way, they can follow a system of correlated strategies which will yield a payoff vector they both prefer to that of the mixed Nash equilibrium and which is itself an equilibrium. One way they can achieve this is to have their friend Ron play a variation of the familiar shell game by hiding a pea under one of three walnut shells, numbered 1, 2 and 3. Kay and Amie both think that each of the three relevant possible worlds corresponding to $\omega_k = \{\text{the pea lies under shell } k\}$ is equally likely. Ron then gives Amie and Kay each a private recommendation, based upon the outcome of the game, which defines a system of strategy combinations f as follows

$$(\bigstar) \qquad f(\omega) = \begin{cases} (s_1, s_1) \text{ if } \omega_k = \omega_1 \,, \\ (s_1, s_2) \text{ if } \omega_k = \omega_2, \\ (s_2, s_1) \text{ if } \omega_k = \omega_3 \end{cases}$$

f is a *correlated* strategy system because the agents tie their strategies, by following their recommendations, to the same set of states of the world Ω. f is also a strict *Aumann correlated equilibrium*, for if each agent knows how Ron makes his recommendations, but knows only the recommendation he gives her, either would do strictly worse were she to deviate from her recommendation.[25] Since there are several strict equilibria of Chicken, f

---

[24]A mixed strategy $\sigma_k(\,\cdot\,)$ is a probability distribution defined over k's pure strategies by some random experiment such as the toss of a coin or the spin of a roulette wheel. k plays each pure strategy s_{kj} with probability $\sigma_k(s_{kj})$ according to the outcome of the experiment, which is assumed to be probabilistically independent of the others' experiments. A strategy is *completely mixed* when each pure strategy has a positive probability of being the one selected by the mixing device.

Nash (1950, 1951) originally developed the Nash equilibrium concept in terms of mixed strategies. In subsequent years, game theorists have realized that the Nash and more general correlated equilibrium concepts can be defined entirely in terms of the agents' beliefs, without recourse to mixed strategies. See Aumann (1987), Brandenburger and Dekel (1988), and Skyrms (1991) for an extended discussion of equilibrium-in-beliefs.

[25]Ron's private recommendations in effect partition Ω as follows: $\mathcal{H}_1 = \{\{\omega_1, \omega_2\}, \{\omega_3\}\}$ and $\mathcal{H}_2 = \{\{\omega_1, \omega_3\}, \{\omega_2\}\}$. These partitions are diagrammed below:

corresponds to a convention as defined in Vanderschraaf (1997). The overall expected payoff vector of f is $(3, 3)$, which lies outside the convex hull of the payoffs for the game's Nash equilibria and which Pareto-dominates the expected payoff vector $(\frac{4}{3}, \frac{4}{3})$ of the mixed Nash equilibrium defined by $\alpha_i = \frac{1}{3}$, $i = 1, 2$.[26] The correlated equilibrium f is characterized by the probability distribution of the agents' play over the strategy profiles, given in Figure 3.3.

Figure 3.3. Correlated Equilibrium Distribution for Chicken

		Kay s_1	Kay s_2
Amie	s_1	$\frac{1}{3}$	$\frac{1}{3}$
	s_2	$\frac{1}{3}$	0

Aumann (1987) proves a result relating his correlated equilibrium concept to common knowledge. To review this result, we must give the formal definition of Aumann correlated equilibrium.

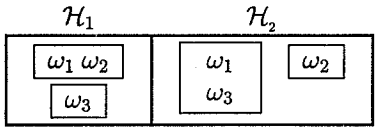
Definition 3.8. Given a game $\Gamma = (N, S, \boldsymbol{u})$ together with a finite set of possible worlds Ω, the vector valued function $f\colon \Omega \rightarrow S$ is a *correlated n-tuple*. If $f(\omega) = (f_1(\omega), \ldots, f_n(\omega))$ denotes the components of f for the agents of N, then agent k's *recommended strategy at* ω is $f_k(\omega)$. f is an *Aumann correlated equilibrium* iff,

(i) $$E(u_k \circ f) \geq E(u_k(f_{-k}, g_k))$$

for each $k \in N$ and for any function g_k that is a function of f_i. □

The agents are at Aumann correlated equilibrium if at each possible world $\omega \in \Omega$, no agent will want to deviate from his recommended strategy, given that the others follow their recommended strategies. Hence, Aumann correlated equilibrium uniquely specifies the strategy of each agent, by explicitly introducing a space of possible worlds to which agents can correlate their acts. The deviations g_i are required to be functions of f_i, that is, compositions of some other function with f_i, because i is informed of $f_i(\omega)$ only, and so can only distinguish

---



Given their private information, at each possible world ω to which an agent i assigns positive probability, following f maximizes i's expected utility. For instance, at $\omega = \omega_2$,

$$E(u_1(A_1) \mid \mathcal{H}_1)(\omega_2) = \tfrac{1}{2} \cdot 3 + \tfrac{1}{2} \cdot 2 = \tfrac{5}{2} > 2 = \tfrac{1}{2} \cdot 4 + \tfrac{1}{2} \cdot 0 = E(u_1(A_2) \mid \mathcal{H}_1)(\omega_2)$$

and

$$E(u_2(A_2) \mid \mathcal{H}_2)(\omega_2) = 4 > 3 = E(u_2(A_1) \mid \mathcal{H}_2)(\omega_2)\,.$$

[26] An outcome $\boldsymbol{s}_1$ of a game Pareto-dominates an outcome $\boldsymbol{s}_2$ if, and only if,

(i) $E(u_k(\boldsymbol{s}_1)) \geq E(u_k(\boldsymbol{s}_2))$ for all $k \in N$.

$\boldsymbol{s}_1$ strictly Pareto-dominates $\boldsymbol{s}_2$ if the inequalities of (i) are all strict.

between the possible worlds of Ω that are distinguished by f_i. As noted already, the primary difference between Aumann's notion of correlated equilibrium and the endogenous correlated equilibrium is that in Aumann's correlated equilibrium, the agents correlate their strategies to some event $\omega \in \Omega$ that is external to the game. One way to view this difference is that agents who correlate their strategies exogenously can calculate their expected utilities conditional on their own strategies.

In Aumann's model, a description of each possible world ω includes descriptions of the following: the game Γ, the agent's private information partitions, and the actions chosen by each agent at ω, and each agent's prior probability distribution $\mu_k(\cdot)$ over Ω. The basic idea is that conditional on ω, everyone knows everything that can be the object of uncertainty on the part of any agent, but in general, no agent necessarily knows which world ω is the actual world. The agents can use their priors to calculate the probabilities that the various act combinations $\boldsymbol{s} \in S$ are played. If the agents' priors are such that for all $i, j \in N$, $\mu_i(\omega) = 0$ iff $\mu_j(\omega) = 0$, then the agents' priors are *mutually absolutely continuous*. If the agents' priors all agree, that is, $\mu_1(\omega) = \cdots = \mu_n(\omega) = \mu(\omega)$ for each $\omega \in \Omega$, then it is said that the *common prior assumption*, or CPA, is satisfied. If agents are following an Aumann correlated equilibrium f and the CPA is satisfied, then f is an *objective* Aumann correlated equilibrium. An Aumann correlated equilibrium is a Nash equilibrium if the CPA is satisfied and the agents' distributions satisfy probabilistic independence.[27]

Let $s_i(\omega)$ denote the strategy chosen by agent i at possible world ω. Then $s\colon \Omega \to S$ defined by $s(\omega) = (s_1(\omega), \ldots, s_n(\omega))$ is a correlated n-tuple. Given that $\mathcal{H}_i$ is a partition of Ω[28], the function $s_i\colon \Omega \to S_i$ defined by s is *$\mathcal{H}_i$-measurable* if for each $H_{ij} \in \mathcal{H}_i$, $s_i(\omega')$ is constant for each $\omega' \in H_{ij}$. $\mathcal{H}_i$-measurability is a formal way of saying that i knows what she will do at each possible world, given her information.

Definition 3.9. Agent i is *Bayes rational with respect to* $\omega \in \Omega$, or *ω-Bayes rational*, iff s_i is $\mathcal{H}_i$-measurable and

$$(i) \qquad E(u_i \circ s \mid \mathcal{H}_i)(\omega) \geq E(u_i(v_i, s_{-i}) \mid \mathcal{H}_i)(\omega)$$

for any $\mathcal{H}_i$-measurable function $v_i\colon \Omega \to S_i$. □

Note that Aumann's definition of ω-Bayesian rationality implies that $\mu_i(\mathcal{H}_i(\omega)) > 0$, so that the conditional expectations are defined. Aumann's main result, given next, implicitly assumes that $\mu_i(\mathcal{H}_i(\omega)) > 0$ for every agent $i \in N$ and every possible world $\omega \in \Omega$. This poses no technical difficulties if the CPA is satisfied, or even if the priors are only mutually absolutely continuous, since if this is the case then one can simply drop any ω with zero prior from consideration.

Proposition 3.10 (Aumann 1987). If each agent $i \in N$ is ω-Bayes rational at each possible world $\omega \in \Omega$, then the agents are following an Aumann correlated equilibrium. If the CPA is satisfied, then the correlated equilibrium is objective.

PROOF. We must show that $s\colon \Omega \to S$ as defined by the $\mathcal{H}_i$-measurable s_i's of the Bayesian rational agents is an objective Aumann correlated equilibrium. Let $i \in N$ and $\omega \in \Omega$ be

---

[27]While both the endogenous and the Aumann correlated equilibrium concepts generalize the Nash equilibrium, neither correlated equilibrium concept contains the other. See Chapter 2 of Vanderschraaf (1995) for examples which show this.

[28]Aumann (1987) notes that it is possible to extend the definitions of Aumann correlated equilibrium and $\mathcal{H}_i$-measurability to allow for cases in which Ω is infinite and the $\mathcal{H}_i$'s are not necessarily partitions. However, he argues that there is nothing to be gained conceptually by doing so.

given, and let g_i: $\Omega \to S_i$ be any function that is a function of s_i. Since s_i is constant over each cell of $\mathcal{H}_i$, g_i must be as well, that is, g_i is $\mathcal{H}_i$-measurable. By Bayesian rationality,

$$E(u_i \circ s \mid \mathcal{H}_i)(\omega) \geq E(u_i(g_i, s_{-i}) \mid \mathcal{H}_i)(\omega)$$

Since ω was chosen arbitrarily, we can take iterated expectations to get

$$E(E(u_i \circ s \mid \mathcal{H}_i)(\omega)) \geq E(E(u_i(g_i, s_{-i}) \mid \mathcal{H}_i)(\omega))$$

which implies that

$$E(u_i \circ s) \geq E(u_i(g_i, s_{-i}))$$

so s is an Aumann correlated equilibrium. □

Part of the uncertainty the agents might have about their situation is whether or not all agents are rational. But if it is assumed that all agents are ω-Bayesian rational at each $\omega \in \Omega$, then a description of this fact forms part of the description of each possible ω and thus lies in the meet of the agents' partitions. As noted already, descriptions of the agents' priors, their partitions and the game also form part of the description of each possible world, so propositions corresponding to these facts also lie in the meet of the agents' partitions. So another way of stating Aumann's main result is as follows: *Common knowledge of ω-Bayesian rationality at each possible world implies that the agents follow an Aumann correlated equilibrium.*

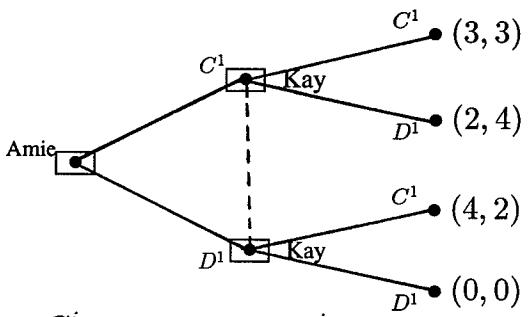
Propositions 3.6 and 3.10 are powerful results. They say that common knowledge of rationality and of agents beliefs about each other, quantified as their probability distributions over the strategy profiles they might follow, implies that the agents' beliefs characterize an equilibrium of the game. Then if the agents' beliefs are unconditional, Proposition 3.6 says that the agents are rational to follow a strategy profile consistent with the corresponding endogenous correlated equilibrium. If their beliefs are conditional on their private information partitions, then Proposition 3.10 says they are rational to follow the strategies the corresponding Aumann correlated equilibrium recommends. However, we must not overestimate the importance of these results, for they say nothing about the *origins* of the common knowledge of rationality and beliefs. For instance, in the Chicken game of Figure 3.1, we considered an example of a correlated equilibrium in which it was *assumed* that Amie and Kay had common knowledge of the system of recommended strategies defined by (★). Given this common knowledge, Kay and Amie indeed have decisive reason to follow the Aumann correlated equilibrium f. But where did this common knowledge come from? How, in general, do agents come to have the common knowledge which justifies their conforming to an equilibrium? Philosophers and social scientists have made only limited progress in addressing this question.

Games of Perfect Information

In extensive form games, the agents move in sequence. At each stage, the agent who is to move must base her decisions upon what she knows about the preceding moves. This part of the agent's knowledge is characterized by an *information set*, which is the set of alternative moves that an agent knows her predecessor might have chosen. For instance, in the extensive form game of Figure 3.4, when Kay moves she is at her information set $I^{22} = \{C^1, D^1\}$, that

is, she moves knowing that Amie might have chosen either C^1 or D^1, so this game is an extensive form representation of the Chicken game of Figure 3.1.[29]

Figure 3.4.



C^i = "cooperate", D^i = "defect"

In a game of perfect information, each information set consists of a single node in the game tree, since by definition at each state the agent who is to move knows exactly how her predecessors have moved. In Example 1.3 it was noted that the method of backwards induction can be applied to any game of perfect information.[30] The backwards induction solution is the unique Nash equilibrium of a game of perfect information. The following result gives sufficient conditions to justify backwards induction play in a game of perfect information:

Proposition 3.11 (Bicchieri 1993). In an extensive form game of perfect information, the agents follow the backwards induction solution if the following conditions are satisfied for each agent i at each information set I^{ik}:

(α) i is rational, i knows this and i knows the game, and

(β) At any information set I^{jk+1} that immediately follows I^{ik}, i knows at I^{ik} what j knows at I^{jk+1}.

PROOF. The proof is by induction on m, the number of potential moves in the game. If $m = 1$, then at I^{i1}, by (α) agent i chooses a strategy which yields i her maximum payoff, and this is the backwards induction solution for a game with one move.

Now suppose the proposition holds for games having at most $m = r$ potential moves. Let Γ be a game of perfect information with $r + 1$ potential moves, and suppose that (α) and (β) are satisfied at every node of Γ. Let I^{i1} be the information set corresponding to the root of the tree for Γ. At I^{i1}, i knows that (α) and (β) obtain for each of the subgames that start at the information sets which immediately follow I^{i1}. Then i knows that the outcome of play for each of these subgames is the backwards induction solution for that subgame. Hence, at I^{i1} i's payoff maximizing strategy is a branch of the tree starting from I^{i1} which leads to a subgame whose backwards induction solution is best for i, and since i is rational, i chooses such a branch at I^{i1}. But this is the backwards induction solution for the entire game Γ, so the proposition is proved for $m = r + 1$. □

Proposition 3.11 says that far less than common knowledge of the game and of rationality suffices for the backwards induction solution to obtain in a game of perfect

---

[29]Note that in the extensive form game trees given in the figures, the agents' information sets are depicted by boxes surrounding the relevant nodes.

[30]In general, the method of backwards induction is undefined for games of imperfect information, although backwards induction reasoning can be applied to a limited extent in such games.

information. All that is needed is for each agent at each of her information sets to be rational, to know the game and to know what the next agent to move knows! For instance, in the Figure 1.2 game, if R_1 (R_2) stands for "Alan (Fiona) is rational" and $\boldsymbol{K}_i(\Gamma)$ stands for "i knows the game Γ", then the backwards induction solution is implied by the following:

(i) At I^{24}, R_2 and $\boldsymbol{K}_2(\Gamma)$.

(ii) At I^{13}, R_1, $\boldsymbol{K}_1(\Gamma)$, $\boldsymbol{K}_1(R_2)$ and $\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$.

(iii) At I^{22}, $\boldsymbol{K}_2(R_1)$, $\boldsymbol{K}_2\boldsymbol{K}_1(R_2)$ and $\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$.

(iv) At I^{11}, $\boldsymbol{K}_1\boldsymbol{K}_2(R_1)$, $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(R_2)$ and $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$.[31]

One might think that a corollary to Proposition 3.11 is that in a game of perfect information, common knowledge of the game and of rationality implies the backwards induction solution. This is the *classical argument* for the backwards induction solution. Many game theorists continue to accept the classical argument, but in recent years, the argument has come under strong challenge, led by the work of Reny (1987, 1992), Binmore (1987) and Bicchieri (1989, 1993). The basic idea underlying their criticisms of backwards induction can be illustrated with the Figure 1.2 game. According to the classical argument, if Alan and Fiona have common knowledge of rationality and the game, then each will predict that the other will follow her end of the backwards induction solution, to which his end of the backwards induction solution is his unique best response. However, what if Fiona reconsiders what to do if she finds herself at the information set I^{22}? If the information set I^{22} is reached, then Alan has of course not followed the backwards induction solution. If we assume that at I^{22}, Fiona knows only what is stated in (iii), then she can explain her being at I^{22} as a failure of either $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(R_2)$ or $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$ at I^{11}. In this case, Fiona's thinking that either $\sim \boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(R_2)$ or $\sim \boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$ at I^{11} is compatible with what Alan in fact does know at I^{11}, so Fiona should not necessarily be surprised to find herself at I^{22}, and given that what she knows there is characterized by (iii), following the backwards induction solution is her best strategy. But if rationality and the game are common knowledge, or even if Fiona and Alan both have just have mutual knowledge of the statements characterized by (iii) and (iv), then at I^{22}, Fiona knows that $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(R_2)$ or $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma)$ at I^{11}. Hence given this much mutual knowledge, Fiona no longer can explain why Alan has deviated from the backwards induction solution, since this deviation contradicts part of what is their mutual knowledge. So if she finds herself at I^{22}, Fiona does not necessarily have good reason to think that Alan will follow the backwards induction solution of the subgame beginning at I^{22}, and hence she might not have good reason to follow the backwards induction solution, either. Bicchieri (1993), who along with Binmore (1987) and Reny (1987, 1992) extends this argument to games of perfect information with arbitrary length, draws a startling conclusion: If agents have strictly too few or *strictly too many* levels of mutual knowledge of rationality and the game relative to the number of potential moves, one cannot predict that they will

---

[31] By the elementary properties of the knowledge operator, $\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma) \subseteq \boldsymbol{K}_2\boldsymbol{K}_1(\Gamma)$ and $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_2(\Gamma) \subseteq \boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(\Gamma)$, so we needn't explicitly state that at I^{22}, $\boldsymbol{K}_2\boldsymbol{K}_1(\Gamma)$ and at I^{11}, $\boldsymbol{K}_1\boldsymbol{K}_2\boldsymbol{K}_1(\Gamma)$. By the same elementary properties, the knowledge assumptions at the latter two information sets imply that Fiona and Alan have third-order mutual knowledge of the game and second-order mutual knowledge of rationality. For instance, since $\boldsymbol{K}_2\boldsymbol{K}_1(\Gamma)$ is given at I^{22}, we have $\boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_1(\Gamma)$ because $\boldsymbol{K}_1(\Gamma) \subseteq \boldsymbol{K}_1\boldsymbol{K}_1(\Gamma)$ and so $\boldsymbol{K}_2\boldsymbol{K}_1(\Gamma) \subseteq \boldsymbol{K}_2\boldsymbol{K}_1\boldsymbol{K}_1(\Gamma)$. The other statements which characterize third order-mutual knowledge of the game and second order mutual knowledge of rationality are similarly derived.

follow the backwards induction solution. This would undermine the central role backwards induction has played in the analysis of extensive form games. For why should the number of levels of mutual knowledge the agents have depend upon the length of the game?

The classical argument for backwards induction implicitly assumes that at each stage of the game, the agents discount the preceding moves as strategically irrelevant. Defenders of the classical argument can argue that this assumption makes sense, since by definition at any agents' decision node, the previous moves that led to this node are now fixed. Critics of the classical argument question this assumption, contending that at when reasoning about how to move at any of his information sets, *including those not on the backwards induction equilibrium path*, part of what an agent must consider is what conditions might have led to his being at that information set. In other words, agents should incorporate reasoning about the reasoning of the previous movers, or *forward induction* reasoning, into their deliberations over how to move at a given information set. Binmore (1987) and Bicchieri (1993) contend that a backwards induction solution to a game should be consistent with the solution a corresponding forward induction argument recommends. As we have seen, given common knowledge of the game and of rationality, forward induction reasoning can lead the agents to an apparent contradiction: The classical argument for backwards induction is predicated on what agents predict they would do at nodes in the tree that are never reached. They make these predictions based upon their common knowledge of the game and of rationality. But forward induction reasoning seems to imply that if any off-equilibrium node had been reached, common knowledge of rationality and the game must have failed, so how could the agents have predicted what would happen at these nodes?

This section has barely scratched the surface of this controversy over common knowledge and backwards induction. The key unresolved issue is of course explaining what happens at the off-equilibrium information sets. To date, there is not a generally accepted theory of what agents having certain mutual or common knowledge will do at off-equilibrium nodes. However, we can at least repeat one generally accepted conclusion: In a game of perfect information, mutual knowledge of rationality and the game which falls far short of common knowledge can suffice to explain why agents follow the game's Nash equilibrium, the backwards induction solution. On the other hand, unlike other examples we have considered in which agents have mutual and even common knowledge without having to reason through levels of knowledge, backwards induction arguments in games of perfect information require that at each information set, the agent who would move were the information set to be reached must reason her way through at least as many levels of knowledge as there are remaining potential moves in the game.

Games of Incomplete Information

One can draw several morals from the e-mail game of Example 1.5. Rubinstein (1987) argues that his conclusion seems paradoxical for the same reason the backwards induction solution of Alan's and Fiona's perfect information game might seem paradoxical: Mathematical induction does not appear to be part of our "everyday" reasoning. This game also shows that in order for A to be a common truism for a set of agents, they ordinarily must perceive an event which implies A *simultaneously* in each others' presence. A third moral is that in some cases, it may make sense for the agents to employ some solution concept weaker than Nash or correlated equilibrium. In their analysis of the e-mail game, Monderer and Samet (1989) introduce the notions of *ex ante* and *ex post* ϵ*-equilibrium*. An *ex ante* equilibrium h is

s system of strategy profiles such that no agent i expects to gain more than ϵ-utiles if i deviates from h. An *ex post* equilibrium h' is s system of strategy profiles such that no agent i expects to gain more than ϵ-utiles by deviating from h' given i's private information. When $\epsilon = 0$, these concepts coincide, and h is a Nash equilibrium. Monderer and Samet show that, while the agents in the e-mail game can never achieve common knowledge of the world ω, if they have common p-belief of ω for sufficiently high p, then there is an *ex ante* equilibrium at which they follow (A, A) if $\omega = \omega_1$ and (B,B) if $\omega = \omega_2$. This equilibrium turns out not to be *ex post.* However, if the situation is changed so that there are no replies, then Diane and Greta could have at most first order mutual knowledge that $\omega = \omega_2$. Monderer and Samet show that in this situation, given sufficiently high common *p-belief* that $\omega = \omega_2$, there is an *ex post* equilibrium at which Greta and Diane choose (B, B) if $\omega = \omega_2$! So another way one might view this third moral of the e-mail game is that agents' prospects for coordination can sometimes improve dramatically if they rely on their common beliefs as well as their mutual knowledge.

REFERENCES

Lewis (1969) is the classic pioneering study of common knowledge and its potential applications to conventions and game theory. As Lewis acknowledges, parts of his work are foreshadowed in Hume (1740) and Schelling (1960).

Aumann (1976) gives the first mathematically rigorous formulation of common knowledge using set theory. Bacharach (1989) and Bicchieri (1993) adopt Schiffer's idea of augmenting sentential logic with knowledge operators, and develop a logic of common knowledge which includes soundness and completeness theorems.

Aumann (1995) gives a recent defense of the classical view of backwards induction in games of imperfect information. For criticisms of the classical view, see Binmore (1987), Reny (1992), Bicchieri (1989) and especially Bicchieri (1993). Brandenburger (1992) surveys the known results connecting mutual and common knowledge to solution concepts in game theory. For more in-depth survey articles on common knowledge and its applications to game theory, see Binmore and Brandenburger (1989), Geanakoplos (1994) and Dekel and Gul (1996). For her alternate account of common knowledge along with an account of conventions which opposes Lewis' account, see Gilbert (1989). Monderer and Samet (1989) remains one of the best resources for the study of common p-belief.

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

Aumann, Robert.: 1976, "Agreeing to Disagree", *Annals of Statistics* 4, 1236-9.

Aumann, Robert.: 1987, "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55, 1-18.

Aumann, R. 1995. "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior* 8: 6-19.

Bacharach, Michael. 1989. "Mutual Knowledge and Human Reason", mimeo.

Bernheim, B. Douglas. 1984. "Rationalizable Strategic Behavior." *Econometrica*, 52: 1007-1028.

Bicchieri, Cristina. 1989. "Self Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis*, 30: 69-85.

Bicchieri, Cristina. 1993. Rationality and Coordination. Cambridge: Cambridge University Press.

Binmore, Ken. 1987. "Modelling Rational Players I," *Economics and Philosophy*, 3: 179-241.

Binmore, Ken. 1992. *Fun and Games*. Lexington, Massachusetts: D. C. Heath..

Binmore, Ken and Brandenburger, Adam. 1988, "Common knowledge and Game theory" ST/ICERD Discussion Paper 88/167, London School of Economics.

Brandenburger, Adam. 1992. "Knowledge and Equilibrium in Games", *Journal of Economic Perspectives* 6: 83-101.

Brandenburger, Adam, and Dekel, Eddie. 1987, "Common knowledge with Probability 1", *Journal of Mathematical Economics* 16, 237-245.

Brandenburger Adam and Dekel, Eddie. 1988. "The Role of Common Knowledge Assumptions in Game Theory", in *The Economics of Missing Markets, Information and Games*, ed. Frank Hahn. Oxford: The Clarendon Press: 46-61.

Carnap, Rudolf. 1947, *Meaning and Necessity: A Study in Semantics and Modal Logic*, Chicago, University of Chicago Press.

Dekel, Eddie and Gul, Faruk. 1996. "Rationality and Knowledge in Game Theory", working paper, Northwestern and Princeton Universities.

Geanakoplos, John. 1994. "Common Knowledge", in *Handbook of Game Theory*, Volume 2, ed. Robert Aumann and Sergiu Hart. Elsevier Science B.V.: 1438-1496.

Gilbert, Margaret. 1989, *On Social Facts*, Princeton University Press, Princeton.

Harsanyi, J. 1967. "Games with incomplete information played by "Bayesian" players, I: The basic model." *Management Science* 14: 159-82.

Harsanyi, J. 1968*a*. "Games with incomplete information played by "Bayesian" players, II: Bayesian equilibrium points." *Management Science* 14: 320-324.

Harsanyi, J. 1968*b*. "Games with incomplete information played by "Bayesian" players, III: The basic probability distribution of the game." *Management Science* 14: 486-502.

Hintikka, Jaako. 1962. *Knowledge and Belief*. Ithaca, New York: Cornell University Press.

Hume, David. (1740, 1888) 1976, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. rev. 2nd. ed., ed. P. H. Nidditch. Clarendon Press, Oxford.

Lewis, C. I. 1943, "The Modes of Meaning", *Philosophy and Phenomenological Research*, 4, 236-250.

Lewis, David. 1969, *Convention: A Philosophical Study*, Harvard University Press, Cambridge, Massachusetts.

Littlewood, J. E. 1953. *Mathematical Miscellany*, ed. B. Bollobas.

Mckelvey, Richard and Page, Talbot, "Common knowledge, consensus and aggregate information", *Econometrica* 54: 109-127.

Milgrom, Paul. 1981. "An axiomatic characterization of common knowledge", *Econometrica* 49: 219-222.

Monderer, Dov and Samet, Dov. 1989, "Approximating Common Knowledge with Common Beliefs", *Games and Economic Behavior* 1, 170-190.

Nash, John. 1950, "Equilibrium points in n-person games." *Proceedings of the National Academy of Sciences of the United States* 36, 48-49.

Nash, John. 1951, "Non-Cooperative Games." *Annals of Mathematics* 54, 286-295.

Pearce, David. 1984. "Rationalizable Strategic Behavior and the Problem of Perfection." *Econometrica*, 52: 1029-1050.

Reny, Philip. 1987. "Rationality, Common Knowledge, and the Theory of Games", working paper, Department of Economics, University of Western Ontario.

Reny, Philip. 1992. "Rationality in Extensive Form Games," *Journal of Economic Perspectives*, 6: 103-118.

Rubinstein, Ariel. 1987. "A Game with "Almost Common Knowledge": An Example", in *Theoretical Economics*, D. P. 87/165. London School of Economics.

Schelling, Thomas. 1960, *The Strategy of Conflict*, Harvard University Press, Cambridge, Massachusetts.

Schiffer, Stephen. 1972, *Meaning*, Oxford University Press, Oxford.

Skyrms, Brian. 1984, *Pragmatics and Empiricism*, Yale University Press, New Haven.

Stinchcombe, Max. 1988. "Approximate Common Knowledge", mimeo, University of California, San Diego.

Sugden, Robert.: 1986, *The Economics of Rights, Cooperation and Welfare*, Basil Blackwell, New York.

Vanderschraaf, Peter. 1995, "Endogenous Correlated Equilibria in Noncooperative Games", *Theory and Decision* 38: 61-84.

Vanderschraaf, Peter. 1995. *A Study in Inductive Deliberation*, Ph.D. thesis, Department of Philosophy, University of California at Irvine.

Vanderschraaf, Peter. 1997, "Knowledge, Equilibrium and Convention", mimeo.

von Neumann, John and Morgenstern, Oskar.: 1944, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.