Uncommon Knowledge

Abstract. This dissertation collects four papers on common knowledge and one on introspection principles in epistemic game theory. The first two papers offer a sustained argument against the importance of common knowledge and belief in explaining social behavior. Chapters 3 and 4 study the role of common knowledge of tautologies in standard models in epistemic logic and game theory. The first considers the problem as it relates to Robert Aumann's Agreement Theorem; the second (joint work with Peter Fritz) studies it in models of awareness. The fifth paper corrects a claimed Agreement Theorem of Geanakoplos (1989), and exploits the corrected theorem to provide epistemic conditions for correlated equilibrium and Nash equilibrium.

Contents

Acknowledgements

This thesis is indebted above all to five people. Daniel Rothschild supervised the BPhil thesis on which Chapter 3 is based. More recently, his comments on Chapter 2 transformed that paper from a scattered collection of observations to what it is now. Timothy Williamson's comments on Chapters 1-3 improved each of those chapters tremendously. His written work was part of the inspiration for the project I pursue in Chapters 1 and 2, and also for the project of Chapter 5. Frank Arntzenius has read Chapters 3 and 5 numerous times. He has given detailed, incisive comments on every part of these papers, as well as on every part of Chapters 1 and 2. If it were not for Frank and Tim's advice and support, I would not be a philosopher today. Frank especially has fielded every manner of administrative and academic query at fantastic speed by email, Skype, and in person. I am sorry I could not give him the Dutch championship he so richly deserves. While officially my Doktorbrüder (or properly, Halbbrüder), Peter Fritz and Jeremy Goodman have in reality been surrogate fathers. The idea for Chapter 1 was inspired by a conversation with Jeremy on one of many memorable walks in Worcester gardens; Jeremy repeatedly crushed the first ideas for Chapter 2 on subway rides downtown from a seminar Bob Stalnaker taught at Columbia (and later, again, at a number of dinners in Oxford). Peter Fritz has always pointed (however gently) to the interest of neighborhood frames of the kind I use throughout the thesis. A conversation with Peter about a much lesser draft of Chapter 4 one afternoon led to the current joint version, which is much better for being joint.

Finally, I owe a great debt to Tiankai Liu for both moral and technical support, especially through the second year of the BPhil and the first year of the DPhil.

This thesis is dedicated to Emma Loftus, with love.

Introduction

This dissertation collects four papers on common knowledge and one on introspection principles in epistemic game theory.

The first two papers offer a sustained argument against the importance of common knowledge and belief in explaining social behavior. A group commonly knows a proposition just in case all members know it, all know that all know it, and so on. A group commonly believes a proposition just in case all believe it, all believe that all believe it and so on. The main argument for the prevalence of common knowledge and belief is an abductive one: common knowledge and belief provide the best explanation for a wide range of social behavior. Chapters 1 and 2 demonstrate that the argument is unsound: there are other, better explanations of behavior.

Chapter 1 considers three examples which have been taken to illustrate the importance of common knowledge and common belief. In each case, I propose an alternative explanation for the behavior we observe (and for the rationality of that behavior).

But does the presence of common knowledge or belief offer a simpler explanation of this behavior than my alternative? I provide an example in which people exhibit behavior which is supposed to be typical of that observed in the presence of common knowledge and belief, but where people cannot achieve common knowledge of the relevant propositions. I use this example to show that the explanation of behavior based on common knowledge or common belief is not as simple as my alternative.

The examples in Chapter 1 are not the only ones which have been thought to motivate the importance of common knowledge or belief. Another motivation comes from linguistics and philosophy of language. The intricacies of conversational phenomena seem to demand that participants not only have beliefs about others' beliefs, but beliefs about others' beliefs about their own beliefs, and so on. This thought—along with more sophisticated arguments—has suggested that the common ground, the information shared by participants in a conversation, has the iterative structure of common belief. In his influential work on the subject, Robert Stalnaker has developed a theory of the common ground according to which it has this structure.

Chapter 2 motivates an alternative to Stalnaker's approach by considering some *prima facie* odd consequences of the view that the common ground has the iterated structure of common belief. I then develop a "minimal" theory, and show how it accounts for a variety of linguistic data. According to my minimal theory, the common ground has the structure of mutual belief, where a proposition is mutually believed just in case all parties believe it. The minimal theory can provide explanations of speaker presupposition, update, and a variety of subtle sentences involving epistemic vocabulary. I argue that these explanations are at least as compelling as the explanations available to those who take the common ground to have the structure of common belief.

Chapters 3 and 4 turn to the role of common knowledge of tautologies in standard models in epistemic logic and game theory. In these models, axioms on belief and knowledge are typically imposed by making them tautologies—true at every state in every model. But in these models, it is also a tautology that every agent knows every tautology. It follows by an easy induction that the agents commonly know every tautology.

Chapter 3 (a version of which has been accepted by the *Review of Symbolic Logic*) studies this issue in the context of Robert Aumann's Agreement Theorem. I use pointed models—in which a single state in the model is designated as the actual world, and assumptions are imposed only at that state—to show how Aumann's result depends on certain assumptions about the agents' common knowledge. I provide examples in which all of Aumann's informal assumptions hold at the actual world, but in which the agents in the model still "agree to disagree".

Chapter 4, which is joint work with Peter Fritz, pursues a related project for models of awareness. Dekel, Lipman and Rustichini (1998) prove a series of triviality results for "state space models" of unawareness. We show that moving to pointed models avoids these triviality results, while preserving the simplicity of standard state space models.

Chapter 5 is a technical exercise. The main novel observation is that Geanakoplos's (1989) claimed Agreement Theorem is false. The correction of the result, and the extension to provide epistemic conditions for correlated equilibrium and Nash equilibrium is mathematically trivial (and also not particularly illuminating). The reader may wish to read this Chapter only through Section 5.3, before skipping to the final section. (The reader unfamiliar with epistemic game theory may find Proposition 5.5.2 of some interest, although this result no longer counts as "new", since it has now appeared in a number of places. Still, the observation that nothing in the proof of this proposition depends on introspection principles may be of some interest to philosophers.) But let me emphasize

again, for readers short of time, that the novel results of Sections 5.4 and 5.5 will not be particularly conceptually rewarding.

Each of the papers is written to stand alone. The reader of the whole dissertation will thus find a number of definitions repeated throughout, which will at least make reading the dissertation faster. The bibliography is at the end of the whole dissertation. I have included back-references in the bibliography, to serve as a partial index for where specific topics are discussed.

CHAPTER 1

Uncommon Knowledge

Abstract. A group commonly knows a proposition if all know it, all know that all know it, and so on (and similarly for common belief). An important argument for the existence of common knowledge and belief is an abductive one: these states provide the best explanation of human behavior. This paper argues that the abductive argument for common knowledge and belief fails.

1.1. Introduction

A group *mutually knows* (or: mutually knows[1]) a proposition if and only if all know it. A group mutually knows^n a proposition if and only if all know that the group mutually knows^{n-1} it. The group then *commonly knows* a proposition just in case, for all natural numbers n, the group mutually knows^n it. Common belief is defined by replacing "know" with "believe" in these definitions.

Common belief and common knowledge have been claimed to play a crucial role in explaining social behavior. But do we ever have these complex attitudes? Normal people certainly do not often think about the infinitely iterated states of common knowledge and common belief at a conscious, representational level. So proponents of common knowledge and belief must hold that an agent's conscious experience of belief is not a decisive factor in determining what he or she believes. For them, in deciding what an agent believes, it is at least as important to look to the beliefs our best theory of the agent's behavior attributes to him or her. So the question of whether we have common knowledge or common belief reduces to the question of whether our behavior is best explained by the hypothesis that we do.

Now in fact, models of behavior which assume a great deal of common knowledge are used widely in a range of disciplines. The formal theory of common knowledge and belief has enabled game theorists to provide transparent analyses of rational action in complex multi-agent situations. And today, these models are widely used in philosophy, linguistics, sociology, and, above all, in theoretical economics and computer science. Proponents of common knowledge in philosophy take these models to provide evidence about human psychology. For example, Dan Greco has recently claimed that if an epistemological

theory is inconsistent with the existence of common knowledge, it conflicts with our best theories from the social sciences, and should be rejected in part for that reason.[1]

But do models which assume a great deal of common knowledge really provide our best theory of human behavior? In this paper, I will argue that they do not. I consider two examples which are supposed to motivate the importance of common knowledge, and I propose much simpler, more familiar psychological patterns to explain the observed behavior in these situations. In pursuing this project I build on a rich body of work in recent economic theory, which has demonstrated that relaxing overly strong common knowledge assumptions can often lead to more accurate predictions of behavior. I then provide an example in which information in "plain view" nevertheless fails to be commonly known. I exploit the example to argue that explanations in terms of common knowledge and belief cannot be as powerful as my preferred explanation. I conclude that the abductive argument for common knowledge and belief fails.

Models which assume a great deal of common knowledge *can* provide insight into the situations they represent, and it is easy to understand why they are so prevalent. These simplified and idealized models do not accurately reflect the realities of human psychology. But in abstracting away from the ugly reality, the models help to reveal the structure of the situations they represent in a particularly simple and illuminating form. Even so, we must be careful not to take the simplifying assumptions too seriously. Greco's argument—that the prevalence of models which assume a great deal of common knowledge demonstrates that arguments against the KK principle are not sound—is a prime example of this fallacy.

The plan of the paper is as follows. Section 1.2 develops the abductive argument in favor of common knowledge in slightly more detail. Section 1.3 then considers the coordinated attack scenario. I first introduce my positive explanation of the rationality of coordinated attack, and then show why an abductive argument based on this scenario fails. Section 1.4 studies the electronic mail game, a much subtler case for both proponents and opponents of common knowledge. Section 1.5 then presents the crux of my argument. I provide an example in which agents exhibit the kind of behavior common knowledge is supposed to explain, but in which (as I argue), the relevant propositions cannot be common knowledge. Section 1.6 completes the main argument, with a survey of some of the models in which common knowledge is used. Section 1.7 concludes by bolstering the conclusion that common knowledge is not particularly important, with

[1]Greco 2014a, 2014b

considerations derived from Timothy Williamson's "antiluminosity argument". Common knowledge, to the extent that it exists, is luminous. But even granting that *knowledge* is luminous, I am unaware of any theory which explains the further, striking fact that if public announcements (for example) lead to common knowledge, then the occurrence of a public announcement is luminous.[2]

1.2. The Abductive Argument

The abductive argument for common knowledge relies on the claim that common knowledge provides the best explanation of behavior. But how should we understand this notion of "best"? I will not attempt here to give a general theory of when one explanation is better than another; the issue is typically best settled by considering concrete alternatives. But at the outset it will be helpful to distinguish a few senses in which one explanation might be simpler than another, and to identify two which will be important later on.

Robert Stalnaker makes some suggestive remarks about the sense in which common knowledge and belief might be "simpler" than other states:

> ...There is no problem with an infinite number of tacit, or implicit beliefs. For example, I believe (tacitly, at least) that no human being weighs more than x pounds, for each x greater than 2000. Instead of thinking of beliefs, and presuppositions in terms of sentences that express them, in the language of thought, stored in the belief or presupposition box of the mind, think of them negatively: it is the live options—the space of possibilities one allows for—that need to be represented. One's beliefs or presuppositions are the propositions that are true in all of those possibilities. What would put an unrealistic computational load on the mind would be a situation in which one believed, up to six iterations, that one believed that, but failed to believe the seventh iteration. That would require representing a possible situation (and a representation of the complex relational structure of possibilities compatible with possibilities that are compatible with possibilities

[2]Some authors (most notably Schiffer 1972) use "mutual knowledge" for what I call "common knowledge". This usage has lost out in the vast literature in economic theory, and I'll go with the current terminology, which reserves "common knowledge" for the infinitely iterated attitude. There are other defined notions of common knowledge, and I show how my argument generalizes to them in Appendix 1.A. Note finally that I'll often slip into writing "common knowledge" when I mean "common knowledge or common belief". When the distinction matters, it should be clear from context.

> compatible with...etc.) in which this mind-bogglingly subtle distinction is made. Infinitely iterated beliefs and presuppositions are much simpler for the subject to represent, and to understand. (2009: 402-3)

I think Stalnaker's concerns about "computational load" are not a promising way of making precise the way in which common knowledge or belief might be simpler than other states. It is hard to see what Stalnaker could mean when he claims that we must "represent" live possibilities in order for them to be compatible with our beliefs. He cannot mean that subjects must have actively drawn the distinctions between each of the possibilities compatible with what they believe. For every integer n greater than 100 trillion, it's compatible with what I believe that there are exactly n stars. These are live possibilities for me. But I have never "represented" many of these possibilities in any substantive sense.[3] So granting Stalnaker that the distinction between higher order beliefs and knowledge is "mind-bogglingly subtle", it's not clear why ruling in possibilities which differ with respect to this distinction, as opposed to ruling them out, imposes a greater computational load on the subject.

But Stalnaker's comments suggest a more promising thought, which is more closely related to the argument introduced in the Introduction. According to this view, the hypothesis that people have common knowledge is the best explanation of behavior because it is a simple one. Later, we will consider two senses in which the hypothesis that people have common knowledge might be simple. First, it might be that the explanation unifies a wide variety of behavior, providing a single explanation of a variety of phenomena. Second, it might be that the explanation simplifies our models of individual situations. In the next section, we'll see an example of the second kind of simplicity. The first kind will become central in Section 1.5.

1.3. Coordinated Attack

I now turn to the first example which is supposed to motivate the importance of common knowledge. The example of coordinated attack has been used to argue that people will coordinate only if they commonly know that they will. There are two ways to understand this claim. In the first, it is claimed that the explanation of the behavior of coordinating crucially relies on the assumption that the agents commonly know that

[3]The same point refutes another interpretation of Stalnaker's remarks, as based on the idea that representing universal generalizations is easier than representing partial ones. The simplest version of this thought is at odds with Stalnaker's invocation for us not to consider beliefs as sentences in the belief-box. Still, one might suppose that he holds that possibilities in which universal generalizations hold are easier to represent—keep live—than ones in which they do not. But again the example of the stars is a counterexample to this claim.

they coordinate. This is not a serious argument, and I do not think proponents of common knowledge would wish to endorse it. But considering an argument of this form will help us to introduce some important ideas. A second, more serious argument based on coordinated attack relies on a claim about simple formal models. This argument will be the subject of Section 1.3.2.

1.3.1. A Version of Coordinated Attack. Here, finally, is the scenario:

> Two divisions of an army, each commanded by a general, are camped on two hilltops overlooking a valley. In the valley awaits the enemy. It is clear that if both divisions attack the enemy the next day at dawn they will win the battle, while if only one division attacks it will be defeated. As a result, neither general will attack unless he believes the other will attack with him. Their only method of communicating in the night is by sending a messenger through the enemy camp.[4]

Will the generals attack? The usual line of reasoning says: No. If General Row sends General Column a message, Row won't believe that Column will attack unless Row sends a confirmation and Column receives it. But Column, in turn, won't believe that Row will attack unless Column believes that Row has received his message, and he will believe this, once again, only if Row sends a message confirming that he received the message. And so on. The argument is supposed to show that common knowledge is required for coordination in this scenario. Since the generals' message-passing mechanism prevents them from achieving common knowledge, they cannot coordinate.

The conclusion of this argument is absurd. Suppose the generals, repeatedly risking the life of their messenger, transmit seventeen messages in the course of the night, each saying that they will attack the next day. But still, they do not attack. What will they say when they return to the court and the King wonders why they did not defeat the enemy on that fateful day? The scene is something out of a comedy. "With only 17 messages, I just wasn't ready to attack."

But why is this outcome so absurd? One reply would be to hold that it is much easier to achieve common knowledge than the story makes it seem. Maybe seventeen messages are enough to generate the needed common knowledge or belief. Another reply is to give

[4]This is a much modified and redacted version of Fagin et al. 1995: 190-191. For earlier versions of the problem, cf. Akkoyunlu et al. 1975: 73-4; Gray 1978. In the standard version of the story of coordinated attack, much emphasis is placed on the timing of the attack, but this won't be important for our purposes in this section. The issues raised by timing in coordinated attack will resurface in the discussion of Sapling, but they arise in Sapling in a way which is more surprising than the more abstract cases involving times.

up on common knowledge, and to seek an alternative explanation for how the generals coordinate.

1.3.1.1. *An Argument for Common Knowledge?* I want to start by undermining the motivation for adopting the first line of response: that is, for preserving common knowledge. I will do so by observing that it is not necessary for (rational) coordination that the generals commonly know that they will attack.

To see this point, suppose the generals pass no messages, but an angel comes to each of them separately, and in each case truthfully says: "The other general will attack." (Suppose the angel has foreseen the future.) We may suppose that both generals now know that the other general will attack. But do they know that they both know this? The angel told them nothing about the other's mind, and the angel could even have told each of them falsely but credibly that the other general has lost his mind, and will attack without thinking. In this case, it would be rational to attack, even though there is no common knowledge. It suffices for one general's attack to be rational that he believes the other will attack. (It is best if these beliefs are true—for then the attack will be successful—but if the angel misled one of them, that general's action could still be explained by his (false) belief that the other would attack.)

So the generals may be rational in attacking even if they do not commonly know that they will attack. What about sufficiency: if the generals commonly know that they will attack, will they then be rational in coordinating? Of course, but not not because common knowledge has any particularly favored position in the explanation of action. The reason common knowledge is sufficient for coordination is that if the two commonly know that they will attack, each also knows that the other will attack, and, as we saw, this on its own was sufficient to explain why they attack. The point can be further illustrated by considering messages with different contents. In the story above, we may suppose that the generals are passing messages which say "I will attack tomorrow." According to the argument which "shows" that the generals won't attack, at the end of the story the proposition that they will both attack is not even believed, let alone known, or commonly known. But the status of a different proposition does alter through the course of the message passing. For if $2n + 1$ messages are passed, it plausibly becomes mutually known^n that the first message was received. And yet even *common* knowledge that this first message has been passed would not on its own be enough to explain the generals' attacking. We can dramatize this point once again by *deus ex machina*. Suppose the angels appear in a burst of light, midway between the generals in the heavens and

		Column A	Column B
Row	A	M, M	$0, -L$
	B	$-L, 0$	$0, 0$

FIGURE 1.3.1. Coordinated Attack

announce "We are telling you both that General Column has received the message." If one of the generals is uncertain whether common knowledge *that the message has been passed* is enough for the other to act, this public announcement will not help him out of his predicament.

This preliminary discussion already shows us two important points. First, attacking is a "stable" outcome of the situation, even if common knowledge is not achieved. In fact, as we'll see in a moment, that profile of strategies is a Nash equilibrium of the game the generals are playing. Second, what is needed to achieve this stable situation is just belief, not common belief. If rational generals fail to coordinate, this is because one of them fails to believe that the other will attack. As it is—if we accept the opening argument, at any rate—the generals don't trust one another's messages. When the one receives a message saying that the other will attack, he doesn't believe that the other will in fact attack.[5]

1.3.1.2. *How the Generals Coordinate.* Now we will turn back to the original example, and I will introduce my positive explanation for how the generals can reasonably come to coordinate. In introducing this explanation, it will be useful to think in terms of probabilities, and to assign real-valued utilities to the outcomes. So let's suppose that ϵ is the probability of the messenger being captured, M is the utility of coordinating, $-L$ is the utility of attacking on one's own, and 0 is the utility of doing nothing. The basic game—without signals—is depicted in normal form in Figure 1.3.1, where action A is attack, and B is "do nothing".

We can now give a simple explanation according to which, for any finite positive values of M and L, there is some finite number of messages after which the generals coordinate. We suppose that each general is uncertain what kind of person the other is. Each general assigns different probabilities to different "types" of the other, where a "type" is associated with a function from external factors to actions, in particular, from the

[5]This explanation also holds the key to some examples in Heal 1978. The framing of those examples makes higher-order beliefs appear important, and so, raises an error-possibility we don't usually consider. The framing undermines the agents' first-order knowledge of what the other agent will do: the agents make the mistake of thinking that higher-order beliefs are important (after all, the case makes them seem as if they are!), and this mistake leads their usually stable patterns of action to unravel.

number of messages received to the actions of attacking or not. Each general i believes to degree ρ_{ij} that if the other general has received j messages, the other general will attack. Our account will assume that both Generals are rational, in the sense that each maximizes expected utility relative to his beliefs. Now we consider the case of General Row. If the actual number of messages Row has *sent* is m, then the number he has received is $m-1$. So if m is great enough that: $(\rho_{Rm}(1-\epsilon)+\epsilon\rho_{R(m-1)})M > (1-[\rho_{Rm}(1-\epsilon)+\epsilon\rho_{R(m-1)}])L$, then General R (Row) will attack. If as m goes to infinity, ρ_{im} goes to 1, then for any finite positive values of M and L, there will be some number of messages after which the generals attack. As the absurdity of the example with 17 shows, for most of us, this value ρ_{im} in fact becomes close to 1 very quickly.

This simple model gives a natural explanation for why it seems so obvious that the generals should attack. The explanation is natural in part because the psychology of the model is so familiar. We often decide what to do on the basis of a somewhat automatic process. We size up a situation, arrive at an "uppermost thought" about what will happen, and then choose our action according to this uppermost thought about the outcome. In this case, we see the number of messages which has been passed, and 17 certainly seems large enough for the other player to attack. So we attack.

In the model just presented, the generals do not engage in any higher-order reasoning. Each only thinks at "depth 1" about the other's actions, and doesn't take account of the other general's beliefs, utilities or rationality at all. But such a model is often appropriate: in many circumstances, it doesn't occur to us to think about what others are thinking; we only think about what they will do.

But in other circumstances, we do engage in this kind of reasoning, so it is worth considering an amendment of the model to account for some form of higher-order reasoning. In addition to the "level-1" and "level-0" types already introduced informally, we add level-2 and higher types, who *do* take account of others' reasoning, via a distribution over the space of types of strictly lower levels. A level-0 type, as we saw, is a function from messages received to actions. A level-1 type treats everyone else as level-0 types. The generals above were level-1 types of this kind: their beliefs were described by a distribution over the space of level-0 types. A level-2 type, in turn, has a distribution over the union of the space of level-1 and level-0 types: that is, a distribution over a measurable space containing all functions from message counts to actions *and* the set of distributions over the set of functions from message counts to actions. And so on for types of higher and higher levels. Here in this new model, suppose—as seems eminently plausible—that

the ρ_{ij} of level-1 types go to 1 as j goes to infinity, that level-2 types only assign positive probability to such level-1 types, and level-3 types only assign positive probability to such level-2 types, and so on. Then once again for any positive finite M and L, there will be an m such that if m messages are passed, the generals coordinate. In fact, as I've said before, it's plausible that most people assign comparatively high probability to types who will act after, say, four or five messages. So once again the model can explain—in a familiar way—how agents will coordinate after finitely many messages have been passed.

This time, in the more complex model, there is room for higher-order reasoning. The level-3 types, for example, do think about how the other general might be thinking. A level-3 type even thinks about how the other general thinks that he himself (the first general) thinks. This makes the psychology of the model even more familiar than it was before. On a given occasion, we find ourselves perhaps thinking about what others think, or thinking about what others think we think, but not much more than this. Moreover, the depth of our own thinking may vary from case to case: in one instance we might be level-0 thinkers; in another level-1, and so on. Both our confidences and how deep our reasoning processes go are somewhat automatically determined in a given situation.[6]

The fact that these processes are automatic does not make using them "irrational". Obeying one's "uppermost thought" in this way is often the right thing to do. We automatically recognize certain relevant features of many situations to which we may not attend consciously (and if we did, we might misunderstand their importance). But even if these automatic processes weren't taking account of extra information in this way, there's a more obvious reason for the generals to trust them, which holds for us too. The generals win their battle if they coordinate. If they trust their capacity to mimic unbounded high-level processes with explicit reasoning, they might conclude that 17 is not enough to warrant attacking. So it is better in this case to be bounded and victorious, than boundless and inert.

1.3.2. The Argument from Simple Models. The first argument from coordinated attack was not a serious argument for the importance of common knowledge or belief. We considered it at such length because it gave us a simple, concrete setting to introduce some important ideas, and in particular a "level-k" model for describing people's behavior. But there is a more serious argument for the importance of common knowledge and belief based on this example, to which we now turn.

[6]For a nice recent study of the depth of reasoning, see Arad and Rubinstein 2012. The classic game-theoretic studies of how far agents reason are: Nagel 1995, Stahl and Wilson 1994; 1995.

The argument proceeds as follows. We begin by specifying a simple class of models which encode plausible assumptions about the above situation. It then turns out to be valid on this class of models that agents coordinate only if they commonly believe that they coordinate. So the simple models seem to show that coordination requires common belief.

Let's consider the argument in more detail. Fagin et al. 1995 prove a theorem for a different purpose which we can use to introduce the problem. They consider a class of Kripke models on which it's valid (true at all worlds in every model in the class) that A attacks if and only if B attacks. They also assume that it's valid on this class of models that A attacks only if she believes she does, and that B attacks only if he believes he does. These assumptions alone are sufficient to prove that A and B attack only if it's common knowledge that they do.[7]

The class of models Fagin et al. consider diverges from the story above in one striking way. At the outset of the story, each agent reasons in part by considering the possibility that he might attack when the other doesn't, leading to a catastrophic defeat. The agents' fear of this possibility is part of what makes the story compelling. But since there are no worlds in any of Fagin et al.'s models in which A attacks but B doesn't, there are also no worlds at which A or B consider it possible that A attacks but B doesn't. So there are no worlds in the models which represent the outcome which the agents fear.

But this unfaithful assumption is in fact unnecessary for the main result: we can replace our assumption that A attacks if and only if B attacks with the assumption that A attacks only if A believes B attacks (and vice-versa). Once again, we consider only those models where A attacks only if she believes she herself attacks (and similarly for B). The expanded class of models now includes some models with worlds where the agents experience catastrophic defeat; they falsely believe the other will attack, and so attack

[7]In this section, I'm going to be using facts about a logic interpreted over a set of Kripke models without formally introducing a language or a class of models. The notation is fairly standard, and those who are interested will be able to check my arguments against standard definitions. Proof: (1)$\models$ $Attack_A \Leftrightarrow Attack_B$ (assumption); (2) $\models$ $Attack_A \Rightarrow B_A(Attack_A)$ (assumption). Now suppose (3) $w \models Attack_A \wedge Attack_B$. Then, by (2), we have $w \models B_A Attack_A$. (1) gives us that the set of worlds at which A attacks is the set of worlds at which B attacks, so if A believes that A attacks, he believes also that B attacks (since those are the same proposition). So $w \models B_A(Attack_A \wedge Attack_B)$, and similarly if we replace B_B for B_A. But now it's clear that attacking is a "self-evident" event for all agents, and thus "public"; it's a standard fact in epistemic logic that public events of this kind are commonly believed (see, e.g. Monderer and Samet 1989; Fagin et al. 1995: Ch. 1. The theorem is often attributed to Lewis 1969 and Aumann 1976, but priority belongs to Friedell 1969). Note that Monderer and Samet also prove a more general theorem which applies to p-common belief: that each agent assigns probability p that each assigns probability p that...So if we interpret belief here as "probability p", the result still goes through. (Note also that no specification of times or runs is needed to get the counterintuitive conclusion.)

without the needed support. But it remains the case that, if the agents coordinate, it's common belief that they do.[8]

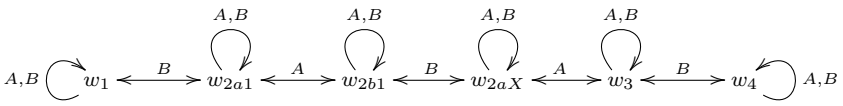
Reasoning about common knowledge and belief is hard. Epistemic models help us to ensure that we correctly assess the consequences of our assumptions about the world, and about agents' knowledge and belief. But the use of multimodal logics interpreted on Kripke models to represent multi-agent belief brings with it an additional risk. Anything which holds at all worlds in a model will be a matter of common belief among the agents in that model. So when faced with a surprising result such as what Fagin et al. dub the "paradox of common knowledge", we must check carefully that we cannot chalk up the "surprising" result to a background assumption of common belief.

Our result is in fact a textbook example of this phenomenon: the resulting common belief is a mere artifact of other modeling assumptions. First, observe that in our class of models, it's a tautology that each agent will act only if they believe the other will (if we had introduced the right sort of propositional language, this would be true at all worlds in all models). But in Kripke models, for any tautology, it's a tautology that each agent knows it. And this property carries over, by a simple inductive argument, to common belief: it's also a tautology that agents commonly believe that (each will act only if she believes the other will).

This common belief is part of what leads to our surprising result. If agents may fail to commonly believe that (each will act only if she believes the other will), then there may be worlds where the agents coordinate even though they do not commonly believe that they do.[9] In fact, this may be so even though agents mutually believe^n that

[8]By assumption we now have that $\models Attack_A \Rightarrow B_A Attack_{A,B}$. Similarly, that $\models Attack_B \Rightarrow B_B Attack_{A,B}$. So if $w \models Attack_{A,B}$, then $w \models B_{AB} Attack_{A,B}$; once again, attacking is a "self-evident" event; for the completion of the proof, see references in previous note.

[9]In the following example, at w_1, both players believe both players are attacking. At w_1, A knows the actual state is w_1, but B considers w_{2a1} possible. From w_{2a1}, A (not B) considers w_{2b1} possible; from w_{2b1} B (but not A) considers w_{2aX} possible. Finally, from w_{2aX}, A considers w_3 possible. At w_{2aX} and w_3, A attacks without believing B attacks (she considers it possible that B is not attacking, as at w_3). As a result the agents mutually believe^3 that they attack, but do not mutually believe^4 that they do, and so don't commonly believe that they do. This is possible because they don't mutually believe^2 (and so don't commonly believe) that (each will attack only if she believes the other will). The example could be made more realistic by including a sequence of worlds, $w_{2a2}, w_{2b2}, w_{2a3} ... w_{2bN}$, between w_{2b1} and w_{2aX}, where each w_{2an} (and each w_{2bn}) has the same relationship to its neighbors as w_{2a1} (respectively w_{2b1}) does. Adding N such worlds makes it so that, at w_1, the agents mutually believe^{N-1} that (each acts only if he or she believes the other does). They would also mutually believe^N that they are coordinating.

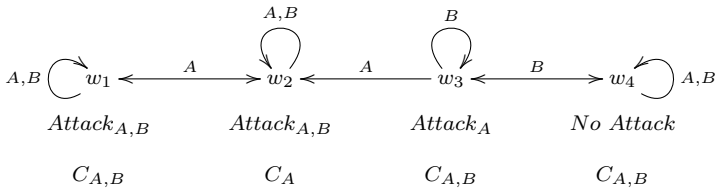


(each will act only if she believes the other will), *for any natural number n* (so long as they don't commonly believe it). The examples which demonstrate this fact do not require relaxing any introspection assumptions or assumptions about logical omniscience. They demonstrate that, in a slogan, it takes common belief to make common belief. The "paradox of common belief"—at least in this example—is not the product of an attractively simple model, but the unintended consequence of an overly simplistic one.

Other plausible weakenings also lead to the same result. For example, if we weaken the assumption that agents' beliefs are closed under entailment, we can give countermodels to the claim that agents coordinate only if they commonly believe that they do, even if agents commonly believe that (each attacks only if she believes the other does).[10] We can interpret these formal models as describing situations in which the agents *are* closed under entailments; in fact they may mutually believe^n for any natural number n that their beliefs are closed in this way. All that is required is that the agents do not commonly believe that each of them believes everything entailed by what he or she believes.

The argument based on coordinated attack began with the observation that simple models deliver the result that the agents coordinate only if they commonly believe that they do. It then holds that we should accept the theorems of such simple models at face value. This is an abductive argument based on the second kind of simplicity introduced in Section 1.2. But the proposed argument cannot be that every theorem of simple models should be accepted as describing reality. In simple models used in classical physics, for example, objects are treated as point masses. It's a theorem that in these models objects have no spatiotemporal extension. But we don't take the success of these models in

[10]To relax closure, we use neighborhood models, where each world is associated with a subset of 2^Ω, and at each world an agent believes only those propositions that are elements of this "neighborhood". We suppose that it's common knowledge that A's beliefs are closed under entailments so that A's beliefs can be represented as usual in a Kripke frame. (Kripke frames are special neighborhood models, where the neighborhood is a filter on the algebra of propositions, that is, the power set algebra of the underlying state space.) At every world except for w_2, B's beliefs are similarly closed under entailments; his neighborhood is a filter on the space of propositions. (We represent this in the depiction below by, e.g. "C_B" to indicate that B's beliefs there are closed.) But at w_2, B's neighborhood has only the following sets: $\{w_2\}\{w_1, w_2, w_3, w_4\}$ (only the first is represented in the figure). At w_2, B believes all tautologies of the model, including $Attack_A \Rightarrow B_A Attack_A$; B also believes that A attacks. But he fails to draw the conclusion that A believes that A attacks. Notice that if w_1 is the actual world, then B is in fact logically omniscient; w_2 is required only to represent the fact that A considers it possible that B may fail to draw this conclusion. Using a tactic related to the one in the previous note, we can easily provide a model where the two have arbitrarily high iterations of mutual belief that they're logically omniscient. Finally, note that in the model it's common belief that (each attacks only if she believes she does).



explaining the phenomena to demonstrate that real objects don't have spatiotemporal extension. This simplifying assumption was just something we built in at the start.

The two models just given show that the common belief present in coordinated attack is a consequence of similar, deliberately idealized models. By building in common knowledge that (each acts only if she believes the other will), along with common knowledge of logical competence, the models assume at the start that coordination requires common knowledge of these aspects of the situation. These simplifying assumptions about common knowledge then breed other, downstream examples of common knowledge. But as with spatiotemporal extension in models of Newtonian physics, these downstream instances of common knowledge are simply components of our idealization, not interesting facts which the models reveal about our reality.

1.4. Electronic Mail Game

One of our first observations about coordinated attack was that it *could* be rational for the agents to coordinate on attacking in that scenario. In particular, we noted that when utilities are added to the game, the coordination outcome where both generals attack is a Nash equilibrium of the game. Since it's a Nash equilibrium, it's also composed of strategies which survive iterated deletion of strictly dominated strategies. So in the first instance we could think of the problem of coordinated attack as a problem about equilibrium selection, not about whether we should play an equilibrium at all. The argument for the purported importance of common knowledge aimed to show that the generals would not achieve the best equilibrium, given the information they were able to share. But in that case the intuitive, better outcome was also an equilibrium profile.

A more powerful argument for the importance of common knowledge begins from a situation in which the intuitive outcome is *not* an equilibrium. In Ariel Rubinstein's electronic mail game (1989), in the absence of common knowledge, the *only* Nash equilibrium is one where players fail to coordinate on the better coordination point. In fact, variations on Rubinstein's game can be given in which, if common knowledge is absent, the only strategies which survive iterated deletion of strictly dominated strategies are the ones in which players fail to coordinate on the best outcome (see n. 11 below). In this situation, intuition still says that the players should coordinate on the better outcome. But all the usual game theoretic notions of rationality say that it is *irrational* for agents to play these actions. Whereas in the earlier example, the mathematical notions were just more permissive than intuition, here the mathematics says that intuition is wrong.

		Column A	Column B
Row	A	M, M	$0, -L$
	B	$-L, 0$	$0, 0$

G_A

		Column A	Column B
Row	A	$0, 0$	$0, -L$
	B	$-L, 0$	M, M

G_B

FIGURE 1.4.1. The Electronic Mail Game

Two players, Row and Column, are uncertain which of two coordination games, G_A or G_B, they are playing (see Figure 1.4.1); in each of the games, the players receive a payoff of M if they coordinate on an optimal action and 0 if they coordinate on the non-optimal action. The two games are in one sense inverses: the optimal pair of acts (A, A) in Game G_A (payoff of M to each) is the *non*-optimal coordination outcome (payoff of 0 to each) in game G_B; conversely, the optimal play (B, B) in Game G_B is the non-optimal coordination point in game G_A. If players fail to coordinate, however, both games favor A as the safer option: in both games, if Row plays B when Column plays A, Row receives $-L$, while Column receives 0; similarly if Column plays B while Row plays A, Column receives the losing $-L$, and Row receives 0. The payoffs obey the inequality $L > M > 0$. Game G_A occurs with probability $p > 1/2$; G_B occurs the rest of the time.

Row will be informed of the true game. If the game is G_B, his computer will send a message to Column's computer. If either player's computer receives a message, it automatically replies. But each message has a positive, equal and independent rate of failure (ϵ); the process ends almost surely. When it is over, each Player's monitor will display the number of messages he or she has received and sent. As Rubinstein showed, there is a single Nash equilibrium in which Row plays A in G_A, and in this equilibrium, both players play A regardless of how many messages they have received.[11]

There are three data which a theory of this example must explain. First, when the game is G_B and we receive a high number of messages, playing B seems intuitively rational. Second, people *do* coordinate when they play the electronic mail game.[12] But,

[11]We can now state the version of the game in which the only rationalizable profile is the intuitively wrong one (it is borrowed from Strzalecki 2010). Here, $p = 1/2$, and the payoffs are as follows:

		Column A	Column B
Row	A	$-2, -2$	$-2, 0$
	B	$0, -2$	$0, 0$

G_A

		Column A	Column B
Row	A	$1, 1$	$-2, 0$
	B	$0, -2$	$0, 0$

G_B

[12]Rubinstein himself predicted this (1989: 389). See now Camerer, 2003, pp. 226–232. Comparable studies (e.g. Heinemann et al 2004) also indicate that the absence of public information is something of an impediment to coordination, but not as much as theory predicts.

third, they do not coordinate as often in playing this game as they do in situations where they don't have a noisy messaging system.

In coordinated attack, the common knowledge-based analysis did not accord with the intuitive verdict about what is rational, nor did it accord with what we expect people to do. It failed on every count. But here it appears to have two points in its favor. First, while the common knowledge-based explanation still fails to agree with our intuitions about what it's rational to do, this counterintuitive verdict is now supported by the fact that the only Nash equilibrium is to play in accord with what the players commonly know. Second, the common knowledge-based explanation promises to explain why people coordinate more often in "open" situations, since they are supposed to have common knowledge in those situations, by contrast to what happens in the electronic mail game.

But we should hope for a theory which can also explain what people do—and ideally one which shows us that our intuitions about rationality are not as hopeless as the common knowledge-based theory claims. I will provide a theory which explains all three of the data: how it is rational to coordinate, why people do, and why they coordinate less often in the electronic mail game than in other situations. In doing so, I'll argue that Nash (and even rationalizability) don't capture the only interesting notion of rationality in the area; the reasoning I propose is rational in a new way, based on an extension of the ideas behind rationalizability. And—it will be no news to the reader—this more satisfactory theory will not employ common knowledge.

I will introduce the theory in two stages. First, I will discuss ways in which the electronic mail game is a very special situation, which does not often arise in ordinary life. Most ordinary situations which involve something like message passing are ones where the rational action is to coordinate after finitely many messages. Second, I will describe the reasoning of players in the electronic mail game, just as I did for coordinated attack. This second stage will in a way rely on the first: the rationality of players' behavior in these circumstances depends to some degree on what are good policies in other, similar situations. This second stage will also include an explanation for why successful coordination on the better outcome is not as frequent in the electronic mail game as in situations where players can communicate "openly".

1.4.1. Why the Electronic Mail Game is Special. The electronic mail game relies on a very special structure to ensure that common knowledge is required for coordination on the best outcome. Two of these special features concern the structure of the payoffs. First, the game requires that L (the loss agents will suffer) be greater than M (the upside to coordinating). In real life the payoffs to coordinating are typically greater than the losses of failing to coordinate. For example, if I want to meet Frank tomorrow at 2, the payoff to failed coordination is not so bad; perhaps I must send him an additional email or wait 15 minutes before realizing he won't come. But, second, even those real-life situations where the appropriate inequality is observed typically differ from the formal example in that L will not be unevenly distributed between the participants. If Frank misses our meeting, then he'll suffer the costs of my unhappiness and generally be forced to compromise in the next round, even if I'm the one who bears the upfront costs of waiting impatiently for him to arrive.

These facts about payoffs are already enough to explain why in ordinary life, we can (rationally) coordinate using texts or email, only after sending a single message. A single message can be credible in normal life, in a way it was not for the generals in coordinated attack or the players in electronic mail.

But Rubinstein's Remark 1 (1989: 388) demonstrates an even more striking way in which the formal example differs from real life. The Remark states that the expectation of limited communication makes coordination possible, even in the absence of substantial common knowledge, and even when the payoffs observe the delicate pattern just described. Rubinstein shows that if ϵ is sufficiently small in the sense that $(1-\epsilon)M > \epsilon L$, then if it is commonly believed that only T messages will be sent, there *is* a Nash equilibrium in which both players play B if they receive confirmation of all their messages. In fact, although Rubinstein states his remark using common belief, the proof only requires *mutual* belief (here: certainty): what needs to be ruled out is the possibility that waiting for another message could be a conditional strategy with a higher expected payoff. Mutual certainty is enough for this, since each player is certain that no such message will come.

To see how this works, suppose Row and Column mutually believe that Column will reply to one message but no more, and mutually believe that Row will send only one message. Then when payoffs are as above, it's a Nash equilibrium for both players to play

B if they receive one message (and to play A otherwise).[13] In this situation, since Column doesn't know whether Row received her confirmation message, they coordinate on B (if they do) without having common belief of whether they're playing G_A or G_B, or common belief that they will play B—in fact they don't even have mutual knowledge[2] of what action they will take.[14] As is now well known, Nash does not require common knowledge even of the other player's strategies—first-order knowledge is enough. So we don't need to require that players commonly know that they are using these conditional strategies, either. If it is mutually known that only finitely many messages are sent, coordinating on the best outcome requires little substantive common knowledge.[15]

Still, even in the modified scenario there *is* a Nash equilibrium (albeit one with a lower expectation for each player) in which the players fail to coordinate on the better outcome in G_B. So one might wonder whether there's a way to understand the coordination we observe in normal life as the *uniquely* rational outcome of the game (or some close modification of it).

In an elegant paper, Ken Binmore and Larry Samuelson (2001) show that there is a way to understand coordination as the uniquely rational outcome. They expand on Rubinstein's Remark, studying versions of the electronic mail game in which message sending is not automatic, but voluntary, and also where attention to messages is costly. If either of these assumptions hold, then as in the example above, there are Nash equilibria where agents coordinate after only finitely many messages. But more importantly, Binmore and Samuelson show that in models where both assumptions hold (message-sending is voluntary and attending to messages is costly) the equilibria in which players play B

[13]Justification: When Column has received her message (and automatically replied), her expected payoff from B is $(1-\epsilon)M - \epsilon L$, which is strictly greater than 0, her expected payoff from A. So Column always plays B when she's received one message, presuming Row will do the same. If Row receives the reply, for his part, he's guaranteed M by playing B, as opposed to 0 from playing A.

[14]Justification: Column isn't certain that Row has received her message; she believes that with probability ϵ Row hasn't received it. So they don't even mutually believe Row will play B. Moreover, Column knows that if Row hasn't received the message back, Row will believe with probability ϵ that Column hasn't received the message, so they do mutually believe that the game is G_B, but they don't mutually believe that they do.

[15]The classic paper is Aumann and Brandenburger 1995, but cf the reply by Polak 1999. Subsequent work has shown that Aumann and Brandenburger's result can be extended significantly to avoid common belief of the kind alleged by Polak (e.g. Barelli 2009, Liu 2010 (unpublished), Bach and Tsakas 2014). An extremely simple form of the theorem can now be found in Dekel and Siniscalchi 2014, and also in Battigalli et al. 2014. It is also Proposition 5.2, in Chapter 5 of this thesis.

after finitely many messages are "evolutionarily stable for sufficiently small costs".[16] Ignoring all messages as in the unique Nash equilibrium of Rubinstein's game, by contrast, is not stable in this way. So Binmore and Samuelson demonstrate that, in an important sense, coordinating is the uniquely rational action in close variants on the original game.

1.4.2. Explaining Electronic Mail, and Why Coordination Rates Differ. We have seen that the electronic mail game is very special: it depends on a special pattern of payoffs, on the automatic character of the message-sending machines, and on the assumption that message-sending and attention are costless. If we relax these assumptions there are Nash equilibria—and, in the final case, even Nash equilibria with a further, special characteristic—in which players coordinate on the best outcome. For most of our lives, it can be rational to coordinate on the better equilibrium in analogous situations, even in the absence of common knowledge.

But this still leaves us with two questions. Since people do in fact coordinate even in the electronic mail game, we want to understand how they reason in these situations, ideally, in a way which explains how their behavior is rational. Moreover, we want a theory which also explains the difference between the electronic mail game and more usual situations in which information is shared without this messaging system.

The explanation I will propose will be familiar already.[17] The simplest model is exactly the one I used for coordinated attack. Different agents assign different probabilities to different level-0 "types" of other players, where a "type" is again a function from the number of messages to actions. Once again, we suppose that people assign varying probabilities to types who act after m messages have been received. To make this slightly more precise, we suppose that each person i has a belief that ρ_{im} of the population are such that they'll play B if they receive at least m messages. Thus, once again, for the case of

[16] In their paper, Binmore and Samuelson defend the importance of this notion, as opposed to the usual notion of evolutionary stability. The problem is that every Nash equilibrium of their revised game is strict, and hence evolutionarily stable in the standard sense. As they point out, the standard definition of evolutionary stability is justified on the basis of the assumption that the "basin" in which a strict equilibrium lies is guaranteed to be sufficiently steep that chance occurrences which would lead away from the equilibrium will be sufficiently rare as to be safely ignored. But in fact, this need not be so; even strict equilibria may lie in very "shallow" basins. The notion of evolutionarily stable for small costs avoids this problem by ensuring that small shocks cannot lead away from the equilibrium. Formally, a strategy is evolutionarily stable for sufficiently small costs iff for every sufficiently small ϵ, there are values $c^*$and d^* so that for all $c < c^*$ and $d < d^*$ and all strategies s' not equal to s,
$P(s, (1-\epsilon)s + \epsilon s', c, d) > P(s', (1-\epsilon) + \epsilon s', c, d)$,
where $P(s, s', c, d)$ is the payoff for a strategy s, playing against s', and c and d are costs to sending messages and to attention, respectively.

[17] In this case, we are fortunate that a full analysis of this kind of model has already been given: see Strzalecki 2010. In my model, the level 0 types differ from those considered in Strzalecki's basic model, in that I have heterogeneous versions of the level 0 types. But under the assumption that higher-level types assign probabilities in the way discussed for coordinated attack, his results carry over straightforwardly to the new setting.

Row, if m is great enough that $(\rho_{Rm}(1-\epsilon)+\epsilon\rho_{R(m-1)})M > (1-[\rho_{Rm}(1-\epsilon)+\epsilon\rho_{R(m-1)}])L$, then Row will play B.

Let's suppose that, for both Row and Column, $\rho_{i1} = 1$. Then if two messages are passed, they're guaranteed to coordinate. Now as I've said, these strategies do not form part of any Nash equilibrium, and in the variant in note 11, they won't survive iterated deletion of strictly dominated strategies. But, I want now to argue that these formal senses of rationality are too restrictive: the actions just described *are* rational. Since every strategy that is part of a Nash profile survives iterated deletion of strictly dominated strategies, I'll focus on arguing that agents can be rational in playing strategies which do not survive iterated deletion of strictly dominated strategies. If my argument succeeds, it will also show that agents can be rational in playing strategies which do not form part of any Nash equilibrium.[18]

To understand the argument, we need to consider the "epistemic" justification of the play of strategies which survive iterated deletion of strictly dominated strategies. It's a standard fact that these strategies are all and only the ones which are "rationalizable", that is, consistent with common knowledge of rationality (Bernheim 1984 and Pearce 1984 Brandenburger and Dekel 1987). The formal sense of rationality employed in this result involves two components: first, that agents maximize subjective expected utility; second, that each player has an infinite hierarchy of beliefs-about-beliefs-about-beliefs and so on. But what if one player maximizes subjective expected utility, but has a limited capacity for forming beliefs-about-beliefs? It seems clear that in such a circumstance an action might be *rationalized* by this player's limited beliefs, even if it is not rationalizable by any infinite hierarchy of beliefs. In fact, new actions could be rationalized even if neither player is in fact limited, but each player *thinks* the other has a limited capacity for forming beliefs. (This is in fact the situation in our example above.) So even if a strategy doesn't survive iterated deletion of strictly dominated strategies, it could be rational to play that strategy if you are limited, or if you believe the other player might be. Usually, as is familiar from Savage, rationality is assessed not by asking whether the beliefs a player has are appropriate, but by asking whether what the player did was rational *relative* to his beliefs. The formal theory of rationalizability diverges from this paradigm, since it doesn't allow players to assign positive probability to the hypothesis that others are limited. But this also makes the formal sense inappropriate in real life circumstances. In

[18]In any case, it's fairly well known the fact that an action is rationalizable is a more compelling *necessary* condition for an action to be rational than the fact that it belongs to a Nash equilibrium. A beautiful example in which a Nash strategy does not appear to be rational is given by Aumann 1985, figure 1.

the real world, it often seems right to assign the hypothesis that others are limited at least some positive probability.

So the fact that this eminently *rational* way of playing the game is not rationalizable reflects a limitation of the formal notion of rationalizability. Moreover, the behavior just described is rational in more substantive ways than just that it maximizes subjective expected utility relative to a specific hierarchy of beliefs. Remember how special the electronic mail game is. In most message-passing situations, coordinating on the better equilibrium is the rational action in the sense of being Nash, and thus rationalizable. Real life is, at least in this way, kind. In real life, reasoning boundedly usually gets us the same payoffs as reasoning unboundedly would. So in real life it makes sense to have simple heuristics which tend to lead to the good Nash outcome, without forcing us to think in detail about what's Nash and what's not.

Our experience tells us that when people pass messages, they coordinate on good outcomes. So when two people come to play the electronic mail game, each will assign high probability to the other coordinating on the good outcome. And it is perfectly reasonable to do so: we know that we have all learned our way around the kind world where bounded reasoning often gets us the best outcome. So good belief-forming policies lead to the belief that other players, too will reason boundedly—not because the other is irrational, but because reasoning boundedly is the right thing to do in situations such as this one.

But then why do people coordinate any less often in this situation, as compared with "normal" public announcement situations? The answer is not far to seek. We far more often encounter situations akin to a public announcement scenario. In such situations, our dispositions are stable, and our models of others are well-formed. The closest analog to the electronic mail game in real life is a scenario of normal communication. So we tend to bring our model of people's behavior "across" from these more familiar situations to deal with the new electronic mail scenario.[19] But we recognize that there's something fishy in the new situation, and some of us, or each of us with some probability, will decide on a different analogy, bringing a different model to the new situation. These mistakes are enough to explain the differences between public announcement scenarios and message-passing versions of the game. But since we display considerable regularity in how we form models of the new situation, we still coordinate at very high rates.

[19]For more precise work along these lines, see Mengel 2012 and Grimm and Mengel 2012.

The electronic mail game seemed to promise a new argument for the importance of common knowledge and common belief. But observed behavior in this situation suggests that the mathematics of higher-order belief in the electronic mail game does not track the psychological reality. If we try to respect this psychological reality, we see that common knowledge is a distraction. A simpler theory explains in one fell swoop why we coordinate so often, why it can be rational to do so, but also why the new game is hard, and sometimes gives us trouble.

1.5. Uncommon Knowledge

So far, we've seen that some important motivating examples for common knowledge are not compelling. I've provided explanations of these situations which I believe are more natural and more faithful to ordinary psychology than the explanations based on common knowledge. But there's a vast array of other examples which might be invoked in defense of the claim that common knowledge is crucial to explaining our social lives. We need some more decisive way of showing that theories which explain social behavior by invoking common knowledge are not as powerful and not as simple as the kind of theory I've been proposing.

In this section I will provide an example which exhibits all of the characteristics that common knowledge is usually invoked to explain. Some information is "public" or "open" in an important sense: it is in plain view for all to see. But the information is not common knowledge.

A simple, powerful theory of behavior will explain all examples of "public" information by the same mechanisms. This class of phenomena is unified, and deserves a uniform explanation. My example shows that the explanation in terms of common knowledge cannot deliver on this demand. I will argue that the alternative "level-k" explanations can deliver on the demand. So the abductive argument for common knowledge is doomed.

1.5.1. Sapling.

> Sapling: You are playing in a game show. Two contestants compete at a time, together. Each contestant has a single button in front of them. The contestants have an unobstructed view of each other's faces (nothing below the neck), and of a sapling in the middle of the stage. They cannot see each other's keyboard; if they speak to each other, they are disqualified. Prior to the show, all players are taught important facts about inches and feet. If the sapling is taller than 6 inches, and

> both players press the button, they receive $\$1,000,000$. If the sapling is not taller than 6 inches, or if only one person presses the button, that person pays the show $\$10,000$.

Suppose that the sapling is in fact $5'10$ (that is, five foot ten inches tall—the single $'$ comes after the feet; when the double is used, it indicates inches). It's clear that you're in luck: you'll win $\$1,000,000$. But, as I'll now argue, the most prominent theories of our knowledge of tree-heights—even those which preserve the KK or "positive introspection" principle—predict that it's *not* common knowledge that the tree is taller than 6 inches. (Recall that KK is the principle that if a subject knows a proposition, the subject knows that she knows it.)

To make things concrete—and to show that my argument doesn't depend on denying positive introspection—I'll give the argument using Robert Stalnaker's view of the situation (Stalnaker, 2009). For Stalnaker, if the height of the tree is in fact within a particular plus / minus range of one's precise "best guess", then one knows that it is, regardless of where the height of the tree falls in that range.[20]

Suppose that both you and your partner have a range of ± 6 inches (we can even suppose that this was publicly announced before the game, after empirical tests). Moreover suppose that your best guess is right on the mark, at $5'10$, so that you know that the tree is between $5'4$ and $6'4$. What do you know about what your partner knows? That depends on your best guess, and on the general accuracy of your best guess about his best guess. Suppose, as is plausible, that there are *some* independent factors contributing to differences in his perception of the tree, but that many sources of potential error are correlated between you, so that your best guess provides evidence about your partner's best guess beyond the evidence given by the height of the tree. In other words, your accuracy about your partner's best guess is much greater than your accuracy about the height of the tree: your range concerning his best guess lies within ± 1 inch around your own best guess. So in the case above, you believe that his best guess is between $5'9$ and $5'11$. (If it in fact lies in this range, then you know that it does.)

Now here's the argument. You know that he believes the height of the tree is within 6 inches of his best guess. So, given your uncertainty about his best guess, you believe (and let's suppose you know) that he believes the tree is somewhere between $5'3$ and $6'5$. But

[20]To simplify the discussion, I'll be considering point-valued best guesses. But the argument goes through as well with interval-valued best guesses, so long as agents do not have precise knowledge of their partner's interval. (Goodman 2013 explores models of interval-valued appearances in a KK-denying framework, but his models could be adapted to Stalnaker's "cliffhanger knowledge" setting.)

the process continues: for you know that his best guess about your best guess depends on his own best guess. Thus if his best guess is $5'9$ (as you think it might be), then he knows that you know that the tree is somewhere between $5'2$ and $6'4$, but no more. Similarly if his best guess is $5'11$, he knows that you know that the tree is somewhere between $5'4$ and $6'6$. So you know that he knows that you know only that the tree is between $5'2$ and $6'6$. By the obvious extension of this pattern of reasoning, it's clear that it is common knowledge only that the tree is of some positive height in inches.

Here's an abstract Kripke model to represent the situation for two agents A and B. Let worlds be elements of $\mathbb{R}^3_{>0}$, triples of positive real numbers representing: an actual tree height, A's best guess and B's best guess. In a simple version, we define knowledge in terms of belief. We do everything for player A, but the same holds for B, with the obvious changes. A's belief relation is as follows:

$$R^A_B(x, y, z) = \{(x', y', z') : |x' - y| \leqslant \sigma, y' = y, |z' - y| \leqslant \epsilon\}$$

(The subscripted B is for belief.) That is, if the player's best guess is y, he or she believes that the tree lies within σ of y, and believes that her partner's best guess lies within ϵ of y. (Note that the relation is serial, euclidean and transitive so that we have a model of $KD45$, Stalnaker's preferred logic for belief (Stalnaker 2006).) We then define knowledge as follows:

Case 1. $R^A_K(x, y, z) = R^A_B(x, y, z)$ if $(x, y, z) \in R^i_B(x, y, z)$;

Case 2. $R^A_K(x, y, z) = \{(x', y', z') : |x' - y| \leqslant \sigma, y' = y, |z' - z| \leqslant |y - z|\}$ if $|y - z| > \epsilon$ and $|y - x| \leqslant \sigma$;

Case 3. $R^A_K(x, y, z) = \{(x', y', z') : |x' - x| \leqslant |y - x|, y' = y, |z' - y| \leqslant \epsilon\}$ if $|y - x| > \sigma$ and $|y - z| \leqslant \epsilon$; and finally

Case 4. $R^A_K(x, y, z) = \{(x', y', z') : |x' - x| \leqslant |y - x|, y' = y, |z' - z| \leqslant |y - z|\}$ if both $|y - x| > \sigma$ and $|y - z| > \epsilon$.

It's then easily checked that the model obeys positive introspection at every point. In the concrete story above, σ was set to $6''$ and ϵ was set to $1''$. The important fact is this: if $\epsilon \neq 0$, then from *any* point, only the universe is common knowledge.

This gives us a model against which we can assess the argument, but it's time to consider the premises of the argument, or the ingredients of the model, in more detail, at a conceptual level. Before doing so, it's important to note that the argument *nowhere* uses a margin-for-error principle about knowledge like Williamson's (2000: Ch. 4 and 5).

We assumed that both agents are positively introspective about their knowledge; this assumption (together with the factivity of knowledge) is in contradiction with Williamson's margin for error principle.

First, we made the simplifying assumption that your range and your partner's are the same (σ is the same for both players, and so is ϵ). But we can relax this assumption without any effect whatsoever on the argument. If each player has a constant σ and constant ϵ at every point, the argument goes through as before.

Second, we assumed that you and your partner are both certain of each other's range. If we'd allowed for uncertainty instead, the argument would be longer, but the result would be unchanged. For conditional on any hypothesis about your partner's (constant) range, nothing substantial would be commonly known about the height of the tree. Adding uncertainty about his range generates more uncertainty about what he knows, but this uncertainty is not compensated for by greater knowledge about the height of the tree.

Our third assumption requires more discussion. We assumed that both you and your partner's ranges about the tree and about one another's best guess do not vary as a function of your own best guess or the actual height of the tree. For example, if your range is $\pm 6''$ when the tree is $5'10$, it's also $\pm 6''$ when the tree is $5'$, or even $4''$. (Clearly in the latter case, your range is in a sense a function of the tree-height, since you would not then consider it possible that the tree was $-2''$ tall (whatever that means). When I say that your range is constant, I mean that it's constant except for this kind of boring exception.)

This is a substantial assumption. If we think in terms of our Kripke structure, it ensures that for every positive real number x, a world at which the tree is x inches tall is "reachable" by a finite number of steps using yours and your partner's (A's and B's) accessibility relation. Since the common knowledge accessibility relation is the transitive closure of the union of your respective relations, it follows that nothing non-trivial is commonly known about the height of the tree.

But rejecting this assumption about reachability is unpromising. Suppose that prior to the play of the game, researchers conduct extensive tests on the reliability of the subjects at assessing tree heights. In each case, subjects report their best guess, and researchers plot where this best guess falls relative to the actual tree height. For each individual, they produce a "score sheet", matching actual tree heights with ranges of

accuracy, perhaps defined by a statistical metric, for the different individuals. Each player of the game then receives both her own score sheet, and her partner's.

We can now ask: What would the score sheet have to be like for agents to commonly know that the height of the tree was greater than 6″? There would have to be some n greater than or equal to 6, such that no finite steps from one player's "best guess" can reach a tree height of n inches. For actual tree heights which are close to but greater than n, the subjects' best guesses would have to be arbitrarily precise, at least on the lower bound of their guess. To make this slightly more concrete, suppose that there was a greatest such unreachable n, and that it was 7 inches.[21] In the imagined scenario, if the tree was 7.1 inches, the score sheet would say that they reliably guessed it to be at least 7. Even if it was 7.00001 inches, they still guessed it to be above 7 inches, and the same for any real number greater than 7.

This hypothesis is absurd. There is clearly no n around which we have arbitrarily precise estimates of tree heights. But to heighten the absurdity, let's suppose that there is such a greatest unreachable n and that it is again 7 inches. Now suppose that instead of asking subjects to coordinate if the tree is taller than 6 inches, we ask them to coordinate if it is taller than 8 inches. If the tree is 5″10, it seems just as clear as before that agents will coordinate. But since 7 was the greatest unreachable n by hypothesis, then in the revised situation where the relevant information is whether the tree is taller than 8 inches, once again, the subjects will not commonly know this fact. The two points made in terms of 7 and 8 apply completely generally. Any choice of greatest unreachable n, or least reachable n leads to the absurd verdict that our assessments are arbitrarily precise close to n. Moreover, any choice of such an n will make subjects' behavior unrealistically sensitive to slight differences in the announcements made by the game show hosts. Denying the reachability assumption is therefore hopeless.

The fourth and final premise of the argument is that the agents do not know whether their partner's best guess is *exactly* the same as theirs. But to assume that you know your co-contestant's precise best guess is to assume an absurd level of interpersonal knowledge of phenomenology. After all, his visual system is subject to *some* errors which are independent of yours; you can't rule out the possibility that his experience of the tree differs subtly from yours.

[21] "Suppose" because of course it might be that the best guesses displayed the structure of an interval which is closed on its lower bound, so that there was no greatest unreachable n, but there was a least reachable one. In the latter case, we'd still have the bizarre feature that when the tree was 7 inches, the players never or very rarely guessed anything below 7.

The argument so far has been given in terms of knowledge. But what about belief? Belief can of course outrun knowledge. It might be that, although we don't know our partner's best guess, we still believe that it has a precise value, and even believe that we know this. Can we save common belief by this route?

We cannot. Just as it's possible for people to form beliefs which outrun their knowledge, it's also possible for people to be more conservative, to form only those beliefs in a given domain which coincide with knowledge. Humans are systematically overconfident in our own abilities. But the coordination behavior in this kind of case seems eminently *rational*. We do not want an explanation of the behavior which relies crucially on our irrational overconfidence in our own abilities. Worse yet: this way of preserving common belief fails to accord with our judgements about the case, since we judge the information in question to be public or shared in the relevant sense, independently of our judgements about the ways in which the participants form beliefs, whether they form them according to what they may know, or not.

So the participants in SAPLING don't commonly know (or, in general, commonly believe) that the tree is taller than $6''$.

1.5.2. What SAPLING Shows. This example spells doom for the abductive argument for common knowledge. It shows that a plausible thought about common knowledge is incorrect. One might think that if information is "in plain view" to all, it is commonly known. The example shows that this *prima facie* plausible thought is wrong. The tree is in plain view to all. But its height is not a matter of common knowledge. In general, this makes it clear that the notion of common knowledge does not carve up social or informational phenomena in the right way. We should hope to explain such "plain view" situations in the same way in all examples. So we should hope for a better explanation than the one common knowledge offers.

But can my preferred theory do better? It can. We encounter such "plain view" situations every day. And every day we succeed in finding the better coordination equilibrium: not the one where we lie inertly wondering which equilibrium others will choose, but the one where we engage in projects and move around the social world. Just as in coordinated attack and electronic mail, our dispositions to act are formed in such situations, and we form beliefs about others' behavior on the basis of this experience. When we encounter such a situation, where it's clear that the tree is taller than $6''$, it's also clear what we should do, and we do it.

1.6. Completing the Negative Argument

The abductive argument for common knowledge relies on the premise that common knowledge provides the best explanation of behavior. In coordinated attack—an example often used to motivate the importance of common knowledge—we saw that the behavior predicted by common knowledge was neither intuitively rational, nor what we observe. A different hypothesis predicted the correct behavior. Moreover, we saw that a formal argument based on a class of simple models of the situation made inappropriate use of idealizations implicit in the model.

In electronic mail, the situation was more complex. It took more work to show how a simple explanation of people's behavior delivered the verdict that that behavior was rational. But the simple explanation satisfied more of our demands in that case than any explanation in terms of common knowledge could.

Finally, SAPLING showed that a simple view of "plain view" situations was incorrect. One might have thought that when information is open to all as it was in this situation, then it is commonly known. But the example shows that this thought is mistaken. Even in such "plain view" situations, the information we share need not be common knowledge.

But if the foregoing arguments are correct, why are models which assume a great deal of common knowledge so prevalent? The answer is that these models *are* simpler, just not in a way which gives us any reason to think that people have common knowledge. We saw this already in coordinated attack. The models in which agents attack only if they commonly know that they do were formally simpler than my alternatives. But they were simpler in ways which made them worse, not better predictors of what we observe. Common knowledge doesn't just simplify our models; it also helps in a further way, which we have not yet encountered, by simplifying the statements of general results.

To see this, consider rationalizability, which I mentioned in the discussion of the electronic mail game. In a pair of papers, Bernheim 1984 and Pearce 1984 showed that the rationalizable strategy profiles—behavior predicted under common knowledge of rationality—are exactly the profiles which survive iterated elimination of strictly dominated strategies. (The result was subsequently extended by Brandenburger and Dekel 1987 to include correlated beliefs.) But the use of common knowledge here is simply to encompass every finite game in one simple statement. Eliminating common knowledge from the statement of the theorem is trivial: informally, the revised statement would say that, for all natural numbers n, if there's mutual belief^n that players are rational,

then the only strategies which receive positive support in players' conjectures are those which survive n rounds of deletion of strictly dominated strategies. For every finite game, there's some finite n such that no strategies are deleted in the $n+1$th round of deletion. So for every finite game, there's some finite n such that if players mutually believe^n that the others are rational, they'll play only those strategies which survive infinitely iterated deletion of strictly dominated strategies (since those are exactly the strategies which survive n rounds of deletion).

The example of the "muddy children", another story often used to motivate the importance of common knowledge, is no different. Some number of the muddy children have mud on their foreheads, and each wants to determine whether she herself has mud on her forehead. The children can't do this unless someone makes an announcement saying that at least one of them has mud on her forehead. Once the announcement is made—making it "common knowledge" among the children that one of them has mud on her forehead—then if the children announce whether they know that they have mud on their foreheads or not, they can each figure out whether they are muddy or clean in a finite number of rounds. But here, too, for finite numbers of children, we need only mutual belief^n for some finite n: common knowledge is only used because it allows us to give a single statement covering all sets of children. (In any case, for sufficiently high n, say, 100 muddy children, we of course shouldn't expect that even logicians *will* reason in this way "in the field", at least if they are unfamiliar with the problem.)

The prevalence of common knowledge and belief in the technical literature is readily explained by these two facts: common knowledge simplifies our models, and simplifies our formal statements. But neither of these reasons provides any evidence for the empirical prevalence of common knowledge or belief. Some, it's true, have explicitly taken common belief and knowledge to give genuine explanations of observed phenomena. But on investigation this is more owing to the popularity (or catchiness) of the phrase than it is to any *need* for the infinitely iterated state. In his book *Rational Ritual*, for example, Michael Chwe argues that a series of social practices can be explained on the assumption that they aim at generating common knowledge (Chwe 2001). But Chwe never motivates the need for this infinitely iterated state, and aside from stylistic reasons, I don't imagine he'd be averse to replacing his statements with "finitely iterated mutual knowledge". When he argues that the progress of the king in medieval England was designed to promote common knowledge that the new king was king, for example, it's perfectly compatible with the data that the progress was designed to promote higher orders of mutual knowledge

among a greater percentage of the population. Replacing Chwe's definitions in this way doesn't change his central point: that sometimes, social practices are best understood as aiming at generating more shared information. The point is a very general one, and carries over similarly to many other explanations which use common knowledge.[22]

The theorists themselves in general do not believe that agents have the common knowledge or belief ascribed to them in standard models. This point can be illustrated by considering recent developments in the field. Recent years have seen a move away from the common knowledge assumptions associated with the high theory of earlier mathematical game theory. In part this is because more care has shown that core theoretical concepts such as Nash equilibrium in fact require no common knowledge at all: neither of the game, nor of other players' actions.[23] This revolution in our understanding of the mathematics has liberated game theorists from many standard common knowledge assumptions. The rise of behavioral economics, too, has led to greater interaction with psychologists, and a search for empirically viable and psychologically realistic models of behavior. The tradition of level-k models invoked throughout this paper is a particularly rich vein of empirical explanations which use only low finite levels of higher-order reasoning.[24] These models are designed to reduce common knowledge assumptions, and accordingly they have enjoyed comparative success at predicting observed behavior.

So the argument from the fact that common knowledge and belief simplify our models and our statements fails in the general case for the same reason that it failed in the case of coordinated attack. This version of the abductive argument fails, just as the one which claims that common knowledge unites a wide range of phenomena.

As noted in the introduction, this verdict has at least one important ramification for mainstream epistemology. Dan Greco has recently suggested the following line of argument in favor of the KK principle: the social sciences invoke common knowledge; arguments against KK typically have the result that common knowledge is impossible; so we should reject those arguments. But the suggested argument—as the reader can see—is at best unpromising. The uses of common knowledge in the social sciences are an accident of (theoretical) social scientists' pursuit of formal simplicity—whether to simplify the models they consider, or to simplify the statements of general theorems. In this case, we cannot infer from the prevalence of common knowledge in the models to its

[22]For example the use of common knowledge by Steven Pinker and his coauthors (2008) to explain the "logic" of "indirect speech" can be similarly explained away.
[23]See above, n. 15 for references.
[24]For standard references, see above, n. 6.

importance to human social life or its psychological reality. To do so would be to commit the error of an engineer who takes mathematical physicists' models of point-like objects to be evidence that bridges have no spatiotemporal extension.

1.7. Antiluminosity and Common Knowledge

This completes the main argument of the paper. The standard abductive argument in favor of common knowledge fails. But could other arguments succeed? In this final section, I close by considering the relationship of Timothy Williamson's antiluminosity argument to debates over the existence of common knowledge. I'll suggest that even those who maintain the KK principle in the face of this argument should not therefore hold that common knowledge of "non-trivial" propositions is possible.

Recall that a condition c is luminous just in case, if a subject s is in c, s is in a position to know that s is in c. Recall furthermore that agents commonly know that p if and only if they mutually know^n that p for all n. As Williamson himself points out (2000), common knowledge is luminous by definition. Since for each natural number n, the agents mutually know^n p they also each know that they mutually know^n that p (this is just the claim that they mutually know^{n+1} that p). Thus if the agents commonly know a proposition, they each know that they commonly know it.

If common belief were prevalent, then it is plausible that it would sometimes occur in such a way that each of the beliefs amounted to knowledge. If these social beliefs never coincided with knowledge, our social lives—or whatever common belief is supposed to explain—would be a practice founded on irrational overconfidence, a practice of forming beliefs which have no hope of achieving their aim. But now, since common knowledge is by definition luminous, if common belief were prevalent, there would be prevalent "non-trivial" luminous conditions. So those who accept Williamson's anti-luminosity argument as applied to all conditions c must reject not just the possibility of common knowledge; they will also be pressed to deny the commonness of common belief.

But the audience for denying common knowledge on the basis of anti-luminosity is much broader than this argument suggests. The most popular strategy for rejecting the antiluminosity argument has been to hold that for certain conditions c, the belief that one is in c is related to the obtaining of c in a special way. In the jargon, for these conditions, there's a "constitutive connection" between the belief and the obtaining of the condition. By implication, the proponents of this class of views *do accept* that conditions which do

not enjoy such a constitutive connection fail to be luminous, that is, an agent may be in the condition but not be in a position to know that she is.

Does common knowledge enjoy this constitutive connection? The following example will show that it does not.

> Three of us are in a small room, and a public announcement is made stating that one of us has mud on our foreheads.

This is a paradigm example of common belief; since we may suppose it is truthful and reliable, here, if ever the proposition announced should be common knowledge. But now imagine a series of slight variations on the original case. In each new variation, my eyelids are a little more closed. Each case differs from the last by a change which you are unable to perceive. At the end of the series, however, my eyes are completely closed, and you believe I may be asleep.

Clearly the content of the announcement is no longer common knowledge (or even belief). The series of slight variations is exactly the kind of series envisioned in the original antiluminosity argument. But here it's utterly implausible that *my* being awake is constitutively connected to your belief that I'm awake, or that your belief that the third person knows I'm awake is constitutively related to this third person's belief that I'm awake.[25] To assume otherwise is to assume a conspiratorial relationship between my state of mind and yours. So the public character of the announcement is not itself luminous. But the state of common knowledge is, by definition, luminous. So we have a gap between the obtaining of the precondition for publicity, and the epistemic features of the situation which are supposed to explain this publicity.

So I think those who reject the antiluminosity argument for the restricted class of conditions where beliefs are "constitutively connected" to the obtaining of the condition should agree that common knowledge never occurs in examples such as the one above. But if common belief explicates what's public about the information in this case, then, as I've argued, sometimes that common belief will amount to common knowledge. So if one accepts even a restricted version of the antiluminosity argument, one should reject the prevalence of both non-trivial common belief and non-trivial common knowledge.

[25]To make this point slightly more precise, it's obvious that neither of the two principles Srinivasan (2013) offers to the "luminist" hold with "she feels cold" and "is in a condition R" replaced by "an announcement is public":

(CON) If S has done everything she can to decide whether she feels cold, then she believes that she feels cold if and only if she feels cold. (15)

(CONFIDENCE-SAFETY) If in case a S knows with degree of confidence c that she is in a condition R, then in any sufficiently similar case a* in which S has an at-most-slightly-lower degree of confidence c* that she is in condition R, it is true that she is in condition R. (16)

Now of course there are still others who not only reject the unrestricted antiluminosity argument, but reject the version of the argument used for *knowledge*. They hold that if an agent knows that p, she knows that she knows that p. Two views of this form are most prominent: Robert Stalnaker's and Dan Greco's. As Greco explains the idea, the principle KK follows from general facts about what it is for a state to encode information. If a state carries information, the fact that it does so is sufficient for it to carry the information that it carries the original information. Since Greco and Stalnaker take belief to carry information in the way that tree rings do, they hold that to believe that p is just to believe that one believes that p. They argue that we should reject the antiluminosity argument as applied to knowledge because the state of knowing is identical with the state of knowing that one knows.

Whatever we may think of this line of thought in the case of knowledge, it simply cannot be extended to the interpersonal case. Any reduction of common knowledge to finitely iterated mutual knowledge cannot proceed by reducing a state with a greater number of iterations to a state with fewer. For even if as a medical matter people were *incapable* of distinguishing between say mutual knowledge^{100} and mutual knowledge^{101}, these states are nevertheless *different* states. In particular, they are *not* necessarily equivalent. The former state just doesn't have enough information to entail the latter, and no finite iterations have sufficient information to entail the infinitely iterated state. That's one of the many things that Rubinstein's beautiful example dramatizes.

So even if one rejects the antiluminosity argument altogether, it's a very difficult challenge to come up with a philosophically motivated way of understanding how common knowledge *could be* luminous. Certainly, we haven't been given such an explanation to date. Stalnaker and Greco's prominent style of preserving introspection principles for individual attitudes doesn't have much promise for explaining why we should preserve them in the interpersonal case. This doesn't show that those responses fail on their own terms—in fact, they might succeed. But I think it does demonstrate conclusively that the considerations adduced in the antiluminosity argument are damning for the possibility of common knowledge, *even if they aren't for knowledge itself.* One of the aims of this paper has been to show that we shouldn't be concerned that we never have common knowledge. Life can go on as usual without common knowledge or belief. Once we see that common knowledge is not needed to explain behavior, we should embrace a world without this outlandish epistemic state.

1.A. Alternative definitions

In the main text, I focused on the definition of common knowledge which takes it to be an infinitely iterated attitude. But other definitions have been offered, and one might think the core arguments of the paper leave these definitions unaffected. In this Appendix I'll make some schematic remarks about why I think this position is mistaken: SAPLING is a problem for alternative definitions of common knowledge and belief, just as it is for the definition as an infinitely iterated attitude.

Consider, for example, the *reason to believe definition*. In a schema that will already be familiar, you and I have mutual reason to believe^1 that p if and only if we both have reason to believe that p; we have mutual reason to believe^n that p if and only if we have mutual reason to believe that we have mutual reason to believe^{n-1} that p. So

> You and I have RB-common belief that p just in case, for all natural numbers n, we have reason to believe^n that p.[26]

In SAPLING, because the agents don't have common knowledge, they don't have RB-common belief. To dramatize the point, suppose two people have read Section 1.5, so that they understand acutely the difficulties imprecise "best guesses" pose for the achievement of common knowledge. This would undermine their mutual reason to believe^n that the tree is taller than 6 inches, for some finite n. So the example undermines this "reason-to-believe" explanation of public information as well.[27]

Another prominent definition, the *fixed point* definition of common knowledge, is somewhat different:

> You and I have FP-common knowledge that p if we both know that p and q,

where q is the proposition that we both know p and q.

A special case of the fixed point definition focuses on a natural way for representing the fixed point q, by way of an indexical:

> You and I have I-common knowledge that p if (1) we both know that p, and *this*,

[26]Lewis's definitions—the obvious omission from the three here—is discussed in an appendix. The discussion of RB common belief is intended to apply also to Heal 1978 and Peacocke 2005.

[27]Margaret Gilbert 1989 proposes the following theory of common knowledge: agents commonly know a proposition if, if fully rational agents were in the same situation, they would have common knowledge of the infinitely iterated variety. But unless fully rational agents have perfect knowledge of each others' best guesses, Gilbert must also deny that common knowledge obtains in the example above.

where "this" refers to the whole state of affairs in (1).[28]

Sapling presents a dilemma for proponents of FP-common knowledge or its I-common knowledge variant. Do the agents above have I-common knowledge or not? If they do not, then the example is also a counterexample to this explanation of what makes information public or "open". If it is claimed that agents do have I-common knowledge in the example, then I-common knowledge is poorly motivated, as I will now argue.

Agents can sometimes have seemingly contradictory beliefs under different "guises". However we explain this phenomenon—by "compartmentalization or "fragmentation", by fine-grained contents, or even by genuinely contradictory beliefs—sometimes action and reasoning will follow one of the "beliefs" as opposed to the other. For example, Kripke's Peter, who assents to the claim that Paderewski the pianist is not Paderewski the president of Poland, will "use" only one of these descriptions when discussing famous pianists (Kripke, 1979). Similarly, Ann and Bob may have I-common belief, even though, under the "guise" involving iterated belief, Ann may believe that Bob doesn't believe that Ann believes that Bob believes that p. In the case above, it's clear that this is true: the agents don't have common belief or knowledge under the more straightforward guise, even though they might infer it if they utilized their common knowledge of the indexical state of affairs. But it seems plausible that it's the straightforward guise which governs their action and reasoning, and not the infinite conjunction they might infer by the appropriate use of the indexical. And if the information can be open in spite of agents' acting on only finitely iterated belief or knowledge, it's unclear why we need a definition which gives us the infinite conjunction in the first place.

These remarks remain schematic. But I think they show that there's no simple retreat to getting infinitely many iterations by some sly route which differs from the explicit definition considered in the main text.

[28]There's an in-house metaphysical dispute between those who favor a version of the fixed point definition and those who prefer the indexical one. For this debate, see Peacocke 2005, Appendix, *contra* Barwise (e.g. 1988). Nothing I say here will turn on this difference.

CHAPTER 2

A Theory of Common Ground

Abstract. A new theory of common ground is proposed.

2.1. Introduction

In the 1970s Robert Stalnaker (and, independently, Lauri Karttunen) described the notion of the *common ground* of a conversation: a body of information shared among participants in the conversation (Stalnaker 1974, 1978, Karttunen 1974).[1] In the first instance, Stalnaker proposed that these were the propositions participants "mutually assume that they take for granted". But in later work (especially 1998, 2002, 2014), he has elaborated this idea, arguing that the attitudes which determine the common ground have the "iterative" structure of common belief, where a group commonly believes a proposition just in case all believe it, all believe that all believe it, and so on.

At the outset, the common ground was envisioned as an enrichment of formal approaches to semantics, allowing for an equally formal explanation of certain pragmatic phenomena. Stalnaker's theory of assertion, for example, uses properties of the common ground to explain divergences between the semantic content of a sentence in context and what is said by an assertion of that sentence in that context. Stalnaker himself has continued to emphasize, broadly following Grice, that the laws of pragmatics—including, for example, the theory of assertion—should be derivable from general laws of rational interaction. Since the common ground is a primitive postulate of Stalnaker's pragmatic theory, he accordingly holds that the common ground itself is a mandatory feature of anything which is to count as rational communicative activity.

Some aspects of the theory of the common ground are hotly contested today, even among those who use the common ground in their own linguistic theories. For example, "dynamic" semanticists argue that the common ground plays an even more central role in communication than Stalnaker believes it does. These theorists have developed the hypothesis that the value of a sentence is a function from common grounds to common

[1]Thanks to Frank Arntzenius, Emil Möller, Daniel Rothschild and Timothy Williamson for detailed comments on this material. Bob Stalnaker generously gave helpful comments on a much earlier draft. Parts of this paper are deeply indebted to numerous conversations with Jeremy Goodman.

grounds (often in this setting called "contexts").[2] Whereas Stalnaker's original theory adhered to a more traditional notion of semantic content, these theories make the much more radical step of holding that the meaning of some if not all expressions cannot be understood in isolation from their effect on the common ground. In a second, equally prominent example, discourse representation theorists argue that the common ground must have a highly articulated logical structure to account for phenomena such as tense anaphora.[3] Whereas the common ground was initially understood as a body of information with only Boolean structure, discourse representation theorists have endowed it with a much more complex structure, which also requires new laws governing the dynamics of update.

But one aspect of Stalnaker's theory has enjoyed almost unanimous support: that if there is a common ground, the attitudes which determine what is common ground have the iterative structure of common belief.[4] Stalnaker and his followers have proposed various different hypotheses about the precise nature of the attitudes which determine the common ground, but in each case they have agreed that the attitudes have this infinitely iterated structure.[5] In other traditions, too, the point is often simply taken for granted. To return to the earlier examples, dynamic semanticists and discourse representation theorists are not often given over to detailed discussion of the epistemology of the common ground (their "contexts"). But the minimal discussion one does find refers favorably to Stalnaker's views on the common ground.[6]

But is this unanimity justified? There are at least two reasons for thinking that Stalnaker's theory of the epistemology or psychology of the common ground is mistaken. First, the "iterated" theory of common ground makes intuitively irrelevant and even unnecessary features of agents' psychology essential to the possibility of communication. To see this, suppose scientists discover that an isolated linguistic community—whom I will call "Luxembourgers"—do not possess a part of the brain which is active whenever non-Luxembourgers think about what others believe they believe others believe they believe (and similarly, for other propositional attitudes relevant to conversation). Moreover, if you ask a Luxembourger about people's higher-order beliefs beyond this level, she will shrug and say that she has no idea about attitudes of this kind. Worse still: except for

[2]For a good survey of this development, see van Eijck and Visser 2012. Standard references include Heim 1983; Groenendijk and Stokhof 1991 and Veltman 1996.
[3]For an excellent overview, see Kamp et al. 2011.
[4]An important complicating example is Veltman 1996.
[5]See, e.g. Clark and Marshall 1981, Clark 1996, Yalcin 2007: 1007 for examples.
[6]See, for a "dynamic" example, Portner 2007: 357; for an (admittedly less clear) example from discourse representation theory, see van Eijck and Kamp 2011: 191.

the consistent shrug, Luxembourgers' dispositions to act regarding higher-order beliefs beyond the specified limit are so inconsistent that by far the best hypothesis concerning their behavior is that they have no such beliefs.

If the common ground has the infinitely iterated structure of common belief, this scientific discovery about Luxembourgers would amount to the discovery that Luxembourgers do not have a practice of assertion. In fact, given Stalnaker's views about the importance of the common ground to communication, some of his remarks suggest that we should think of this discovery as a discovery that Luxembourgers' practice of communication is fundamentally defective. But both of these judgments are implausible. It seems far more likely—at least on a first pass—that the members of the hypothesized community might be very normal, even with respect to their communicative practices. And this point holds regardless of whether we think people in fact generally *do* commonly believe propositions which are relevant to their conversations. The point is simply that the common ground should not be defined so that it requires an infinite hierarchy of attitudes.

A second example of the same form may help to clarify the point. Suppose it turns out that building computers which can "think" about others' thoughts is considerably more resource intensive than building them with the ability to think about others' actions and concrete physical objects. We may suppose we are in a situation where it would require a whole new set of programming tools, as well as new hardware, to endow computers with the capacity to describe others' thoughts. Moreover, to make the point as clear as possible, let us suppose that with each embedding of attitudes (thinking about others' thoughts about others' thoughts...), the programs and hardware required to implement the possibility of machines engaging in this sort of "cognitive" activity become harder and harder to design. And, finally, suppose that we know that, even if we did successfully write these programs, and build this hardware, the computing power utilized by the new processes would swamp all other processes.[7]

Here again, it seems that we would not need to overcome all of these design hurdles in order to create a computer which could mimic various basic communicative activities.

[7]According to some philosophers—and to a large body of work in psychology—the hypothesized Luxembourgers and these computers are in fact much like real people: using theory of mind in reasoning requires a great deal of effort, and humans' dispositions regarding embedded attitude reports are highly unstable. For the first point, see Lin et al. 2010; for the second see especially Kinderman et al. 1998 (cf. Stiller and Dunbar 2007). In fact, one might even go so far as to suggest that the computational difficulties described in the main text correspond to evolutionary pressures away from devoting cognitive resources to unimportant questions about iterated attitudes (unless the capacity for them was a consequence of some other, more fundamental capacity).

The computers might be capable of performing a wide range of normal conversational behavior, even if they do not have an infinite capacity for thoughts about thoughts about thoughts. Depending on the precise limitations of these machines, they would be more or less obviously different from us—for example, in their ability to parse attitude reports or make inferences about others' beliefs. But these differences in degree do not seem to be ones which would prevent them from engaging in many of our ordinary linguistic practices.

The second reason for doubting Stalnaker's theory of the epistemology or psychology of the common ground is that the theory predicts a sharp divergence between one-on-one, face-to-face communication and other kinds of communication. Consider the following example:

> Sam is at the airport, about to leave town for a few days. He knows that his neighbor has set up a device which will destroy Sam's house either today at noon or tomorrow at noon. Sam knows that the day of destruction depends on the outcome of a coin flip the device will perform at 11:30 today. Sam knows that his wife will stop by today at 1pm, to stay at home for the night. He has left a note telling her about his neighbor's plans and the device. If she sees it, the house will not have been destroyed, but will be destroyed only the following day.

In this case, the contents of Sam's message cannot be commonly believed, since Sam does not believe that his house has not yet been destroyed, and so does not believe that his wife believes that Sam knows about the neighbor's dastardly plans. If the message gets through, Sam's wife will know that Sam knows about the plans, but she will also know that Sam does not know whether she knows that he knows about the plans.

If the common ground has the structure of common belief, this kind of message passing will be an unusual or defective form of communication. Written communication does seem to involve a variety of complexities which may not be present in face-to-face conversations, but these complexities do not seem to mean that, at the appropriate level of abstraction, note passing is a fundamentally different form of communicative activity than conversation. To mention just one consideration, agents who have a practice of assertion would thereby be able to understand taped recordings of speakers of their own language. And it seems they would be able to do this, even if they did not have the capacity to pretend to have a common ground (or to have whatever other complex

attitudinal states one might wish to invoke to preserve the iterated theory in the face of noisy message passing examples such as the one above).

More mundane examples than Sam's odd situation also generate a related problem. People who make speeches at weddings know that some listeners will be whispering to their neighbors, but their speeches are generally unaffected by the fact that they don't know who their audience is. Even more clearly, it doesn't bother speakers in such situations that their audience doesn't know who the audience is—although such knowledge *would* be required if the group were to achieve common knowledge. This example is no different from that of writers who send their books into the world to be read by an unspecified audience. As in our first example of written notes, these more mundane examples are still instances of intricate social practices which undoubtedly involve very complex psychological states and cultural backgrounds. But once again, these forms of communication do not, at the appropriate level of abstraction, strike us as forms of communication that are fundamentally different from one-on-one, face-to-face communication.

These two reasons for doubting Stalnaker's theory of the epistemology or psychology of the common ground are certainly not full-fledged arguments against the theory. I don't wish to pretend that the proponent of this "iterated theory" of common ground has no way of responding to them. But these two problems with the Stalnaker's view do motivate the search for a different theory of common ground, which delivers more intuitive verdicts about these cases. The aim of this paper is to provide just such a theory of common ground.

In developing this theory, I will follow Stalnaker in considering only propositional attitudes. This means that the theory I propose here will be limited in important ways. For example, my models of common ground cannot be used to describe some of the more complex structural features of the common ground we find in discourse representation theory. But just as with Stalnaker's theory, I intend my theory as a framework or starting point, from which other developments may begin. Much of the structure missing in my theory could easily be added. To take a simple example, the universe of objects utilized in discourse representation theory can be explicated in a first-order modal language with a distinguished predicate of "salience". We say that if it's common ground that o is salient, then o belongs to the universe of objects. Many other structural features—for example ordering sources, probability distributions, and self-locating attitudes—can be included by similarly small extensions of the theory presented here.

But still these extensions will be left for another day. Our question is: what attitudes must agents be capable of having about each other in order to have something recognizable as the common ground? Since this question can be posed using a purely propositional language, we will only need a propositional theory to answer this question.

The plan of the paper is then as follows. Section 2 introduces the notion of conversational acceptance, and canvasses various different theories of acceptance. The bulk of the paper is then dedicated to showing how the minimal theory can be used to explain three important conversational phenomena: speaker presupposition (Section 3); update of the common ground (Section 5); and the infelicity and felicity of certain patterns involving epistemic sentences (Section 6). Section 4 takes a brief digression to consider a direct argument for the iterated theory as opposed to the minimal one. Section 7 concludes. The Appendices present a variety of formal facts used in the main text. The main text can be read without referring to the appendices. The interested reader may find it most convenient to read a given appendix immediately after reading the main text which refers to it.

2.2. Defining the Common Ground

I will develop the theory of common ground using the notion of "conversational acceptance", or "acceptance for the purposes of conversation". When I use the word "accept", it will mean "conversationally accept". Using this notion, I define a sentential operator "it's common ground that", as follows:

Common Ground: It's common ground that p if and only if all participants accept that p.

But what is acceptance? Although I will use this notion extensively in the paper, I will not settle this important question here. Instead, my aim will be to understand the notion—and its role in theories of conversation better—by canvassing various theories of acceptance. Accordingly, I'll begin with a taxonomy.

2.2.1. Acceptance. Our first two theories of acceptance make use of the idea that a conversation has what we might call a *conversational tone*.[8] This conversational tone determines what attitude is appropriate to the propositions introduced in the course of the conversation. This attitude might be belief, knowledge, supposing, pretending, or something else. I will make the simplifying assumption that conversational tone determines a unique attitude, although in real conversations different attitudes may be

[8]The phrase comes from Yalcin 2007.

appropriate to different (subsets of the) propositions which are relevant to the progress of the conversation.

The three theories of acceptance I will consider are as follows:

(1-Identity) To accept a proposition for the purposes of conversation is to take the attitude determined by the conversational tone to that proposition.

(2-Necessity) To accept a proposition for the purposes of conversation it is necessary but not sufficient that one take the attitude determined by the conversational tone to that proposition.

There are three different ways in which this might be so:

(a-Attention) To accept a proposition is to attend to it, and to take the attitude determined by the conversational tone to it.

(b-Table) To accept a proposition is to believe that the proposition is "on the table", and to take the attitude determined by the conversational tone to that proposition.

(b1-Reducible) To believe a proposition is "on the table" is to believe that all others take the attitude determined by the conversational tone to this proposition.

(b2-Primitive) The property of "being on the table" is taken as a primitive property, which is understood independently of others' attitudes, perhaps explicated as a mixture of relevance, salience, importance and other features of appropriateness.

(c-Distinct) Acceptance is a distinct psychological attitude, which happens to be accompanied by the attitudes determined by the conversational tone. Whereas the foregoing views of type (1) and (2) define acceptance in terms of other attitudes, this final theory does not

take such a strong view of the relationship between acceptance and the attitude determined by the conversational tone.

(3-Primitive) Acceptance is a primitive attitude, and one need not take other attitudes to propositions one accepts. This view may deny the idea of a conversational tone altogether, or it may simply hold that conversational tone is not very important. If the idea of a conversational tone is maintained, this last view holds that one may accept a proposition without supposing it, even when the conversational tone determines supposition as attitude appropriate to that proposition.

It will be useful to begin with a few remarks about the advantages and disadvantages of these theories of acceptance, and its role in the common ground.

The first, "Identity" theory of acceptance takes conversation to be a psychologically heterogeneous phenomenon: there is no single attitude which is common to all conversations. For some, this feature of the theory may be undesirable: the phenomenon of conversation may seem sufficiently uniform that it demands a unified psychological explanation. A second, perhaps more interesting problem with the identity theory concerns examples in which belief is the attitude determined by the conversational tone, but in which one does not accept every proposition one believes. For example, even if everyone at a fancy party can see that there is toothpaste on the hostess's forehead, the polite party-goers may still not presuppose this proposition in the course of their conversations. The first theory of acceptance owes us an explanation of this phenomenon.[9]

The Necessity theories of acceptance offer various different answers to the first of these questions—about the psychological uniformity of conversation. For example, both (2a) and (2b1) offer a conservative story about the attitude which unifies conversations. Neither of these theories posit a new, distinct attitude which we employ only in conversation; instead they hold that conversation involves a combination of attitudes which are familiar from other areas. In the Attention theory, this attitude is attention, plus the attitude determined by the conversational tone; in (2b1), the story involves higher-order attitudes, although only one iteration—not the infinitely iterated attitudes mentioned in

[9]I don't take either of *these* arguments to be decisive against the identity theory. But we'll see in a moment that the Identity theory is unattractive on Stalnaker's own view of the common ground. Later, in Section 5, we'll see that the minimal theory also has trouble given this view of the nature of acceptance. So there is some reason to doubt the prospects of this theory.

the introduction. (The precise strategy for avoiding iteration in (2b1) is somewhat subtle; later we will spend more time understanding its mechanics.)

But both of these fairly conservative theories are similar to the Identity theory, as far as the second kind of example is concerned: neither offers a simple solution to the story about the hostess's toothpaste. In the example, it's plausible that all parties may not just believe that the toothpaste is there, staring them in the face, they may also be attending to it, however discreetly. Similarly, everyone may not just believe that the toothpaste is there, but believe that all others believe it is there, as well. In any case, each of these three theories may advocate the plausible position that the phenomenon should be explained in terms of general social norms, independent from the norms of conversation.

(2b2) sits at a threshold between the more conservative theories above it, and the more radical ones which follow. On the one hand, it uses only the familiar attitude of belief, but on the other it invokes a new, primitive property. A theorist of this kind must offer us some explanations of this new property, and in particular some explanation for why it is not to be understood in terms of others' attitudes. As so often, this explanation need not be particularly informative: it may be that the logic of the theory of acceptance provides the best explication of what it is for a proposition to be "on the table", and that we cannot hope for anything more informative.[10]

The final two theories of acceptance (2c) and (3) do not suffer from the putative difficulties about the unity of conversation or the presence of toothpaste, because they may take the attitude of acceptance to have whatever properties they like. This flexibility, however, comes at a fairly severe cost: of introducing a previously undiscovered attitude. According to these two theories, acceptance is supposed to be a very important attitude, which pervades our social lives, and yet somehow we do not have a word for this attitude. One way of responding to this problem is to identify acceptance with an attitude for which we do have a name. For example, one might think that presupposition is an attitude which cannot be explained in terms of others. One might then replace "accepts" in my statements with "presupposes".[11] But in any case, the "Distinctness" theory, (2c), suffers from this problem less severely than (3), the "Primitive" theory, since (2c) can explain the absence of a word for this attitude by the fact that our acceptances are always masked by the

[10]In this paper, I won't define how propositional quantification would work officially, in a defined object language. This is mainly because I won't introduce a language or logic at all. Unofficially, I will be able to quantify over propositions since propositions are just sets of worlds. In any case, it should be clear that formalizing this theory by introducing propositional quantifiers and constants would be easily done.
[11]For the relationship of this view to Seth Yalcin's theory of the common ground, see below, n. 15.

attitude determined by the conversational tone. Once again, however, this explanation comes at a certain cost: the Distinctness theory, (2c), owes us an explanation for why these different, distinct attitudes are always found in pairs. Is this a matter of metaphysical necessity, or a contingent feature of our psychology, or some other, alternative form of connection?

The Primitive theory, (3), has a further interesting feature. Moore sentences, such as "it's raining but I don't believe it is", are widely agreed to be unassertable in all contexts. A natural explanation, available to all theorists of the first and the second kind, is that in contexts where one asserts a sentence, if one has the attitude determined by the conversational tone toward a proposition then one also believes that proposition. These sentences are then unassertable because they cannot be truly believed. (Other sentences, involving knowledge or certainty, might force stronger views of the attitude determined by the conversational tone, but assuming that true belief is required will be enough for our purposes in this paragraph.) But the theorist of our third kind—of primitive acceptance—cannot straightforwardly adopt this kind of explanation. He or she must explain the unassertability of Moore-sentences by different means—perhaps by a separate norm governing assertion: for example, that one should assert a proposition only if one believes it. This theory would separate the norms for assertion from the norm that one should converse in such a way that one's contributions have the appropriate relationship to the common ground. This in itself is not a problem for the theory, it simply marks a difference between the primitive theory and the earlier theories which place emphasis on the notion of the conversational tone.

These different theories of acceptance will be useful in later sections—especially Section 6 on "epistemic sentences"—when we come to consider different styles of explaining important conversational phenomena. As I've said, I will not attempt to decide between these theories here, but merely to understand their different properties in the context of the minimal theory of common ground.

2.2.2. Stalnaker's Theory. Stalnaker has offered a variety of different theories of the common ground, which it will be useful to have before us for the sake of comparison.

Recall that, where ϕ is an arbitrary propositional attitude, a group mutually ϕs (or: mutually ϕ^1s that p) if all ϕ that p, and mutually ϕ^ns that p if they mutually ϕ^1 that they mutually ϕ^{n-1} that p. They then commonly ϕ that p just in case, for all natural

numbers n, they mutually ϕ^n that p. Common belief, common knowledge, and common acceptance are all defined in this way, substituting "believe", "know" and "accept" for ϕ.

Stalnaker's first theory takes acceptance to be determined by conversational tone, as above in (1), and then defines

(CB-A): It's S_1-common ground that p if and only if the parties commonly believe that they accept that p. (1998, 2002)

One could also use knowledge in place of belief:

(CK-A): It's S_2-common ground that p if and only if the parties commonly know that they accept that p.

(This definition could be seen as derived from Yalcin 2007, although see below.)

More recently, Stalnaker uses acceptance, apparently as a primitive attitude, although he does not give a detailed account of its nature:

(CA): It's S_3-common ground that p if and only if the parties commonly accept that p, (2014)

where acceptance is a primitive attitude of type (3). (When I don't want to disambiguate between the S_1, S_2 and S_3 sentential operators, I'll simply write "it's S-common ground that".)

Each of the first two of these theories could be developed in tandem with any of the theories of acceptance described in the previous section. The proponent of (CA), however, cannot adopt either of the theories of acceptance (1) or (2). The reason is fairly easy to see. If I ask you to suppose that we do not exist, then it may become common ground that we do not exist. But if our supposition is to be consistent, we cannot then also be supposing that we are supposing that we do not exist, for that would be a way of supposing that, after all, we do exist. Since it cannot coherently be commonly supposed that we do not exist, one must, on this theory of acceptance, be able to accept a proposition without bearing the attitude determined by the conversational tone to it. (Rejecting closure for supposition is not a real way out of this problem, since while our suppositions certainly often fail to be logically closed, we *might* engage in a conversation where we were aware of the conflict in supposing that we do not exist and supposing that we are supposing that we do not exist.)

This is a deep problem for related theories of the common ground. To see this, consider the proposal of Seth Yalcin 2007. Yalcin takes "presuppose" as a primitive, and holds that it validates the axiom: S presupposes that p if and only if S presupposes that

the others presuppose that p.[12] Yalcin then defines common ground as the propositions that subjects commonly know that they presuppose, making his theory at least formally analogous to CK-A. He does not articulate the claim that one presupposes that p only if one bears the attitude determined by the conversational tone to p, but he does say that the conversational tone determines the attitudes one should bear to the propositions in the common ground. When we coordinate on a conversational tone, we commonly know that this is the attitude we should bear to the propositions in the common ground. But then, at least on a Stalnakerian, coarse-grained view of content, the various iteration principles Yalcin endorses lead him back into the problem just described. For suppose it's commonly known that: if p belongs to the common ground, I am supposing it. Then by definition of the common ground, it will be commonly known that, if it's commonly known that we all presuppose that p, I am supposing it. But because one presupposes that p only if one presupposes that others do, then if it's commonly known that all presuppose that p, it's also commonly known that all presuppose that all presuppose that p. So if it is common ground that p, it will be common ground that all presuppose that p. So in addition to supposing that p, I will have to suppose that all presuppose that p. But if I wish to suppose that we do not exist, then I am engaging in an incoherent supposition, for I am also supposing that all presuppose that they do not exist, and hence that they both do and do not exist.

Once again, my point is not to discard CA, Yalcin's theory or any other. I wish merely to point out that neither CA nor Yalcin's theory can endorse theories (1) or (2) about the relationship of acceptance (respectively, presupposition) to the attitude determined by the conversational tone. Later, we'll see that the minimal theory should also reject (1), although it may be some small point in favor of the minimal theory that it can accept theories of type (2), and thus is not forced to posit a new social attitude of acceptance or presupposition.

Before moving to the main argument, it will be useful to have Stalnaker's notion of the "context set" or *the* common ground. To state this idea, we use the framework of multi-agent Hintikka-Kripke models in epistemic logic, and in fact we will restrict attention to models which have a finite state space.[13] Each agent has three binary accessibility

[12]Presumably "others" includes oneself, so that one presupposes that p only if one presupposes that one presupposes that p, but what I say in the main text won't rely on this assumption.

[13]The Appendices to this chapter develop the full formal apparatus behind this paper. There, I use the more general setting of neighborhood models of belief and knowledge. The Appendices are designed so that they could be read on their own, or could be read, one at a time, as they are referred to by the main text. See now Appendix A.

relations, associated with the attitudes of acceptance, belief, and knowledge. We think entirely from the perspective of the model, without introducing a language or logic, and define belief, knowledge and acceptance by functions from propositions to propositions, that is, sets of worlds within a universe Ω. Thus for example, $B_i(E)$ takes the proposition that E to the proposition that i believes that E; a world w belongs to the proposition $B_i(E)$ just in case $R_i^B(w) \subseteq E$, where $R(w)$ is the set of worlds accessible from w.

Common knowledge, belief or acceptance is defined using the transitive closure of the union of the agents' accessibility relations, whether for acceptance, belief or knowledge. If this relation is denoted $R^*(w)$, we define (for example) the common belief function on propositions by $w \in CB(E)$ if and only if $R^*(w) \subseteq E$.

We use these definitions to define, at last, a common ground accessibility relation. For example, according to the minimal theory of common ground, the common ground accessibility relation is the union of the agents' accessibility relations for acceptance. According to **CA**, by contrast, the relevant relation is the transitive closure of the agents' accessibility relations for acceptance. We can then define:

Context Set: The context set or the common ground is the set of worlds reachable in one step by the common ground accessibility relation.

When discussing Stalnaker's theories, I will sometimes speak of the S-context set or the S-common ground to avoid confusion. It should also be emphasized that the noun phrase "the common ground" does not have the same denotation as the sentential operator "it's common ground that". In our simplified Kripke models, the two have the following relationship: a world belongs to the context set just in case it's not common ground that that world does not obtain. In more general settings, however, as in the next chapter, this relationship would not be preserved.

2.3. Presupposition and Accommodation

The next four sections will consider and respond to a series of purported arguments for Stalnaker's theory as opposed to the minimal one. In each case, I show that these arguments do not succeed. In responding to the arguments, I develop the mechanics of the minimal theory in more detail. I also show how the arguments which supposedly tell in favor of Stalnaker's theory lead to considerations which in fact tell against his theory and in favor of the minimal one.

The first argument for the iterative conception of the common ground is based on Stalnaker's theory of speaker presupposition. Stalnaker has proposed that a speaker S-presupposes that p if and only if the speaker believes (or, now, accepts) that it's S-common ground that p. If this elegant theory of speaker presupposition depends on Stalnaker's iterated theory of the common ground, then one might take this fact to support this theory of the common ground. Since the theory of speaker presupposition seems attractive in its own right, we should prefer a theory of common ground which allows us to explain speaker presupposition in this way.

My main strategy for defusing this argument will be to show that Stalnaker's theory of speaker presupposition does not in fact depend on Stalnaker's iterated theory of the common ground. But before turning to this main response, let me first make one point about Stalnaker's theory of speaker, since I think this point already diminishes the force of the proposed argument. The point is that Stalnaker's theory of *speaker* presupposition is at least not clearly a theory of the interesting phenomenon in the vicinity of presupposition. Speakers have a clear grasp of sentential presupposition, which is independent of their understanding of the attitudes of a specific speaker on a specific occasion. But the deep phenomena in the area seem to be facts about sentential presupposition—not speaker presupposition. So it is at best unclear how powerful an argument based on Stalnaker's theory of *speaker* presupposition could be.

While I think this is an important observation about Stalnaker's theory of presupposition, in the present context it is just an aside. We can respond to the proposed argument without settling the importance of speaker presuppositions or their potential relevance to the theory of sentential presupposition, because, as I will now argue, the minimal theory can exactly reproduce Stalnaker's theory of speaker presupposition.

Presupposition: A speaker S presupposes that p if and only if S accepts that p and believes that the others accept that p.

Sometimes, speakers can introduce new information by presupposing it. This phenomenon of presupposition *accommodation* is one of the main data which make it seem that the theory of speaker presupposition captures an important component of conversational behavior. So it is important to show that the minimal theory can also explain accommodation. And in fact, it can: the minimal theory can replicate Stalnaker's law governing the dynamics of presupposition accommodation.

Accommodation: If B believes that A presupposes that p, and B comes to accept that p, then B presupposes that p.[14]

Since Presupposition and Accommodation are the main principles of Stalnaker's theory of speaker presupposition, this completes the argument that the minimal theory can also deliver the main components of that theory. There are of course many further issues about presupposition. But I will not discuss them here. It suffices for our purposes to see that the core of Stalnaker's theory of presupposition can be replicated by the minimal theory. The defects of Stalnaker's theory of speaker presupposition will also be defects of the new theory, but so will its virtues. There is thus no argument based on speaker presuppositions in favor of the iterated theory of common ground as opposed to the minimal theory.

There are two principles which my theory of presupposition does not deliver, but which Stalnaker's CA theory, the theory of S_3-common ground does. Most important is a principle of "social" introspection. Discussing presupposition, Stalnaker writes: "What is most distinctive about this propositional attitude is that it is a social or public attitude: one presupposes that φ only if one presupposes that others presuppose it as well." (2002: 701) Following Stalnaker, Seth Yalcin agrees that "one presupposes that p only if one presupposes that one's interlocutors presuppose that p" (Yalcin 2007: 1007). Related to this principle of "social" introspection is the principle of "positive introspection": which Stalnaker want the logic of presupposition to obey: "if Alice presupposes that ϕ, then she presupposes that she presupposes that ϕ" (Stalnaker 2002: 708).

Now in general, where a is some attitude, if a group commonly a's that p, they commonly a that they commonly a that p. This "positive introspection" principle simply follows by definition: a group commonly a's that p if and only if, for all natural numbers n they mutually a^n that p. But then it's clear that for all natural numbers n, they mutually a^n that they mutually a^n that p. This is why Stalnaker's (2014) theory, the CA theory delivers both of these principles.[15]

[14]See, e.g. Stalnaker's 2002; the idea stretches back to his 1974, and has been developed in a number of places. This is as it should be, since one of the aims is to show that the minimal theory can account for the principles he holds are crucial to the notion of common ground.

[15]The emphasis on the "social introspection" and "positive introspection" principles in Stalnaker's (2002), however, is somewhat surprising, since the theory of presupposition in that paper invalidates both of these supposedly crucial principles. There, it is understood that belief is not sufficient for conversational acceptance: there may be settings in which the conversational tone determines attitudes other than belief or knowledge as the attitude appropriate to the conversation. Presupposition is defined as above: we say that one S_1-presupposes that p if and only if one believes that it's S_1-common ground that p. So if one S_1-presupposes that p, it follows that one *believes* that others S_1 presuppose that p, but since one need not *accept* this, nor believe that others accept it, it does not follow that one S_1 presupposes that one S_1-presupposes that p, nor that one S_1 presupposes that others S_1-presuppose that p. As we've seen above Yalcin 2007 takes presupposition as a primitive attitude, so he's not subject to this particular

It may be helpful to see why the minimal theory fails to validate this principle, in the subtlest case, the theory of acceptance (2b). Recall that according to this position, to accept a proposition is in part to believe that it is on the table, where for a proposition to be on the table is for all participants to adopt the attitude determined by the conversational tone to that proposition. Consider a conversation in which the conversational tone determines the attitude of belief. According to this view, it's common ground that p in this conversation just in case the participants mutually believe that p, and mutually believe^2 p. (Note that the proposition that all believe that p may not belong to the common ground, since even though all believe that all believe that p, it may not be the case that all believe that all believe that all believe it.)

Now suppose in this same context that a speaker s presupposes that p: that is, s believes that p, s believes that the participants mutually believe that p , and s believes that the participants mutually believe^2 that p. To presuppose that p, s need not believe that the participants mutually believe^3 that p. But if s presupposes that all presuppose that p, s would have to believe that the participants mutually believe^n that p, for $3 \leqslant n \leqslant 6$. So the question whether a speaker presupposes that p and the question whether a speaker presupposes that others presuppose that p are simply independent questions, which may receive different answers.

But if anything the failure of the minimal theory to validate this principle seems an advantage, not a disadvantage. Suppose in a context where the conversational tone determines the attitude of belief, I tell you "I can't come to the meeting, I have to pick up my sister from the airport." With all the qualifications already mentioned about the notion of speaker presupposition, it's plausible that I presuppose in the relevant sense that I have a sister. But do I presuppose that you believe that I believe that you believe that I have a sister? Utterances involving this kind of iteration are hard to process, so giving a concrete example of an *obviously* infelicitous sentence of this kind does not seem possible. (At any rate, I haven't been able to come up with an example.) But the putative presupposition of the speaker in this example is odd. This is especially clear if we consider again the relationship between speaker presupposition and sentential presupposition. The sentential presuppositions of the sentence "I have to pick up my sister from the airport" do not make reference to the attitudes of others at all. One might be attracted to the schema: if a sentence s presupposes (sententially) that p, a speaker utters s felicitously only if

problem, although, as we'll discuss in the next section, he still does not validate positive introspection for the common ground itself.

the speaker presupposes that p. But this schema cannot be used to explain the bizarre *speaker* presupposition in our example. For the sentence has no sentential presuppositions whatsoever about the attitudes of the participants in any conversation.

In fact, there is some reason to suspect that Stalnaker and Yalcin's endorsement of the iteration principle for presupposition stems from a misunderstanding of a much more obviously desirable principle. Consider the following "maxim" presented as an imperative, following Grice:

Aim-P: Presuppose that p if and only if your interlocutors presuppose that p![16]

An important goal of many conversations is to coordinate on a body of information. We do this in part by attempting to bring our presuppositions in line with each other. Presuppositions about others' presuppositions are not important to achieving this aim of coordination. What is important is that our presuppositions in fact agree. One very good way to ensure this kind of agreement is to know what others presuppose, and bring one's presuppositions into accord with what one knows they presuppose. But even in this good case, it is knowledge of others' presuppositions, not presuppositions about them, which play the role of ensuring coordination on the relevant body of information. In any case, one might reasonably hold that the activity of coordinating on this body of information is rarely if ever as sophisticated as even the knowledge-based picture suggests. Much more usually, we adjust our presuppositions, one step at a time without thinking about it. We try a variety of alternatives, perhaps in parallel, until we find a hypothesis about others which seems to work in processing the conversation we are having. Sometimes, these adjustments involve beliefs about the attitudes of others, but perhaps just as often they involve adjustments in our beliefs about the world. In this latter case, we do not even need beliefs or knowledge about others' presuppositions: we coordinate on a body of information by presupposing only what is true.

For some of our theories of acceptance, the maxim Aim-P can be explained by a second, semi-Gricean maxim, governing acceptance:

Aim-A: Accept that p if and only if your interlocutors accept that p!

Typically, when we move around the world, we update our beliefs directly by acquiring new information. According to the minimal theory, if Aim-A is satisfied in a given context, we can follow a similar practice in conversation. Since in this case, for each speaker s, it's common ground that p if and only if s accepts that p, s doesn't need to think about

[16]This is related to Stalnaker's notions of "non-defective" and "equilibrium" contexts. See the appendix to the next chapter for more detailed discussion of these notions.

the common ground to decide what has been said or to update his own beliefs and acceptances. Instead, he must simply refer to what he himself accepts.

This "Gricean" maxim for acceptance is not available on the "Identity" theory of acceptance, (1). Aim-A will sometimes conflict with the aims of other attitudes, for example, belief. But the other theories of acceptance can endorse this maxim, perhaps claiming it as an aim of the special attitude of acceptance. This approach is perhaps most natural for the theories which take acceptance as primitive or as defined in terms of a new primitive, that is, theories (2b2), (2c) and (3). According to (2b2), for example, one should believe that a proposition is on the table if and only if the others do, too. For each of these theories, Aim-A could state a special aim of acceptance, just as truth or knowledge is the aim of belief.

2.4. Informativeness and Positive Introspection

My main aim is to show how the minimal theory accounts for some basic conversational phenomena, as a way of responding to various objections to this minimal theory. Sections 5 and 6 will continue to pursue this project. But in this section, we take a brief digression to consider some differences between Stalnaker's "iterated" theory and the minimal one I have been proposing. These differences have been thought to provide the basis for an argument against the minimal theory, so although this discussion will be a digression from the project of explaining conversational behavior, we are still dealing with a topic which is central to the goals of this paper.

In Stalnaker 2014, we find a number of places where Stalnaker offers the beginnings of an argument for the claim that the common ground has the structure of common belief:

> When one speaks, one presupposes (takes it to be common ground) that one is speaking, and this means that the conversation is taking place, not only in the actual situation, but in each of the possible situations in the context set. This fact is reflected in the formal representation of the common ground by the iterative structure. (2014: 92)

This passage, from Chapter 7, comes closer to a direct argument:

> When we are engaged in conversation, it is common ground that we are engaged in that particular conversation. This is crucial for the role of common ground in constraining the means used to communicate, and

> is built into the iterative structure of common ground. (The iterative structure implies that information presupposed includes information about what is being presupposed in the conversation in question.) (2014: 178)

It is important to distinguish the claim Stalnaker is making here from something that sounds similar, but which he cannot be saying. When people are speaking they tend to believe that they are. Presumably all of the theories of acceptance on offer above would agree that people do not just believe that they are speaking, they accept this, too. But if all parties accept that Bob is speaking, for example, then on the minimal conception of the common ground, too, it will be common ground that Bob is speaking. "In each of the possibilities in the context set", this act will be taking place. Similarly, since we do tend to accept propositions about others' beliefs in the context of a conversation (and about what others accept), our presuppositions will typically include "information about what is being presupposed in the conversation in question." The fact that "we are engaged in that particular conversation" is typically part of the common ground in the minimal sense of common ground, just as much as it is in Stalnaker's.

Instead Stalnaker is, I think, adverting to the fact that his theory S_3 delivers the principle of presupposition discussed in the previous section: that if one presupposes that p, one presupposes that all others also presuppose that p. In fact, as we've discussed, S_3 also validates the principle that if it's common ground that p, it's common ground that it's common ground that p. The minimal theory does not validate this principle. (For reasons mentioned earlier, the theories S_1 and S_2 do not deliver the relevant principle for presupposition *or* for common ground. Even on Yalcin's version of S_2, where presupposition is taken as a primitive, and assumed to be "socially" introspective, we are not guaranteed to satisfy the principle that if it's common ground that p, it's common ground that it's common ground that p, since participants may commonly know that p but nevertheless fail to presuppose that they commonly know that p.)

From a formal perspective, the fact that Stalnaker's latest theory validates this "positive introspection" axiom for common ground has what might seem to be important consequences. If we think again in terms of our Hintikka-Kripke models of belief, knowledge and acceptance, the positive introspection axiom has the consequence that from any point within the context set, the context set from that point will be a subset of

the actual one.[17] In this sense, as Stalnaker says, positive introspection for the common ground ensures that the common ground "constrain[s] the means used to communicate".

But Stalnaker buys this particular formal constraint at the cost of preventing the common ground from constraining communication in much more important ways. On his view, the common ground is constrained with respect to what it "thinks" the common ground is. But it is much less constrained with respect to many other matters of importance, which do not concern the common ground. To see this, consider the following example.

Louis comes from Luxembourg. I overheard him saying he's from Luxembourg the other day, so I know he's Luxembourgish. Louis slyly saw me listening to his conversation when Luxembourg came up; he knows I know he's from Luxembourg. But I saw him looking at me only later, when he was talking about the Netherlands, so I think he thinks I think that, in spite of his name, he's Dutch. Louis saw me see him see me watching him only when the conversation turned to France, so he thinks I think he thinks I think he's from France.

Now if Louis and I find ourselves in a conversation where the attitude determined by the conversational tone is belief, the minimal theory predicts a much more informative common ground than Stalnaker's theory does. Here, it's S-common ground that Louis is from Luxembourg or the Netherlands or France, and that's it. (The example could of course be extended so that we *commonly* believe nothing about Louis' origins. It would thus only be S-common ground that he's from some country.) But according to the minimal theory, it's common ground that Louis is from Luxembourg. In fact, if we take the simplest version of the minimal theory, where to accept a proposition is to take the attitude determined by the conversational tone to that proposition, then it will also be common ground that it's common ground that Louis is from Luxembourg. (It won't, however, be common ground that it's common ground that it's common ground he's from Luxembourg.) So while S-common ground does ensure that facts about the common ground will belong to the common ground, it does so at the cost of the common ground being much less informative than it might be.

The point is not specific to this example: it's easy to see that in *any* situation, if it's S-common ground that p, it's common ground (in my sense) that p, while the converse doesn't hold in general. If facts about the world *are* commonly believed, then they're

[17]The point is subtler in the neighborhood frames used in the Appendices, since the definition of the context set does not have such a simple relationship to the operator "it's common ground that".

also believed, so if the conversational tone determines the attitude of belief, they will be common ground in my sense, just as much as in Stalnaker's sense. It's just that the minimal theory of common ground says that much more is common ground, since we may mutually believe many propositions we fail to commonly believe.

In fact, this kind of example allows us to add a new motivation for seeking an alternative to Stalnaker's theory of common ground. This new example will also help us to develop a more concrete sense for the difference between the two theories.

Suppose that in the serious context described above, where one should accept a proposition only if one believes it, I say, "Wow, Louis, you must be very glad you're going home soon!" He replies "Yes, my country is very beautiful."

What does Louis say? In a context like this one, it's S-common ground that p if and only if it's common belief that p (or common knowledge that we believe that p). The most informative proposition of this kind, at least as far as Louis' origins are concerned, is that he's from Luxembourg, France or Holland.

Our example will now concern Stalnaker's theory of assertion. Recall that the diagonal proposition expressed by a sentence contains exactly those worlds w such that the semantic content of the sentence *as uttered at* w contains w itself. According to Stalnaker when phrases such as "your country" are uttered in a context where it's not common ground which country that is, the semantic content of the sentence cannot be asserted. Instead, we take the asserted content to be the diagonal proposition. Thus in the situation just described, Stalnaker would hold that Louis asserts the diagonal proposition, the proposition that if he's from Luxembourg, it's beautiful and if he's from the Netherlands, it's beautiful and if he's from France it's beautiful. Louis doesn't say very much.

On my view, by contrast, it is common ground that Louis comes from Luxembourg. So we don't need to consider the diagonal proposition: Louis asserts the semantic content of the sentence. He says that that Luxembourg is beautiful. Since the context is much more informative, the content of what is said is much more informative, too.

As with our first two examples considered in the introduction, this example is hardly an argument for my view as opposed to Stalnaker's. It's hard to disentangle our judgements about the semantic value of Louis's assertion from what we take the content of his assertion to be. And even if one agrees with my judgement about what Louis says in this context, one might respond to the example by rejecting the use of diagonalization, while holding fast to Stalnaker's theory of common ground. But still, the example adds to the

growing body of strange predictions made by the "iterative" conception of the common ground. My theory of common ground, by contrast, guarantees what seems the right result: Louis successfully communicates the proposition that Luxembourg is beautiful, and not the far weaker claim that, if he's from a country, it's beautiful.

2.5. Update

We now end the digression and return to the project of demonstrating how the minimal theory of common ground can explain important conversational phenomena. In this section I consider perhaps the most important desideratum for a theory of common ground: that it be able to explain how the common ground alters in the course of the conversation. In the next section, I will then turn to a final question: how to explain the felicity and infelicity of certain patterns involving epistemic vocabulary.

An influential cluster of arguments for the "iterative" conception of the common ground concerns update, and, in particular, the "transparency" of update. In this section, I will consider two such arguments, and show that they are not really arguments for the iterative conception of the common ground at all. I will do this by demonstrating that these demands can be met within the minimal theory just as easily as within the iterative one.

The first argument takes the perspective of the speaker. It begins from the observation that linguistic communication involves intentional action. It is then claimed that a necessary condition for intending to perform an action is that one believes one can perform it, if one chooses to do so. In the context of communication, it is urged, if speakers are to be capable of the intentional action of, say, assertion, there must be some effect on context such that they believe their utterances will have that effect.

A second line of thought takes the perspective of the listener, and is more direct. Instead of beginning with abstract claims about intentional action, it begins from the claim that in ordinary circumstances we are not in fact at a loss as to how to update context on hearing others' utterances. We do not generally find ourselves in a state of confusion when someone makes an assertion or asks a question. If we do find ourselves confused, we count the utterances "defective", and our theory should aim to predict their defectiveness.

These demands become stronger if they are paired with specific commitments about how the "effect" of an utterance is calculated from the semantic value of a sentence together with features of the common ground itself. Stalnaker holds that it's a norm of

rational communication that one perform an utterance u in a context c only if, for all w in c, what is said by u in c is the same as what is said by u in c_w. The only way to guarantee this identity within Stalnaker's theory by ensuring that, for all w in c, $c = c_w$. I will be arguing that *this* demand for the relationship of *contexts* to their "candidate contexts" is much stronger than what is required to make sense of the possibility of updating contexts.[18]

2.5.1. Intentional Action? Our first task is to think more carefully about intentional action.

Consider the following game. There is a row of equally marvelous and indestructible prizes arrayed on a shelf. One wins a prize just in case one manages to knock it off the shelf. To do so, one must operate a mysterious machine, by pressing one of two buttons, the green or the red. The mysterious machine has a number of states, corresponding to the marvelous prizes. If the machine is in state 1, and one presses the green button, then the machine will knock off the first prize. If it is in state 2, it knocks off the second prize, and so on for all of the prizes. Players of the game are given no information about the state of this mysterious machine. But they know that the red button, regardless of what the state of the machine is, knocks off no prizes, whereas the green, regardless of the state of the machine, will knock off a prize.

Players of this game may be unable to coherently form certain intentions. For example, because of their limited information about the state of the machine, it seems plausible that there is no prize such that they can coherently intend to knock off that prize. They may want to knock off one rather than another, but they are uncertain of the outcome of their actions, and so cannot have intentions about which one they will knock off. On the other hand, they can have other intentions: they can intend, in pressing the green button, to knock off some prize.

I propose that conversation is not so different from this mysterious machine. We can have the appropriate kind of intentions about what we are saying, even if we do not

[18]The fact that Stalnaker's theory of assertion requires negative introspection for the common ground was observed by Hawthorne and Magidor 2009. Appendix 2.B gives more detailed discussion, correcting one important formal error in their argument. It may be worth remarking here on one feature of the discussion inspired by their paper (Stalnaker 2009, Almotahari and Glick 2010, Hawthorne and Magidor 2011). A casual reader of that debate would be forgiven for thinking that the central issue between those authors concerns introspection principles for individual attitudes such as knowledge and belief. But the problem about access to asserted content is independent of this debate (as Hawthorne and Magidor recognized (2009: 387)). Even granting that belief and acceptance satisfy positive and negative introspection, or that knowledge satisfies positive introspection, *common belief* and *common knowledge* still won't satisfy negative introspection. Stalnaker needs further assumptions to achieve that result. But also, even if one denies that individual attitudes satisfy positive introspection, common belief and common knowledge *will* satisfy positive introspection, as a matter of logic. So the real issue concerns negative introspection for the common ground, and this issue is discussed in Appendices 2.C and 2.D.

always know enough to know what precise effect our utterances will have on the common ground.

The idea will be that there is a set or "cloud" of contexts in play at any moment of a conversation.[19] We may say different things in each of these contexts. But the typical situation is that the differences in each context are not great enough or are not of the right kind to impede our choice of the "gist" of our utterances. (When the differences are great or important, an utterance may be infelicitous, as I'll discuss below.) Our capacity for intentional action in communication should be assessed against our capacity to choose this "gist" of our utterances, not our capacity to choose the precise effect of what we say. Communication is more like the pursuit of what Rawls called an "overlapping consensus" as opposed to a pursuit of precise agreement. We attempt to select our utterances, and interpret others as selecting utterances, so that the differences in what our utterances express according to the different contexts won't matter to the main points we are making.

But where does this cloud come from? It's time to make these vague ideas a bit more precise.

2.5.2. Minimal Transparency. Part of being a competent speaker is knowing how to compute a new context set from two inputs: the old context set and an utterance. The following principle makes official an idealized version of this idea:

Transparency: For any competent speaker i, context c, and utterance u, if c updated on u is c', then i knows that c updated on u is c'.

Transparency is only an idealized principle because it's implausible that we need to have settled dispositions about *every* utterance and context to be competent speakers. Moreover, for many utterances, there can be room for reasonable disagreement about the effects of expressions on different contexts; the meaning of expressions is not as transparent as Transparency suggests. But this kind of semantic uncertainty affects all theories which claim that a successful assertion requires that one know what one is saying. For example, on Stalnaker's theory of assertion, too, if one does not know how to compute the semantic value of an utterance at a world, one will be unable to diagonalize. If we assume Transparency, we can abstract from this kind of semantic uncertainty, to focus on issues related to the common ground itself.

[19] I'm borrowing the word "cloud" from Kai von Fintel and Anthony Gillies (see, e.g. 2011 118-119). They borrow it, in turn, from Kratzer. I suppose it ultimately dates back to very natural picture-thinking about Lewis's view in Lewis 1975, or to early discussions of supervaluationism.

Now, given this principle, we want to understand the origins of the cloud of contexts. The following definition introduces two natural notions of "candidate" contexts:

Candidate Context: A proposition c_w is an *i-candidate context* or a *candidate context for i* just in case i does not believe a proposition which entails that c_w is not the context set.

A proposition c_w is a *candidate context* if it's not common ground that p, for any p which entails that c_w is not the context set.[20]

It is worth emphasizing that the definition of an i-candidate context uses *belief*, not acceptance. What's important is not that i *accept* that the context has a certain form, but what i *in propria persona* thinks the context might be. In general, when we consider the agent's individual view of the situation, it is no longer appropriate to consider what the agent *accepts* about what the common ground is, but rather what the agent *believes* it is.

Transparency has the consequence that, for any *set* of contexts, the agent "knows how to update" each of its members. In particular, the agent will know how to update all of the candidate contexts, and so will correctly map the cloud of contexts to a new cloud, in which each context has been updated in the appropriate way.

But in fact, once we have the idea of candidate contexts, we can relax the idealization of Transparency. Instead, we can define a class of circumstances in which i will understand what is said:

Minimal Transparency: An utterance u is *transparent* for an agent i if, for any i-candidate context c, if c updated on u is c', then i knows that c updated on u is c'.

u is *ultra-transparent* for i if, for any candidate context c, if c updated on u is c', then i knows that c updated on u is c'.

[20]The use of the "entails" clause is needed for them to be usable in the neighborhood frames in the Appendices. In Kripke frames, we could have just written: "A proposition c_w is an *i-candidate context* or a *candidate context for i* just in case i does not believe that c_w is not the context set," and made an analogous change to the definition of a candidate context. Note also that I could have written "c_w is the context", and made no reference to the specific world-based implementation of the idea of the common ground. Thus a context here could be a very different object, with much more structure than just a set of possible worlds.

In other words, i does not need to know this for every utterance whatsoever, but when utterances succeed (or succeed as far as i is concerned), it is because i knows what to do in the relevant "candidate" contexts.

With these definitions out of the way, we can come to our real task: to make the earlier analogy to the game with marvelous prizes more precise. Say that an agent i *i-globally asserts* that p if, for every i-candidate context c, if c were the context, she would assert that p. In other words, one globally asserts that p if, when we "supervaluate" over candidate contexts, the contexts agree that one asserts that p. My proposal is then that one may satisfy the requirements of intentional action by choosing an utterance which ensures that one *globally* asserts that p, even if one is uncertain which context set is the actual one. The content of the utterance may differ as regards other propositions in the different candidate contexts. There may be some i-candidate contexts in which one asserts p and q but not r, and others where one asserts p and r, but does not assert q. This kind of uncertainty affects one's ability to have coherent intentions regarding an assertion that q, but it doesn't affect the possibility of asserting that p, and intending to do so. Just as in the game described in the previous subsection, we may intend to perform one action even if our epistemic position is too weak for us to coherently have other, more specific intentions.

So even if one is uncertain about the context, it may be coherent to have the intention to assert that p. This of course does not mean that for any set of contexts C, it will be possible to assert that p, if one knows only that the context is some member of C. But it does demonstrate that the motivation from intentional action does not force us to the position that speakers always know, of some context c, that it is the context.

We still have one further duty to discharge: to show that the "clouds of contexts" model can offer a reasonable picture of update. I'll turn to that task first, and then close by considering some lingering objections.

2.5.3. Updating Sets of Contexts. In the clouds of contexts model, there are two forms of update. The first is to alter which contexts are candidates, whether simply i-candidates or candidates; the second is to update each i-candidate or candidate context according to what is said in that context.

Some updates can occur only if the common ground satisfies some preconditions. For example, on one theory of epistemic modals, "it might be that p" is felicitous only if there's a world in the context set where p. In the sets of contexts model, this translates to the

requirement that there be *some* i-candidate or simply candidate context which satisfies the precondition. Update now breaks into two cases. If no candidate context satisfies the precondition, the utterance is infelicitous, and the conversation breaks down. If some candidate context satisfies the precondition, we rule out those i-candidate or candidate contexts which fail to satisfy it, and update those which do.

In the second kind of update, the situation is equally simple. We update each candidate context by what would be said in that context were it actual. We map the old set of contexts into a new set of contexts according to the usual rules. We tend to identify what has been said by i with what has been said i-globally: what is said in every i-candidate context. This reflects the fact that we often identify what i said with what i meant to say: i's false beliefs about what the context is are sometimes more important in determining what i said, than what the context in fact was.

But identifying what has been said is of course more delicate and flexible than these rigid rules suggest. If, for some listener j, what is said in two different j-candidate contexts is *very* different, especially concerning questions of great import to the conversation, it forces j to choose between two ways of proceeding. If one of the two j-candidate contexts yields a much more plausible view of what has been said, then j may simply update her beliefs by ruling out the candidate which yielded the implausible verdict. This may be so, even if the context in question did not yield a strictly speaking infelicitous utterance. On other occasions, however, what is said in different candidate contexts may be equally plausible, so that j is forced to ask for clarification of the ambiguous utterance. (Or she may proceed even though she recognizes that there are two possible disambiguations of what i just said; she trusts that the issue will be resolved in due course.)

This discussion brings us, however, to a lingering problem. What I've said might seem plausible when we restrict to small numbers of candidate contexts. But won't the number of candidate contexts typically be very large? In that case, how can we possibly represent all the candidate contexts we need to represent? I'll close by considering this issue, and also a related issue about learning.

2.5.4. Representation and Learning. The minimal theory appears to impose huge representational demands on speakers and hearers, since they have to compute what would be said in a vast array of candidate contexts. Do we really have to think about all of these i-candidate contexts to understand what has been said in a conversation?

To dramatize the problem, suppose we've both read *Our Mutual Friend* and remember what happens at the end. But suppose that I don't know that you've read it, and you don't know that I have. Now suppose we're discussing the World Cup, so that the conversational tone determines (as it should) the very serious attitude of belief or knowledge. Now according to the first theory of acceptance (Identity), we're uncertain about what the context is. This problem doesn't just arise from Victorian literature: anything you might believe will be relevant to determining what the context in fact is. In such serious contexts, for example, there will be variability regarding my grandmother's birthday, the color of my favorite copy of my favorite book, and so on.

The minimal theorist can offer two different responses to this problem. First, he might note that the uncertainty here clearly does not impede our ability to communicate. We don't need to know *everything* about the context to update it. For example, suppose you respond to my comment about the German performance, by saying "They played very well last night". It would be absurd to hold that the value of this sentence depends on whether you've read *Our Mutual Friend* or not. So I can add the appropriate proposition to the common ground even though I'm uncertain what propositions exactly belong to the common ground. Since most utterances are sensitive only to a select range of features of the common ground, listeners don't have to know what the common ground is exactly in order to update it according to these utterances. In this sense, our beliefs about context are similar to our beliefs about many things: we can get by just fine even if we represent contexts only partially. We don't have to think about different possible contexts at a conscious or representational level in order to update them. We simply update the context which we represent partially, by adding new facts to our partial representation.

A second way of responding is to hold that in the example above, it's not in fact common ground that we've read *Our Mutual Friend*. This may be so either because we're not attending to this proposition (2a), or it's not on the table (2b), or we're not accepting it for some other reason (2c) and (3). Each of these theories could lead to a view on which it's not common ground that we've read *Our Mutual Friend*, and so, to a theory according to which we're not in fact uncertain about the context in the situation above.

Both of these responses are, I think, both powerful and plausible. But the discussion brings us to a second objection, about learning. Both Transparency and Minimal Transparency make use of speakers' and listeners' ability to compute updates on contexts "pointwise". But how did we acquire this ability? If we are generally uncertain about what

the context is, then we will learn to update sets of contexts, not individual contexts. So it seems that the minimal theory of update undermines itself.

My response to this problem will depend on a preliminary observation. Throughout the preceding discussion I have been slowly moving the reader to thinking in terms of i-candidate contexts as opposed to candidate contexts more generally. The reason for this shift should be clear, but it is nonetheless an important point. If i thinks there is only one possible context, and j's utterance is felicitous in that context, then even if (owing to some of k's beliefs, for example) j's utterance was infelicitous, i won't have any trouble updating on it. To think in terms of candidate contexts is to take an objectionably "context-eye" view of the conversation. The conversation is not about what the context "thinks", it is about what the participants think.

This observation is important because it shows that what is required to meet the demands of learning is not that the *context* know what the context is, but that each agent know what the context is. Stalnaker holds that it's crucial that the same thing be asserted at every point in the context set. But it should now be clear that this demand is too strong. It's much more important that, for each individual i, the same thing be asserted in every context i thinks might be the actual one.

This suggests the answer to our problem. All we need to give a theory for how agents learn how to update individual contexts is a set of cases in which individual agents know what the context is. If a child encounters situations in which she knows, of some c, that c is the context, and if later behavior confirms to her that she acted appropriately when she updated it as she did, then she will learn how to update individual contexts. It needn't *always* be the case that speakers and hearers know, of some set c that it is the context, for them to learn how to update contexts appropriately.

This response has a different character depending on our theory of acceptance. As noted before, the theories (2) and (3) of acceptance may hold that contexts are not very complex objects. There may be comparatively few propositions in a given context, since we may only accept relatively few propositions. So it can be fairly easy to know, on these theories, what the context is.

The "Identity" theorist has a harder time explaining the prevalence of contexts in which we "know what the context is". The most plausible version of this view is to hold that one only needs to know select facts about a context in order to know how to update it on a given utterance. Learning how to update "pointwise" does not then require that we know exactly what the context is in a very strong sense; we need only know some

features of the context. A child who knows that it was common ground that the Germans played a soccer game may learn how to update any context which satisfies this condition (even if he or she didn't know which context was the actual one). (This version of the Identity theory, however, is in a sense a notational variant of some version of theory (2) of acceptance. For why should we not, in the situation above, simply say that the common ground contains only the *relevant* propositions, and that the child did in fact know the exact context?)

I want to close now by noting one structural relationship between the simple-minded version of the identity theory and the iterative theory. Suppose we define a class of situations where it's common ground that p if and only if all agents know that it's common ground that p. These are the situations in which we "know what the context is". If we adopt the "Identity" theory (1) of acceptance, plus this simple-minded view of what it takes to know what the common ground is, then if these situations can be ones in which the conversational tone determines the attitude of knowledge, it will be common ground that p if and only if it's common knowledge that p. In these situations, in other words, the minimal theory will collapse into the iterative theory.

This is an interesting formal property, but, I think not an important conceptual fact. As I've argued, theorists of type (1) should hold that the contexts which are crucial for learning how to update are not ones where it's common ground that p if and only if participants know that it is. In these special contexts, they should hold that it's common ground that p *and* relevant that p if and only if participants know that it's common ground that p. As I've noted above there's a way in which this is a notational variant of a view such as (2b) where "on the table" is explicated by a primitive property of relevance or something like relevance. But notational variant or not, it seems to be the only hope for a minimal theory of this kind.

In any case, the formal collapse only occurs within the Identity theory (and given the simple-minded and implausible view of what it takes to know what the context is according to that theory). If in contexts where the conversational tone determines the attitude of knowledge, knowing a proposition is still not sufficient for accepting that proposition, then individual agents can know the common ground without the common ground having the iterative structure of common belief. More carefully: even in situations where the conversational tone determines knowledge, and it's common ground that p if and only if all agents know that it's common ground that p, it may still fail to be the case that if it's common ground that p, it's common ground that it is common ground that

p. Those who accept one of the theories (2) or (3) for acceptance may thus define a set of contexts where *all agents* know what the context is. These special contexts will still not have the structure of common belief.[21] Minimal theorists of this kind may hold that we learn how to update the common ground pointwise by engaging in conversational situations where we know what the common ground is. And they may do so without invoking the existence of any contexts where the relevant attitudes have the structure of common belief.

I've shown how the minimal theory can explain some important demands on the common ground. The minimal theory can still deliver a sense in which communication is a form of intentional action, since one may know the important effects one's utterance will have on the context without knowing what the context is. Moreover, listeners may know how to update the context even if they don't know which context is the actual one. I closed by responding to some putative problems about representation and learning. I showed that, appearances to the contrary, the minimal theory does not raise special problems about the difficulty of representing contexts or about learning how to update individual contexts. Finally, we saw how once we take an agent's-eye view of the conversation, there need be no collapse of the special contexts where we know what the common ground is into cases where the attitudes which determine what's common ground have the structure of common belief.

2.6. Epistemic Sentences

We now turn to our final topic: sentences which include epistemic vocabulary. I'll consider three example sentences, and how they are to be explained on each of the theories of common ground and update. My strategy here will be somewhat different from the strategy in previous sections. Instead of opting for a specific explanation of the phenomena, as I did with presupposition and update, I will aim to map the territory around these sentences. Our primary interest in these sentences or patterns is whether they can form the basis of an argument against the minimal theory of common ground. The short answer is that they cannot. There are two very plausible ways of explaining the sentences which force no alteration to the theory offered thus far. At the end of the responses, I'll note the ways in which the responses affect the KK principle, since some recent authors have attempted to use these epistemic sentences to argue for this principle.

The three sentences—or, properly, schemas—are:

[21]The formal argument for this point is given in Appendix 2.E.

(I) "p and I don't believe that p"; "p and I don't know p".

(II) "p and I might not believe that p."

(III) A pattern: The question "how ϕ?" presupposes that ϕ. But in most settings it's acceptable to challenge someone's assertion that p by asking: "how do you know that p?"

In general, the various views of (I) will be much the same, regardless of one's theory of the common ground. The other sentences are subtler, and this section will mainly be devoted to explaining these sentences.

2.6.1. Replies Based on the Logic of Acceptance. (I) We saw already that theory (3) of acceptance—that acceptance is a primitive attitude—can't explain Moore sentences via the logic of acceptance. For this subsection and the next, then, I won't consider theories of type (3).

On the other theories of acceptance, where speakers accept a proposition only if they also bear the attitude determined by the conversational tone to the proposition, the logic of acceptance itself explains (I). According to Stalnaker's theory of common ground, as well as to my own, one of the aims of assertion is to add a proposition to the common ground. If the conversational tone of a conversation where one makes assertions is knowledge, then both sentences are explained immediately. Since one accepts a proposition only if one knows it, then one can accept the proposition that p in an assertional context only if one knows that p, and so it cannot be the case that in that context one also doesn't know that p.

If (II) and (III) are to be explained in terms of acceptance, however, we need some elaboration of the logic of acceptance. The obvious principle is:

KA: An agent accepts that p only if she knows that she accepts that p.

(II) Let's first see how this principle explains (II). Suppose that the conversational tone for a given conversation is knowledge, so that one accepts that p only if one knows it, according to theories (1) and (2) of acceptance. Next, by KA, one accepts a proposition only if one knows that one accepts it. Now, we assume that agents also know that if they accept that p in such a context, they know that p. If they draw the inference on the basis of what they know, it follows from **KA** that they accept that p in such a context only if they know that they know it. We are now very close to the conclusion. To assert that p, one must accept it in a way appropriate to assertion. So then by **KA**, one must know

that one knows it. So one cannot both assert that p and "I might not know that p", since the latter entails that one does not know that one knows that p.

(III) Does this principle also explain the felicity of sentences such as (III)? The question "how do you know?" stops a conversation in its tracks. The usual way of understanding the phenomenology of this "arrest" is to take it to be the same as that of accommodation. Let me defend this claim briefly.[22]

One reason for thinking that accommodation is involved in this utterance derives from the "hey wait a minute" test. Suppose that after I assert that p, you ask me: "how do you know?" Now I can fairly easily say, "I didn't say that I know that p, I just said that p". What usually happens in this circumstance is that we reach an impasse. You have indicated that you do not accept my utterance, and I have not satisfactorily answered your challenge. But in making this concession, I have not retracted my assertion, and this is the crucial point. My assertion has been rendered inefficacious because of your attitude to the proposition. I will be forced to find other ways of making my point. But if I am stubborn, I don't have to concede that I shouldn't have made this assertion in the first place. So in making my initial assertion I did not presuppose that I knew p; your question introduced this presupposition and asked me to accommodate to include it in my own presuppositions. (Of course, there may be circumstances where I admit that I shouldn't have made the assertion. But this is usually where I concede that I don't know that p itself, not ones where I merely concede that I don't know whether I know p.)

Given this picture of accommodation, the principle **KA** also helps to explain **(III)**. After the assertion that p, all agents accept that p, and all know that they do. So while p is presupposed by all, it's not necessarily the case that it's also presupposed that the asserter knows that p. But since (according to **KA**), all presuppose that p, it's fairly easy to accommodate, and start presupposing that the asserter knows that p.[23]

---

[22]In fact, in asking questions of the form "how ϕ?" *speakers* presuppose that the answerer knows how ϕ, and thus knows that ϕ. Is this a problem for the explanation in the main text. I don't think it is. First, this speaker presupposition is surely not the kind involved in semantic presupposition. Second, this presupposition need not be accepted by the answerer. I say "It's going to rain." You say: "How do you know?" I shrug, "I dunno, look at the sky." I've clearly rejected your presupposition that I know how I know, and rejected the appropriateness of the *presupposition* that I know that I know (I need not be denying the truth of this claim; I just don't think it should be added to context). So while an assertion provides a basis on which a reasonable conversant can and often does subsequently presuppose that I know that I know, the assertion doesn't make that fact uncontroversially belong to the common ground. In a sense "how do you know?" makes a similarly contentious *presupposition* that I know (once again, the contentiousness of the presupposition needn't derive from a concern about the truth or falsity of the claim). But in this case, the situation is different: I can't reject the presupposition that I know without retracting the assertion. And this is why it seems crucial to have a theory which accounts for this latter datum.

[23]A different principle from **KA** is **AA**. But as I've noted before, this can't be true for the theories (1) and (2) of acceptance. If in a conversation where the conversational tone determines supposition, I

2.6.2. Replies Based on the Logic of the Common Ground. It's possible to explain these problematic sentences within the minimal theory of common ground, by expanding the logic of acceptance to include the **KA** principle. An alternative form of explanation is to expand the logic of the common ground, without expanding the logic of acceptance. Doing this requires giving up on the most basic form of the minimal theory, but it also does not force us to go all the way to the iterative conception of the common ground. (Recall that in this subsection, as in the previous, I won't be thinking about theories of acceptance of type (3).)

Instead of the the **KA** principle, we can use:

2-Common Ground: It's 2-common ground that p if all accept that p and all know that all accept that p.

Given that the conversational tone of an assertional context is knowledge, then the explanation goes as in the previous section. One asserts p appropriately only if one knows that one knows that p, and then the explanation is as before. The difference between this theory and ones based on **KA** is that 2-Common Ground tells us nothing about the psychology of acceptance. But this revised theory of common ground should be unattractive to those who endorse theories (1) or (2) of acceptance, for a reason I've noted often before. If we suppose that we don't exist, so that it becomes common ground that we do not exist, then on pain of having incoherent suppositions, we cannot also be supposing that we're supposing that we do not exist.

2.6.3. Replies Based on Speech Acts. A third response to these problem sentences is perhaps the most powerful. According to this response, we do not need to alter the logic of either acceptance or common ground. Instead, the response invokes general facts about speech acts, deriving from yet more general principles governing certain actions.

For certain actions, one must have a particular standing in order to perform the action. If one does not have this standing, performing the action is inadmissible or even impossible. Timothy Williamson (2013: 82-3) considers the example of a commanding officer. To issue a command to his soldiers, an officer must have certain standing as an officer. He cannot issue this command, while (publicly) questioning his own standing as a lieutenant. (Of course, he can have private doubts about his ability or right to command,

---

accept that we do not exist, then I will be supposing that we do not exist. But then I cannot also be supposing that we are supposing that we do not exist, on pain of engaging in an incoherent supposition.

but he cannot in one breath command, and also question his standing to do so, out loud to his underlings.)

In Williamson's view, assertion is no different. To assert that p one must know that p. Just as with the lieutenant's command, then, it is incoherent to assert that p and (publicly) question one's own standing to assert that p. But this is precisely the performance one engages in if one asserts that p and then says "I might not know that p"; one questions one's own standing to make the assertion. Williamson suggests that this is a very general feature of communication. In support of this view, he adds a further example, of someone who says: "How are you feeling today?—and I have no interest in your answer."

This response to our sentences does not impinge at all on the minimal theory of common ground, or on the theories of acceptance which we have considered. Instead, it invokes general laws about speech acts, and considers how they apply to the speech act of assertion. (The explanation of the question "how do you know?" proceeds as discussed in 2.6.1. An assertion licenses the inference that the asserter has the standing to perform the assertion. So it's easy to accommodate a respondent's new presupposition that the speaker has that standing.)

It's worth noting again that Stalnaker's own theory of the common ground must invoke an explanation of this form for Moore sentences in general. Since, as we've seen, Stalnaker can't require that one accepts that p only if one bears the attitude determined by the conversational tone to p, he cannot claim that it follows from the logic of acceptance that the serious context of assertion demands the attitude of belief to the components of the common ground. So Stalnaker must offer an explanation of even standard Moore-sentences by appealing to facts about speech acts in general, and assertion in particular. Similarly, when we come to the extended Moore-sentences, Stalnaker must also think in terms of general norms of speech acts, as Williamson suggests we should.

This is not a problem for either theory of these patterns. But it does mean that Moore-sentences can't be used to argue in favor of Stalnaker's most recent theory of the common ground as opposed to the minimal theory. Since Stalnaker himself must adopt a theory of Moore-sentences which explains them as infelicitous because of features of certain speech acts (independent of the common ground), his theory and a minimal theory which adopts this response will be on a par. There is thus no argument based on these sentences against the minimal theory.

2.6.4. The KK Principle: Summarizing the Discussion. Before concluding the whole discussion, I want to discuss briefly the relationship between these responses and the so-called KK principle. This is of interest because the data sentences in this section have recently been argued to pose a problem for a package of views endorsed by Williamson.[24] Williamson (2000) argues that knowledge is the norm of assertion. His arguments are directed at showing that *knowledge* that p, as opposed to (say) belief that p governs the practice of assertion. Williamson has also argued forcefully against the so-called "KK" principle, that if one knows a proposition one knows that one knows it.

The first point to note is that the explanation in 2.6.1 can be given within the minimal theory of common ground, and without invoking KK. Once again, this is a very important point which I think has been overlooked in the literature.

The explanation based on the KA principle is subtler. If we adopt the Identity theory of acceptance, then in a serious context where to accept a proposition is to know it, the KA principle will entail that if one accepts a proposition, one knows it, knows that one knows it, and so on. Now in general even this phenomenon does not lead to the return of the KK principle, which is a very general principle saying that *whenever* one knows a proposition, one knows that one does. But still the phenomenon would not be amenable to the spirit of Williamson's views since such an infinitely iterated hierarchy of knowledge would amount to a non-trivial "luminous" state.

But we've seen a number of reasons to doubt the Identity theory in any case. And if we adopt either theory (2) or (3) of acceptance, the KA principle will not imply the existence of such an infinitely iterated hierarchy. For since in a serious context knowing that p will be necessary for accepting that p (but not sufficient), then one may know that one accepts that p without accepting that one accepts it. Moreover, since the KA principle should be thought of as a norm, not as a fact about the psychological state of acceptance, there need be no requirement that acceptance itself is luminous.

Many will find this discussion of the possibility that one might fail to know one's own mind perverse. But for those who do believe that there are some failures to know one's own mind, it will be natural to think that certain propositions enjoy a special status, distinguished by the fact that one not merely has a given attitude toward them, but knows that one has this attitude. And it is clear that this status, while in general less important than the first order attitude itself, is an important one in the context of communication, where one's own standing with respect to the contents of one's acceptances may be called

[24]Cohen and Comesana 2013, Greco 2014b

into question, and where others will use one's own utterance as strong evidence for facts about one's own acceptances. If I myself do not know that I believe a proposition, for example, it would be bizarre for me to convey the contents of this belief to others, since I would then be giving *them* evidence about the state of my own mind which I myself do not possess.

2.6.5. Conclusion. The epistemic sentences discussed in this section don't provide the basis of an argument against the minimal theory of common ground. We may preserve the minimal theory by expanding the logic of acceptance or by invoking general laws about speech acts. On the most promising theories of acceptance, the expansion of the logic of acceptance does not entail even a restricted KK principle.

2.7. Conclusion

I have argued that the main reasons people have favored the iterative conception of common ground are uncompelling. The minimal theory of common ground can explain speaker presupposition just as well as Stalnaker's theory; it yields a theory of common ground according to which common ground is guaranteed to be at least as informative as it would be on the iterative theory; it can account for the basic phenomena of update just as well as the iterative theory; and it can be used to explain extended Moore-like sentences.

Recall now the challenge with which we began. In certain cases of communication it is not possible for the attitudes of the participants to become a matter of common knowledge or even belief among the participants. In successful communication, the attitudes of the participants toward a given subject matter change. I tell you that Mary is coming to town, and you learn something about Mary. But according to the iterated theory of common ground, this change is not enough for normal or successful communication. It must also be that your change in attitude becomes public between us. Our opening examples put pressure on this idea. Communication seems normal, even when it is not possible to make the change in attitudes public between "speaker" and "hearer".

The minimal theory can be thought of as a return to the core Gricean thought which has animated Stalnaker's theory of language throughout. Communication induces complex changes in participants' attitudes. But communication is not in the first instance directed at making these changes in attitude public. This may often be a consequence of the ways in which our attitudes change throughout a conversation, but the fact that our change in view becomes public is not a core feature of communication. In my view, the

common acceptance or common belief based theories of communication adulterate this pure Gricean thought by making the publicity of our changes of attitude central to the theory of communication. Part of my aim in this paper has been to show that Stalnaker's most profound discoveries about the nature of communication do not depend on his view about the publicity of our changes of attitudes. Once Stalnaker's theory is liberated from the unnecessary encumbrance of common belief or acceptance, we can see more clearly the elegance, simplicity and robustness of his insights into the structure of conversations.

2.A. Models

In the Appendices, I use the formalism of pointed neighborhood models. I first introduce these models, and then discuss briefly why this formalism is appropriate.

2.A.1. Models. A pointed conversational model is a structure

$$\langle \Omega, a, I, (B)_{i\in I}, (K)_{i\in I}, (A)_{i\in I}\rangle,$$

where "propositions" are subsets of the set of states Ω, and $a \in \Omega$ is interpreted as the "actual state". I is then a set of agents, the participants in the conversation. The functions B_i, K_i, and A_i represent these participants' beliefs, knowledge, and acceptances: B_i maps a proposition E to the proposition that i believes E; K_i takes a proposition E to the proposition that i knows E, and A_i takes E to the proposition that i accepts E. As usual, I will use $\neg E$ to abbreviate $\Omega \setminus E$, and $E \to F$ for $(\neg E) \cup F$. Note that this last definition allows an analog of contraposition, since obviously $\neg E \cup F = \neg(\neg F) \cup \neg E$. (I will often also write propositions with lower case p's and q's, but the reader should remember that we don't have an object language in the normal sense.)

I will use $\Box$ as a variable over the operators A_i, B_i and K_i. The functions representing these attitudes can be used to define neighborhood functions, $N^{\Box} : \Omega \to \mathscr{P}(\mathscr{P}(\Omega))$, where $E \in N^{\Box}(w)$ just in case $w \in \Box E$. Neighborhood models generalize standard Kripke semantics for modal logic. Kripke models associate with each world an ultrafilter on the power set algebra of the state space. (That is, a filter on the algebra of propositions.) In our setting, these are the propositions an agent believes knows or accepts. Neighborhood models, by contrast, associate each world with an arbitrary subset of $\mathscr{P}(\Omega)$, an arbitrary set of propositions. This may be an ultrafilter on the algebra of propositions, but it needn't be.

In addition to the above, pointed conversational models satisfy three additional requirements:

Necessitation $\cap_{i\in I}\Box_i\Omega = \Omega$

Anti-Necessitation $\cup_{i\in I}\Box_i\emptyset = \emptyset$

Factivity $\quad K_i(E) \to E = \Omega$.

The first two conditions ensure that there's always something the agent believes, and, moreover, the agent doesn't believe the absurdity. In Kripke models (to be introduced in a moment), the first condition has the consequence that the **D** axiom is satisfied ($\Box E \to \neg\Box\neg E$). The third condition is self-explanatory.

We define group operators as follows:

$$\Box_E(E) := \cap_{i\in I}\Box_i E.$$

(Here $\Box_E$ involves a slight abuse of notation; the idea is that "everyone accepts" is defined so that $A_E(E) := \cap_{i\in I}A_i E$, and similarly for "everyone knows" and "everyone believes".)

We then define the common ground:

$$CG(E) := A_E(E).$$

Finally, fixing a pointed conversational model, the event of it being commonly $\Box$'ed that E is defined recursively. Letting $\Box_E^1 E = \Box_E E$ and $\Box_E^n E = \Box_E\Box_E^{n-1}E$, we define: $C\Box = \cap_{n\in\mathbb{N}}\Box_E^n E$. This definition gives us a common knowledge operator CK, common belief, CB, and common acceptance CA.

2.A.2. Stalnaker Models. We now define propositions corresponding to the satisfaction of some standard axioms:

(K): $\cap_{i\in I}\cap_{\Box\in\{B_i,K_i,A_i\}}\cap_{E\subseteq\Omega}\cap_{F\subseteq\Omega}(\Box(E\to F)\to(\Box(E)\to\Box(F)))$

(C): $\cap_{i\in I}\cap_{\Box\in\{B_i,K_i,A_i\}}\cap_{E\subseteq\Omega}\cap_{F\subseteq\Omega}(\Box(E)\cap\Box(F))\to\Box(E\cap F)$

(4): $\cap_{i\in I}\cap_{\Box\in\{B_i,K_i,A_i\}}\cap_{E\subseteq\Omega}\Box(E)\to\Box\Box(E)$

(5): $\cap_{i\in I}\cap_{\Box\in\{B_i,K_i,A_i\}}\cap_{E\subseteq\Omega}\neg\Box(E)\to\Box\neg\Box(E)$

(BA4): $\cap_{i\in I}\cap_{E\subseteq\Omega}A_i(E)\to B_iA_i(E)$

(BA5): $\cap_{i\in I}\cap_{E\subseteq\Omega}\neg A_i(E)\to B_i\neg A_i(E)$.

A Hintikka-Kripke conversational model is a conversational model with $\mathbf{K}\cap\mathbf{C} = \Omega$. A $KD45$ conversational model is a Hintikka-Kripke conversational model with $\mathbf{4}\cap\mathbf{5} = \Omega$ (note that D is guaranteed by the semantic form of necessitation above). A Stalnaker model is a $KD45$ Hintikka-Kripke conversational model with $\mathbf{BA4}\cap\mathbf{BA5} = \Omega$. (Given

these properties, it is unclear why we should not write this B as K, since one's introspective beliefs at least are always true, but we'll maintain this version of the axioms for present purposes.) Finally, we define an a-Stalnaker model as a conversational model such that $a \in \mathbf{K} \cap \mathbf{C} \cap \mathbf{4} \cap \mathbf{5} \cap \mathbf{BA4} \cap \mathbf{BA5}$. Analogously a specific operator $\Box$ satisfies a-K if $a \in \cap_{E \subseteq \Omega} \cap_{F \subseteq \Omega} (\Box(E \to F) \to (\Box(E) \to \Box(F)))$, and similarly for the other conditions.

We can then define three notions of Stalnakerian common ground. Fixing a pointed conversational model:

(CB-A) $CG_{S_1} E = CBA_E E$

(CK-A) $CG_{S_2} E = CKA_E E$

(CA) $CG_{S_3} E = CAE$.

Note that these notions are defined for any conversational model.

2.A.3. The Context Set. It will also be useful to have the notion of a "context set". As above, in Kripke frames, we simply defined this as the set of worlds reachable by the common ground accessibility relation. In neighborhood frames, we need to do a bit more work to define this accessibility relation, since the individual agents' acceptances cannot necessarily be defined from such a relation. So we define a relation from a given function as follows, for an arbitrary function $\Box$:

$$R^{\Box}(w) = \{w' : \exists E[(\exists F \supseteq E)(F \in N^{\Box}(w)) \wedge (\nexists G \subsetneq E)(G \in N^{\Box}(w)) \wedge w' \in E]\}$$

(Note that this $\Box$ is no longer restricted to the B_i, K_i, A_i; it can be defined also for A_E, for example.)

For each operator, we can use this accessibility relation to define a new function:

$$\Box^C E = \{w : R^{\Box}(w) \subseteq E\}.$$

Since this is an operator defined from an accessibility relation, as in Kripke models, for any w, the set of propositions $\{E : w \in \Box^C(E)\}$ is a filter in the power set algebra of Ω. Moreover, if we started with a relational (Kripke) model of the operator $\Box$, it's obvious from the definition that if $w \in \Box(E)$ then $w \in \Box^C(E)$. (This will mean that if the A_i are

given by an accessibility relation as in the main text, the definition of context set here will be equivalent to the one in the main text, since everything there was finite.)[25]

To understand the interpretation of the defined operator, consider the case of acceptance, where we read A_i^C as "i assumes that".[26] In our general models, assumption is independent of acceptance. In this setting, propositions may be accepted but not assumed, for example, if an agent accepts only three propositions, Ω, E, F, where $E \subsetneq F$ and $F \subsetneq E$. Conversely, a proposition may be assumed but not accepted, for example, if an agent accepts only E and Ω, where $\Omega \setminus E \neq \emptyset$. One need not consider a proposition in order to assume it. "You're right: I hadn't thought about it," I might say, "I was simply assuming it" or "I was simply taking it for granted". Assumptions lie behind each of our surface, logically ill-behaved acceptances.

Now we define a context set for an arbitrary world

Context Set $c_w = \bigcup_{i \in I} R^{A_i}(w)$

(Note again, that in finite Kripke models, this definition is equivalent to the definition used in the main text.)

2.A.4. Neighborhood Models. The use of neighborhood models in this setting can be motivated in two ways. First, from a purely formal perspective, the generality of neighborhood frames gives us insight into what is important in the models. When we attempt to do logic without the law of excluded middle, familiar equivalences no longer hold. This forces us to be particularly careful about which assumptions are required in which derivations. If we only knew of models of our logic in which the law of the excluded middle held, we would have difficulty assessing which aspects of our logical systems were due to the law of the excluded middle, since we would not be able to consider models where that principle failed. The logic of belief and knowledge is no different. Making fewer background assumptions helps us to understand which assumptions are and are not needed for particular applications.

The use of these more general models can also be motivated from a psychological perspective. As is well known, standard Hintikka-Kripke models represent agents with unrealistically idealized logical competence. In Hintikka-Kripke models, if an agent believes that p, and believes that q, then she believes that $p \wedge q$. But many reject this

[25] In finite relational frames, of course, the exact converse holds, since there can be no infinite descending chains. In infinite relational frames, we have only an inexact converse: if $w \in \Box^C E$, then $(\forall F)($ if $F \subsetneq E$ then $w \in \Box(F))$.

[26] Note that assumption is sometimes used in a very different way in epistemic logic, e.g. Brandenburger and Keisler 2006. Our usage has no relationship to theirs.

view of belief. For example, proponents of the Lockean thesis, that belief is sufficiently high confidence, reject it because one's confidence in a conjunction may be lower than one's confidence in the conjuncts, and so fall below the threshold required for belief. In more banal examples, people may behave as if they do not believe the conjunction of two propositions they believe because they have never considered the conjunction. Similarly, in Kripke models, if a subject believes that p, and p entails q, then the subject believes that q as well. But what if the subject has not acquired the conceptual resources to consider the proposition that q under any guise? In this case, her behavior also will not accord with her supposed "belief" that q.

The relaxation of the assumption in Kripke frames that every agent believes every theorem of propositional logic is in general less significant in neighborhood models. Neighborhood frames still do not allow us to represent distinct modes of presentation, so if an agent believes *any* instance of any theorem of propositional logic, the agent will be represented as believing all theorems. But it's very plausible that we each believe some instance of the law of excluded middle (for example). So we should represent agents as believing at least one propositional tautology. In my models, accordingly, I'll be assuming that every agent accepts (and believes, and knows) the propositional tautology, so in every model I consider, agents will still be omniscient with respect to propositional logic.

2.A.5. Pointed Models? In general, I will be focusing on pointed models, where the axioms are satisfied only at the actual world a. But is this the right choice? We could have treated propositional logic and the logic of belief on a par, by requiring that every world in every model satisfy the proposed axioms of common ground, acceptance, belief and knowledge. This, after all, is what happens in Hintikka-Kripke conversational models, and in Stalnaker models, too.

But this "global" method has the immediate consequence that, if an agent believes a propositional tautology, he or she will believe and know that the target axioms hold. It has a more drastic, less obvious consequence, too. Recall that agents mutually know^1 that p if all agents know that p. Agents mutually know^n that p if they mutually know^1 that they mutually know^{n-1} that p. Then agents commonly know that p if they mutually know^n that p for all natural numbers n. Now suppose, as is extremely plausible, that at every world in every model every agent knows *some* instance of a propositional tautology ("Look", I say to you "it's either raining now, or it's not"). Then the axioms will not only be known, but, more strikingly, they will be a matter of common knowledge. Knowledge is

not special in this respect; we could have run the same argument with belief or acceptance in place of knowledge (or with certainty, if our models were endowed with probabilities).

The assumption that the axioms are commonly known will be most obviously uncongenial to those who understand features of the "logic" of belief as not really features of logic at all, but contingent facts about human psychology and biology. For example, psycholinguists who conduct empirical studies of the common ground will hold that it is neither an a priori nor a necessary matter whether the beliefs of people pattern in a given way. There are of course *some* features of belief which are necessary or a priori—for example that it has contents—it's just that the interesting features of belief we aim to study are not among those special features. If one takes this view of the properties of belief which interest us here, it is not just natural, but mandatory that the agents in our models *not* be apprised of the full psychological and biological truth about belief. After all, it's precisely that contingent truth that we're trying to discover.

But pragmatic reasons also militate against this method of building models, reasons which apply even if we should suppose that the relevant laws of belief are an a priori or necessary matter, so that any creature who had beliefs at all would have beliefs which obey these laws. The method forces us to prejudge important questions about the role of iterated beliefs, precisely the target of our investigation. Since the class of models cannot draw the appropriate distinctions between situations in which the agents mutually believe[1] the axioms and one in which the agents commonly believe the axioms, they do not allow us to distinguish when a particular phenomenon is due to belief in the axioms, and when it is due to common belief. We should not choose models which are unable to distinguish importantly distinct hypotheses about our subject matter.

Moreover, granting again the hypothesis that the relevant laws of belief are a priori or necessary, we may still wish to treat these truths differently from the laws of propositional logic. It seems fairly clear that agents in the world do not behave as if they have perfect knowledge of propositional logic. It's a shortcoming of our models that they don't allow us to represent distinctions among different theorems of propositional logic: between the "obvious" ones and the "unobvious" ones. But we should be grateful that, in the case of axioms on belief, we are not compelled to adopt this unrealistic assumption. Pointed models allow us the chance to isolate a class of propositions for which the assumption of logical omniscience is not forced on us. Whatever our views of belief, we should not forego this chance.

2.B. Stalnaker's Theory of Assertion

Stalnaker's theory can be stated as involving three principles:

Uniformity: In cases of rational communication, the same proposition is asserted in every candidate context.[27]

Default: If no norm of rational conversation is violated, the content of an assertion of a sentence s is the semantic content of s in the actual world.[28]

Recall that Stalnaker holds that a standard repair strategy is to identify asserted content with *diagonal content.* The diagonal content of an utterance is the proposition that is true at a world if the semantic content of the utterance as made at that world is true, and false otherwise. Stalnaker's hope seems to be that if we identify asserted content with diagonal content, compliance with Uniformity will be restored (at least if the assertion has any chance of success). We strengthen this principle about diagonal content into a generalization:

Diagonalization: If the content of an assertion of s is not the semantic content of s, then the asserted content is the diagonal content of s.

To state these principles more formally, we need a more precise notion of utterances. For a given conversational model M, the set of utterances U (Kaplanian characters) will be the set of functions $u : \Omega \to 2^{\Omega} \setminus \{\emptyset\}$, where each utterance is understood to be a function from worlds to the semantic content of the utterance at that world (we assume, for simplicity, that contradictions are never uttered). For $u \in U$, we write $u(w)$ to denote the semantic content of u as uttered at w. We often think of this proposition as identified with its characteristic function, letting $u(w)(w') = 1$ iff $w' \in u(w)$, and 0 otherwise. The diagonal proposition is then constructed for each u as the set $u^d = \{w : w \in u(w)\}$. Note that we assume the set of utterances is very rich; every function from worlds to propositions is associated with an utterance.

Fixing a model and an utterance u, we write the asserted content of an utterance as $a_u : \Omega \to 2^{\Omega}$.

Now, fixing a pointed conversational model, the axioms Hawthorne and Magidor attribute to Stalnaker are:

Uniformity: $\forall w(\forall v, x \in c_w)(a_u(v) = a_u(x))$

HMDefault: $\forall w\ (\forall v \in c_w)a_u(v) = u(w) \Rightarrow a_u(w) = u(w)$

---

[27] Hawthorne and Magidor 2009: 380; cf. Stalnaker 1978 [1999]: 88

[28] In fact, the formulation of this principle is due to Hawthorne and Magidor, but Stalnaker does not contest it in his reply (2009) to their paper.

HMDiagonalization: $(\forall w)[((\exists v \in c_w)(a_u(v) \neq u(w)) \Rightarrow a_u(w) = u^d]$

(I use $\Rightarrow$ for the material conditional in the meta-language. It should not be mistaken for a strengthening of the set-theoretically defined $\rightarrow$.)

These laws depart from the spirit of Stalnaker's theory in giving a world w an important place in what is asserted, even if $w \notin c_w$. This is especially clear in HMDiagonalization. Even if $w \notin c_w$, $u(w)$ has an important effect on what can be asserted in that context. The following combination (together with Uniformity) seems more in line with Stalnaker's original:

Default: $\forall w\ (\forall v, x \in c_w)(a_u(v) = u(x)) \Rightarrow \forall x \in c_w(a_u(w) = u(x))$
Diagonalization: $\forall w\ (\exists v, x \in c_w)(a_u(v) \neq u(x)) \Rightarrow (a_u(w) = u^d)$

This pair of axioms yields odd consequences if there can be $v \in c_w$ such that both $v \notin c_v$ and $c_v \not\subseteq c_w$. But Stalnaker's S_3 theory of context rules out this possibility by definition, since it validates Positive CG-Introspection:

Positive CG-Introspection: $\Omega{=}CG(E) \rightarrow CGCG(E)$

We can now ask for what class of pointed conversational models M it is the case that for any $u \in U$, u satisfies these axioms. This question was first posed by Hawthorne and Magidor 2009, who answered it in a more restricted setting.

(In fact, they claim that one-step transitivity or symmetry violations are enough to lead to conflicts in these three principles, but this is not quite enough, since Uniformity, Default and Diagonalization really impose conditions on the transitive closure of the context-accessibility relation, as the following proof shows. Their argument on the basis of transitivity violations (383) is also incorrect: they claim that the presence of a world $w \in c_a$ where one diagonalizes is insufficient to force diagonal content to be asserted in c_a, but this contradicts their own formulation of Uniformity: any such world forces diagonal content to be asserted in c_a as well. This is made clear by the "sublemma" in the proof below.)

Fixing a conversational model, let R^* be the transitive closure of the union of the R^{A_i}, defined in Appendix 2.A.3. (We could equivalently define it as the transitive closure of R^{A_E}.) Then we show:

PROPOSITION 2.B.1. *Let M be a pointed conversational model.*

(A) All utterances u satisfy Uniformity, Default and Diagonalization if and only if (*) $\forall w(\forall v \in c_w)[R^*(w) = R^*(v)]$.

(B) All utterances u satisfy Uniformity, HM-Default and HM-Diagonalization if and only if (*) $\forall w(\forall v \in c_w)[R^*(w) = R^*(v)]$ *and* ***(T)*** $(\forall w)(w \in R^*(w))$.

PROOF. (A) *If*: If $(\forall v, x \in c_w)(u(v) = u(x))$, then for all such x, $a_u(w) = u(x)$, and Uniformity is trivially satisfied. If $a_u(w) = u^d$, then there's some $v, x \in c_w$ such that $a_u(v) \neq u(x)$. Since $R^*(v) = R^*(w)$, both v and x are in $R^*(v)$. We show that $a_u(v) = a_u(x) = u^d$ by way of a general lemma (which itself requires a sublemma).

The statement of the lemma is: If $\exists b, c \in R^*(z)(a_u(b) \neq a_u(c))$ then $a_u(z) = u^d$.

There are finite paths from d to b and d to c using R^{A_E}: we induct on the maximum of these two path lengths. For the base case: if $b, c \in R^{A_E}(z)$ then there are two cases. If both $a_u(b) = u(b)$ and $a_u(c) = u(c)$, then since $a_u(b) = u(b) \neq u(c) = a_u(c)$ $a_u(z) = u^d$. If either $a_u(b) \neq u(b)$ or $a_u(c) \neq u(c)$, then these single worlds fulfill the antecedent of Diagonalization, so again $a_u(z) = u^d$. Now we assume we have shown that if both b and c are reachable in at most n steps, then $a_u(z) = u^d$, and show the claim for $n + 1$ steps.

For this we need a sublemma: if $\exists f \in R^*(g)(a_u(f) = u^d)$ then $a_u(g) = u^d$. Again, we prove this by induction on path length. For the base case, if $u(f) \neq u^d$, then f fulfills the antecedent of Diagonalization, so $a_u(g) = u^d$. If $u(f) = u^d$, then the only way Default could be fulfilled is if all $h \in c_g$ have $u(h) = u^d$, but this would still make $a_u(g) = u^d$. Now we suppose the claim holds for worlds reachable in n steps and show it for $n + 1$. By hypothesis there's some n-reachable f' so that $f \in c_{f'}$, and $a_u(f) = u^d$. We proceed exactly as in the base case to show that $a_u(f') = u(d)$. Since f' is reachable in n steps, it follows by the induction hypothesis that $a_u(g) = u^d$.

Now we return to the induction of our lemma. We know there are b', c', reachable in at most n steps from z such that $b \in R^{A_E}(b')$ and $c \in R^{A_E}(c')$. Now there are two cases. If $a_u(b') = u(b') = a_u(b) = u(b)$ and $a_u(c') = u(c') = a_u(c) = u(c)$, then since $a_u(c) \neq a_u(b)$, and c', b' are reachable in n steps, we're done. And if for any $h \in \{b, b', c, c'\}$, $a_u(h) = u^d$, then the sublemma gives us the desired result, that $a_u(z) = u^d$.

With the lemma and the sublemma in hand, the main *If*: claim follows trivially. If $a_u(v) = u^d$ or $a_u(x) = u^d$, then we use the sublemma to show that both of them must be u^d. The supposition that both of them have $a_u(v) = u(v)$ and $a_u(x) = u(x)$ leads to contradiction using the main lemma. So Uniformity is satisfied.

Only if: Now suppose $(*)$ fails: there's a w, w' so that $w' \in c_w$ but $R^*(w) \neq R^*(w')$. This can only be because $R^*(w') \subsetneq R^*(w)$ (and so, $w \notin R^*(w')$). Then define a u so that

for all $w'' \in R^*(w')$ $u(w') = u(w'')$, but for some $w''' \in R^*(w) \setminus R^*(w')$, $u(w''') \neq u(w')$. Then u^d is asserted at w, but at w', $u(w') \neq u^d$ is asserted, in contradiction of Uniformity.

(B) If: If $(\forall v \in c_w)(a_u(v) = u(w))$, then $a_u(w) = u(w)$, and Uniformity is trivially satisfied. If on the other hand $(\exists v \in c_w)(a_u(v) \neq u(w))$, so that $a_u(w) = u^d$, we now show that $(\forall v \in c_w)(a_u(v) = u^d)$ as well. We do this with a version of the earlier "sublemma", proving by induction on path length in R^{A_E} that if $(\exists x \in R^*(v))(a_u(x) = u^d)$ then $a_u(v) = u^d$ as well. For the base case, if $u(v) = u^d$ already, then we're done. If not, then since we're assuming $x \in c_w$ HMDiagonalization immediately gives that $a_u(v) = u^d$. The induction step is similar. So since $a_u(w) = u^d$ and $w \in R^*(v)$, it follows that $a_u(v) = u^d$ as well.

Only if: If (*) fails, the argument is exactly as before.

Now suppose (T) fails: $w \notin R^*(w)$. Define a u so that for all $w', w'' \in R^*(w)$, $u(w') = u(w'')$, but for w itself, $u(w) \neq u(w')$. Then u^d is asserted at w, but at w', $u(w') \neq u^d$ is asserted, in contradiction of Uniformity. □

This result gives us another reason to prefer Default and Diagonalization as Stalnaker's intended axioms. For it is often claimed that the following two principles are sufficient to ensure that Stalnaker's theory is consistent:

Positive CG-Introspection: $\Omega{=}CG(E) \rightarrow CGCG(E)$
Negative CG-Introspection: $\Omega = \neg CG(E) \rightarrow CG\neg CG(E)$;

but if we use the HM principles, these are necessary but not sufficient to ensure that the theory is consistent.

In Hintikka-Kripke frames, Positive CG-Introspection ensures that $c_w = R^*(w)$. Negative CG-Introspection is needed to ensure that $(\forall v \in c_w)(c_v = c_w)$. As I've observed often enough Stalnaker's S_3-theory (though not, I repeat, S_1 and S_2) validates positive introspection by definition, Proposition 2.B.1 demonstrates that Negative CG-Introspection is the key to the question of whether his theory of assertion is consistent with his theory of the common ground. The next appendix studies this axiom in some detail.

2.C. Negative Introspection

To guarantee negative introspection, Stalnaker needs the following condition (the name for the condition is mine):

Genuinely Non-Defective: $a \in \cap_{i\in I}(A_iCAp \rightarrow CAp)$ and $a \in CA(\cap_{i\in I}(A_iCAp \rightarrow CAp))$[29]

In Stalnaker models, this axiom delivers negative introspection for common ground (an analogous axiom delivers it for presupposition). I'll state and prove a more general version of the theorem which shows this, since the more general statement is illuminating.

PROPOSITION 2.C.1. *Let M be a Hintikka-Kripke conversational model. Then for any $n \in \mathbb{N}$, if the model satisfies*

(i) $a \in \neg A_ip \rightarrow A_i\neg A_ip$

(ii) $a \in A_iA^n_Ep \rightarrow A^n_Ep$

(iii) $a \in A_i(A^n_Ep \rightarrow A_iA^n_Ep)$

(ia) $a \in A^n_E(\neg A_ip \rightarrow A_i\neg A_ip)$

(iia) $a \in A^n_E(A_iA^n_Ep \rightarrow A^n_Ep)$ *and*

$(iiia)$ $a \in A^n_E(A_i(A^n_Ep \rightarrow A_iA^n_Ep))$ *then*

$a \in \neg A^n_Ep \rightarrow A^n_E\neg A^n_Ep$.

Stalnaker's version of the proposition replaces my A^n with CA. Even when we make these replacements, (i) and (ia) hold trivially in Stalnaker-models, since $\mathbf{5} = \Omega$. In Stalnaker models, all agents are negatively introspective, *and* it's commonly accepted that they are. Moreover, both (iii) and $(iiia)$ hold by definition when we replace A^n_E with CA. Finally, Stalnaker models are of course Hintikka-Kripke conversational models.

PROOF. It's a standard fact that any operator formed by iterating and conjoining normal modal operators is itself normal: that is, the versions of **K** and **C** for A^n_E also equal the universe Ω. We first show that (i),(ii),(iii) entail $a \in \neg A^n_E \rightarrow A_E\neg A^n_Ep$. Assume $a \in \neg A^n_Ep$. By contraposition of (ii) we have $a \in \neg A_iA^n_Ep$, by (i) we have $a \in A_i\neg A_iA^n_Ep$. By contraposition of (iii) and the fact that **K** holds for A_i at a, we have $a \in A_i\neg A^n_Ep$. We repeat for each $i \in I$ to derive $A_E\neg A^E_np$. Thus by conditional proof, we have (using (i), (ii), (iii) as abbreviations for the statements) $\Omega = (i)\wedge(ii)\wedge(iii) \rightarrow [\neg A^n_Ep \rightarrow A_E\neg A^n_Ep]$. Now we use RN to derive $a \in A^n_E([(i) \wedge (ii) \wedge (iii)] \rightarrow [\neg A^n_Ep \rightarrow A_E\neg A^n_Ep])$. We use (ia), (iia), $(iiia)$ and **C** to give us that $A^n_E[(i) \wedge (ii) \wedge (iii)]$. So then by K we have $(*)$ $A^n_E(\neg A^n_Ep \rightarrow A_E\neg A^n_Ep)$. But now suppose that $\neg A^n_Ep$. By $(i), (ii), (iii)$ we have $A_E\neg A^n_Ep$. and by $(*)$ and K, we have $A_E\neg A_Ep \rightarrow A_EA_E\neg A_Ep$, and so on up to n. □

The minimal theory requires fewer assumptions to prove the same conclusion, that negative introspection holds for common ground:

[29]See next appendix for discussion of Stalnaker's own definition of non-defective contexts.

PROPOSITION 2.C.2. *(Minimal Negative Introspection) Let M be a pointed conversational model where each A_i obeys $a - K$ ($a \in A_i(p \to q) \to A_i p \to A_i q$). Then if*

(i) $a \in \neg A_i p \to A_i \neg A_i p$

(ii) $a \in A_i A_E p \to A_E p$

(iii) $a \in A_i(A_E p \to A_i A_E p)$ *then*

$a \in \neg A_E p \to A_E \neg A_E p$.

PROOF. If $a \in \neg A_E p$, then by *(ii)*, $a \in \neg A_i A_E p$. By *(i)* $a \in A_i \neg A_i A_E p$. By contraposition of *(iii)* and $a - K$ for A_i, $a \in A_i \neg A_E p$. Since i was chosen arbitrarily the claim holds for all i, and so $a \in A_E \neg A_E p$. □

This proposition drops three controversial (in my view, implausible) assumptions: (1) that agents have to accept (in fact commonly accept) that acceptance obeys negative introspection; (2) that agents' acceptances are closed under conjunction in the problematic direction: $AE \cap AF \to A(E \cap F)$; (3) that the context not just *be* non-defective, but it be commonly accepted that it is non-defective.

Stalnaker does not find these assumptions problematic, so this difference won't convince him. He believes that the introspection axioms are necessary facts about belief, and so, given his coarse grained conception of content, he is committed to the claim that we do commonly know these facts about belief. Moreover, he thinks belief does satisfy this closure principle, and, third, he may simply take it as a deliverance of his theory that genuine non-defectiveness requires common acceptance that the context is non-defective.

On its own, then, negative introspection leaves us at an impasse. Both Stalnaker and I require additional assumptions to ensure this introspection axiom. The assumptions needed on the minimal theory are weaker than those needed on Stalnaker's, but Stalnaker will not find the difference in strength conceptually important.

This brings us back to Positive CG-Introspection. Positive CG-Introspection is in fact a logical watershed between Stalnaker's theory and the minimal one. The reason has already been touched on in passing. If we add Positive CG-introspection to the minimal theory, then according to the new theory, if it's common ground that p, it is also S_3-common ground that p.[30] This holds for Positive CG-Introspection in any theory where common ground is defined by iterating mutual acceptance.

[30]I should also note that this version of collapse is a particularly implausible theory. There are clearly cases where you and I both accept something, but we fail to mutually accept that both accept it (this is a point on which Stalnaker and I agree). But the theory under consideration in the main text would (bizarrely) rule out this commonplace phenomenon.

In light of Proposition 2.B.1 this difference might seem critical. If Stalnaker's theory of assertion is to be consistent for all the set of all utterances, we must have *both* positive and negative introspection.

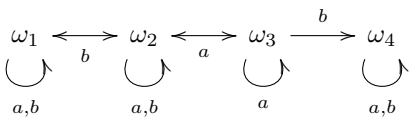
But I don't think the argument that the minimal theory fails here is a very strong one. The minimal theorist has two replies which do not require guaranteeing positive introspection in all contexts. The conservative option is to hold that Diagonalization was merely intended as a restricted principle governing some utterances in some contexts. The aim of the theory was surely not to claim that every utterance can always be interpreted! A less conservative option is to give up on Stalnaker's theory of assertion full stop. This option should not, I think, be overlooked. The theory of Diagonalization, Uniformity and Default has interesting structure, but I do not believe that its predictions are so well confirmed for its rejection to count as a clear cost to the minimal theory of common ground.

2.D. Non-Defective Contexts

Instead of "Genuine Non-Defective", Stalnaker originally defined the following notion of context:

Non-Defective: A context is non-defective if, for all i, if i S-presupposes that p, then it's S-common ground that p.

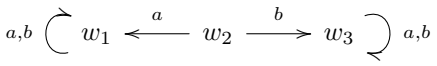
In early draft material for this paper, I provided countermodels to the following claim (Stalnaker 2009: 401) "If a speaker believes the context is nondefective, then that speaker's presuppositions will satisfy both negative and positive introspection." The construction of counterexamples to the claim is sufficiently delicate that I include an example here:



At ω_1, the context is non-defective in the defined sense, and both a and b believe it is. And yet, while it's not common ground that $\{\omega_4\}$, it's also not common ground that it's not common ground that $\{w_4\}$. (In fact, this example is simpler than my originals; thanks to Bob Stalnaker for showing me a related one, and also to Matt Mandelkern who caught some errors in an earlier draft.)

In response to these earlier counterexamples, Stalnaker now adopts a version of the condition which I called "Genuinely Non-Defective", and a variant for presupposition

(it's presupposed that the context is non-defective), to ensure negative introspection for common ground and presupposition (2014: Appendix main text at n. 14). In fact, while the condition Stalnaker now provides for presupposition is correct, the one for common ground is still not quite right: he says "if it is common ground that the context is nondefective, then the negative introspection principle will hold for the common ground." But in fact, the context must also *actually be* non-defective, as the following model shows. In the figure, at w_2 it is common ground that the context is non-defective (the common ground is $\{w_1, w_3\}$), and not common ground that $\{w_1\}$, but also not common ground that it is not common ground that $\{w_1\}$ (since at w_1, the common ground would be $\{w_1\}$).



In both cases, Stalnaker may have been led to his claims on the basis of a result by Bonanno and Nehring (2000), which shows that adding the *axiom* $B_iCBp \rightarrow CBp$ to a normal multi-modal doxastic logic where each individual's beliefs obey $KD45$ returns negative introspection for common belief. Their result depends on the fact that we can use Necessitation with respect to this axiom itself. In model-theoretic terms, it's not just that our beliefs about what we commonly believe are true *at the actual world*; they're true at every world in the model. As a result, for any world $B_iCBp \rightarrow CBp$ is both true at that world and a matter of common belief there. Proposition 2.C.1 is an easy generalization of their theorem, but the statement of it demonstrates the role Necessitation plays in the proof, by making explicit the beliefs about beliefs one must have for the proof to go through.

In any case, this point about definitions is a minor detail; the revised condition serves Stalnaker's theoretical purposes just as well as the one he adopts in the book.

2.E. Unique i-Candidate Contexts and Agent-Relative Introspection

In the main text I noted that it is important that the minimal theorist can define a class of contexts in which agents do know what the context is. This appendix fulfills that promise. We want the following two axioms.

Agent-Relative Positive CG-Introspection: $CGp \rightarrow B_iCGp$

Agent-Relative Negative CG-Introspection: $\neg CGp \rightarrow B_i\neg CGp$.[31]

[31]For reasons mentioned repeatedly above, replacing "B_i" in these conditions with A_i is unpromising, at least if one accepts that p only if one also adopts the attitude determined by the conversational

In fact, we might as well replace the B_i with K_i. By essentially the same argument as used for Proposition 3, we then have

PROPOSITION 2.E.1. *(Contextual Omniscience) Let M be a pointed conversational model where each B_i obeys $a - K$ ($a \in B_i(p \to q) \to (B_ip \to B_iq)$). Then if the model obeys:*

(i) $a \in \neg A_ip \to B_i\neg A_ip$

(ii) $a \in B_iCGp \leftrightarrow CGp$

(iii) $a \in B_i(CGp \to A_iCGp)$ *then*

$a \in CGp \to B_iCGp \cap \neg CGp \to B_i\neg CGp$.

There is one slight difference between this proposition and Proposition 2.C.2. Condition *(ii)* has now been strengthened to a biconditional, delivering Agent-Relative Positive Introspection by stipulation.

I take this result to show that, as far as the intuitive demand about knowing what the context is, there is little to choose between the two theories. Given the appropriate additional stipulations, each can deliver the needed form of omniscience about context. The case of Stalnaker's theory of assertion is a subtler question. It is unsurprising that Stalnaker's theory of context is tailored to that theory of assertion. But it is unclear whether that amounts to an argument, never mind a powerful one, against the minimal theory of common ground.

---

tone to p. Note that in our general neighborhood frames, these axioms will not suffice, since $R^{B_i}(w)$ is defined for the version of belief which corresponds to assumption, not for the original belief operator (which corresponds to acceptance). We then have two options. We can either impose an axiom directly on B_i^C for each agent, or we can alter i-Uniformity to be defined as: $(\forall w)(\exists w')(a_u(w') \in N_i^B(w))$. It then suffices to have simply: $\forall w \exists v(\{w' : c_v = c_{w'}\} \in N_i^B(w)$. The former will be my approach in what follows in the main text.

CHAPTER 3

People with Common Priors Can Agree to Disagree

Abstract. Robert Aumann presents his *Agreement Theorem* as the *key conditional*: "if two people have the same priors and their posteriors for an event A are common knowledge, then these posteriors are equal" (1976: 1236). This paper focuses on four assumptions which are used in Aumann's proof but are not explicit in the key conditional: (1) that agents commonly know, of some prior μ, that it is the common prior; (2) that agents commonly know that each of them updates on the prior by conditionalization; (3) that agents commonly know that if an agent knows a proposition, she knows that she knows that proposition (the "KK" principle); (4) that agents commonly know that they each update only on true propositions. It is shown that natural weakenings of any one of these strong assumptions can lead to countermodels to Aumann's key conditional. Examples are given in which agents who have a common prior and commonly know what probability they each assign to a proposition nevertheless assign that proposition unequal probabilities. To alter Aumann's famous slogan: people *can* "agree to disagree", even if they share a common prior. The epistemological significance of these examples is presented in terms of their role in a defense of the *Uniqueness Thesis*: If an agent whose total evidence is E is fully rational in taking doxastic attitude D to P, then necessarily, any subject with total evidence E who takes a different attitude to P is less than fully rational.

3.1. Introduction

Robert Aumann presented his seminal *Agreement Theorem* as the claim that: "If two people have the same priors, and their posteriors for an event A are common knowledge, then these posteriors are equal" (1976: 1236).[1] According to this *key conditional*, if two agents commonly know what probability they each assign to a proposition, their probabilities agree.[2] In Aumann's phrase, these agents cannot "agree to disagree".

One often reads that Aumann *proved* the key conditional, or that this key conditional requires no further assumptions worth questioning. Thus, Stephen Morris writes:

[1]Thanks to Frank Arntzenius, Daniel Rothschild, and Tim Williamson for detailed and insightful comments on numerous versions of this paper. I am grateful to two anonymous referees for extremely helpful comments and suggestions. Thanks finally to Peter Fritz, Jeremy Goodman and Tiankai Liu for many conversations both about this material and not.

[2]Aumann's result holds for potentially infinite numbers of agents, but it is difficult to state clearly in English with more than two. I will say that agents "commonly know" or "commonly believe" a proposition as an abbreviation for "it is common knowledge among the agents that", and "it is common belief among the agents that".

> "Perhaps the most compelling argument against the common prior assumption is the following *reductio* argument. If individuals had common prior beliefs then it would be impossible for them to publicly disagree with each other about anything, even if they started out with asymmetric information. Since such public "agreeing to disagree" among apparently rational individuals seems to be common...an assumption which rules it out is surely going to fail to explain important features of the world."[3]

Similarly, the computer scientist Scott Aaronson remarks that his "conclusion that complexity is *not* a major barrier to agreement" "strengthens Aumann's original theorem substantially, by forcing our attention back to origin of prior differences."[4]

But in fact, Aumann's original proof depends on a number of strong assumptions which are not stated explicitly in the key conditional. This paper focuses on four:

(1) that agents commonly know, of some prior μ, that it is the common prior;

(2) that agents commonly know that each of them updates on the prior by conditionalization;

(3) that agents commonly know that if an agent knows a proposition, she knows that she knows that proposition;

(4) that agents commonly know that they update only on true propositions.

I make each of these four assumptions explicit, and then show that natural weakenings of any one of them allows for countermodels to the key conditional. Four examples are given in which agents who have a common prior and commonly know the probabilities each assigns to a proposition nevertheless assign that proposition unequal probabilities. The examples show that the *reductio* articulated by Morris, and which lies behind Aaronson's remark requires these further, strong assumptions.

In fact, the examples show something subtler as well. The intuitive idea behind the common prior assumption is that if rational agents have the same information or evidence, they have the same beliefs. But when one formalizes this assumption, one typically imposes the common prior "globally", at every state in a given model. This way of formalizing the assumption has consequences which show that it is unfaithful to the original intuitive idea. In particular, the formal implementation of the assumption has the consequence that agents do not just *have* a common prior, they commonly know, of

[3]Morris 1995: 227-8

[4]Aaronson 2005: 3 (of the full version); his emphasis. See also Cowen and Hanson 2014 (forthcoming).

some prior μ, that it is the common prior. (This was the first implicit assumption noted above.) A similar point holds for each of our four target assumptions. In each case, the informal motivation suggests at most that a given law holds *at the actual state*, not that it is common knowledge, nor that it holds at every state in the model. And in fact, as my examples will show, the key conditional may fail even when *all* of Aumann's informal assumptions hold at the actual state. The apparent paradox of the Agreement Theorem can thus be resolved without giving up on *any* of Aumann's intuitive assumptions; all that must be relaxed are common knowledge assumptions which were unattractive to begin with.

3.1.1. Epistemological Motivation. One aim of this paper is to make the Agreement Theorem more accessible to philosophers. To do this, much of the paper will be devoted to presenting a particular epistemological interpretation of Aumann's theorem. This discussion is unnecessary for understanding the formal content of the paper. Readers primarily interested in these formal results may wish to skip these more discursive sections of the paper.[5]

Recently, a number of epistemologists have advocated versions of the Uniqueness Thesis:

Uniqueness Thesis: If an agent whose total evidence is E is fully rational in taking doxastic attitude D to P, then necessarily, any subject with total evidence E who takes a different attitude to P is less than fully rational.[6]

We will consider a Bayesian elaboration of the Uniqueness Thesis:[7]

Bayesian Uniqueness: There is some prior μ, such that, for any agent i and for any proposition P, if i's evidence is determined by a set of propositions the conjunction of which is a proposition E, then i is fully rational in her attitude to P only if she has a degree of confidence in P equal to $\mu(P|E)$.[8]

---

[5]Technically minded readers should simply read Sections 3.4 and 3.5 in detail. Sections 3.6 and 3.7 may then be skimmed with an eye to the "pointed" interpretation of the models presented there. The discussion of Moses and Nachum's (1990) problem may also be of interest, and is found in n. 36.

[6]White (2013), cited in Kelly (2013). An earlier version is found in White (2005: 445): "given one's total evidence, there is a unique rational doxastic attitude that one can take to any proposition." For a different development, see Feldman 2007. A charitable reading of this thesis requires prefixing "necessarily". Otherwise, if the conditional is material, then if no one is ever fully rational in taking a given doxastic attitude, the principle is true, but vacuously so.

[7]In Section 3.2.1, I will formalize the Uniqueness Thesis directly. In the introduction, however, I focus exclusively on Bayesian Uniqueness, since Aumann's original theorem applies most directly to that thesis.

[8]For the purposes of this paper, I will consider only finite models, where the evidential prior is assumed to be *regular*, that is, if μ is the evidential prior and Ω our finite universe, $\mu(E) > 0$ for every non-empty $E \in \mathscr{P}(\Omega)$. Thus I will be abstracting from substantial issues about undefined conditional probabilities. For some discussion, see Williamson 2007, Hájek 2010, Easwaran 2014. Note that Bayesian Uniqueness is vacuous if there is no agent whose evidence is "determined by" a set of propositions. Later, I rule this

To motivate a version of Bayesian Uniqueness, consider the evidential prior of Williamson (2000). In Williamson's view, conditionalizing the evidential prior on what one knows determines the degree to which one's body of evidence supports a given proposition. Williamson does not claim that the degree of support a proposition enjoys on one's evidence determines the attitude that one should take to that proposition. But a theorist of Bayesian Uniqueness who identifies μ with the evidential prior *would* endorse precisely that stronger claim.[9]

The key conditional can be used as a premise in a *reductio* of Bayesian Uniqueness which is directly analogous to Morris's argument. I will call this form of argument the "argument from Agreement", since it uses Aumann's key conditional as its main premise. The argument runs as follows. The key conditional says that if agents have the same prior, then they cannot "agree to disagree". Bayesian Uniqueness says that rational agents, insofar as they are rational, have the same prior. Together with the key conditional, Bayesian Uniqueness entails that rational agents cannot agree to disagree. But this result is absurd. Scientists who have exchanged their views in person and in print over the course of decades are paradigmatic examples of those who commonly know one another's opinions. Jurors who have sincerely debated the verdict in a long trial, and sincerely discussed their decisions with each other, are yet another paradigmatic example of those who commonly know one another's beliefs. Even so, these scientists may have different degrees of confidence in specific scientific theories, just as these jurors may cast different votes in a given trial. But these disagreements do not give us grounds to fault the scientists' or jurors' rationality.

---

out explicitly by adding a further thesis (Evidential Determination), which takes a stand on the nature of evidence.

[9]Some may agree with the spirit of Bayesian Uniqueness but reject its commitment to there being, for any proposition, a precise degree of confidence one should have in that proposition. One version of such a "mushy" or "imprecise" view holds that one's attitudes at a time should be represented by a set of probability functions, often called the "representer". Similarly, one's attitude to a given proposition P at a time is represented by a set of values, which by extension we call "the agent's representer on P". We would then have:

Mushy Uniqueness: There is some family of priors $\{\mu_\alpha\}_{\alpha\in A}$, for an index set A, such that, for any agent i, and for any proposition P, if i's evidence is determined by a set of propositions, the conjunction of which is a proposition E, then i is fully rational in her attitude to P only if i's representer on P is $\{\mu_\alpha(P|E)\}_{\alpha\in A}$.

In this setting, the key conditional would become: "if it is common knowledge among two agents what one agent's representer on P is, and common knowledge what the other's representer on P is, then their representers on P are the same." (It is easy to see that, since Aumann's original theorem holds for each element of the representer, the relevant modification of the original theorem will yield this modification of the key conditional. Theorem 3.2.1 and its corollary thus apply straightforwardly to Mushy Uniqueness.) I will not explore these imprecise or mushy ideas further here, except to note that the examples of later sections also provide ways of defending Mushy Uniqueness from the analogous argument from Agreement (see next paragraph in the main text). Since Bayesian Uniqueness is just a special case of Mushy Uniqueness (where the representer is a singleton), a model which satisfies Bayesian Uniqueness is also a model which satisfies Mushy Uniqueness.

This argument from Agreement threatens more than just Bayesian Uniqueness. Some reject Bayesian Uniqueness in its most general form, but still hold that very often, or when confined to certain subject matters, there is a uniquely rational attitude that each individual should have. Many subjective Bayesians, for example, hold that, although *some* agents' priors *sometimes* differ, very often or at least when restricted to propositions in certain subject matters, we do have the same priors. If the argument from Agreement is sound, it imposes a surprising and severe restriction on the prevalence of shared priors. It shows that common priors are only as common as agreeing to disagree is rare. For if you and I agree to disagree, then, given the key conditional, we have different priors.

The dialectic of this paper will thus run as follows. I take my "opponent" to be someone who claims that the argument from Agreement shows that Bayesian Uniqueness is false (and that agents share a common prior less often than the subjective Bayesian of the previous paragraph claims). The countermodels to the key conditional will then offer the theorist of Bayesian Uniqueness different ways of replying to the argument from Agreement. The formal examples show that the argument from Agreement requires further, unstated premises. The motivations for the examples show that these further premises are difficult to defend.

Before we begin, let me set aside one natural response to the argument from Agreement. According to this response, the scientists or jurors in the above examples do not in fact commonly know one another's opinions. Instead of challenging the key conditional itself, this response merely questions how often the key conditional applies. This response to the argument from Agreement is of considerable interest. I myself believe common knowledge is comparatively rare. But I won't consider this line of response further in the paper. Many philosophers, and many if not most economists and computer scientists believe that common knowledge is extremely common. If denying that scientists or jurors have achieved common knowledge represents the only resort of the Uniqueness Theorist, Aumann's Agreement Theorem would already have provided an astonishing result: that the Uniqueness Thesis is inconsistent with standard views about the prevalence and attainability of common knowledge.

3.1.2. Outline. Section 3.2 sets up the basic model. I use the formalism of decision functions to provide a version of the argument from Agreement which applies directly to the qualitative Uniqueness Thesis (and not just to its Bayesian elaboration). I further strengthen my opponent's position by proving a novel qualitative Agreement Theorem,

which uses only $S4$ (not $S5$) for knowledge. In Section 3.3 I describe how this theorem carries over to the Bayesian setting, and to Bayesian Uniqueness in particular (the proofs are in the Appendix).

Sections 3.4-3.7 develop the main observations of the paper, relaxing each of the target assumptions in order. In each of these examples, the agents have a common prior, and commonly know one another's posterior beliefs. But still, they disagree.

3.2. The Qualitative Theorem

Our basic models will use the framework of epistemic logic introduced by Hintikka (1962). We begin from an interactive Kripke frame, $\mathcal{F} =< \Omega, N, \{R_i\}_{i \in N} >$, containing: a set Ω of states or "worlds"; a countable set, N, of "agents"; and a set of accessibility relations (one for each agent) $R_i \subseteq \Omega \times \Omega$. For the purposes of this paper, we restrict attention to frames in which the set of states, Ω, is finite (that is, $|\Omega| = n$, for some positive integer n); this ensures that our results do not turn on paradoxes of infinity. Informally, we think of worlds as coarse-grainings of the space of maximally specific situations, by equivalence classes with respect to what is relevant in the case we are modeling. The relations R_i are then interpreted as describing what an agent "considers possible" at any given world.

We use these R_i to define *possibility correspondences*, $P_i : \Omega \to 2^\Omega$, with $P_i(\omega) := \{\omega' \in \Omega \mid \omega R_i \omega'\}$. These functions take a world ω to the set of worlds the agent i "considers possible" or "sees" at that world.[10] A *proposition* is a set of states, $E \subseteq \Omega$.[11] An agent is said to believe a proposition if and only if it holds at every world he or she considers possible. Formally, we represent this interpretation by defining a function $B_i : 2^\Omega \to 2^\Omega$, as follows: $B_i E := \{\omega \in \Omega | P_i(\omega) \subseteq E\}$. Each function B_i takes each proposition E to the proposition that i believes E.

We will be interested in agents with consistent beliefs. Accordingly, we require that, if an agent believes a proposition, he or she does not believe the negation of that proposition. This condition corresponds to the requirement that the accessibility relation be *serial*; an agent who has a serial accessibility relation will be said to satisfy **(D)**.[12] By extension,

[10]Note that we could have taken possibility correspondences as primitive and defined the accessibility relations R_i in terms of the possibility correspondences: $\omega R_i \omega' := \omega' \in P_i(\omega)$. These are equivalent and interchangeable ways of describing the models.

[11]It is a substantive and controversial thesis that the objects of "propositional attitudes" (e.g. belief-that and knowledge-that) can be represented by—never mind identified with—sets of states. Nevertheless, I adopt this assumption throughout, in accordance with the results I will be discussing.

[12]Seriality is the condition that: $(\forall \omega)(\exists \omega')\omega R_i \omega'$. It is equivalent to the condition that for all ω, $P_i(\omega) \neq \emptyset$. It is also equivalent to the condition that for all $E \subseteq \Omega$, $B_i E \subseteq \neg B_i \neg E$ (where "$\neg$" is an abbreviation for set complement, so that "$\neg E$" abbreviates $\Omega \setminus E$). For more on this, and the conditions discussed below, see again Fagin, Halpern, Moses and Vardi 1995.

a frame satisfies **(D)** (or one of the conditions introduced below) if every agent in the frame satisfies **(D)**. Every frame discussed in this paper satisfies **(D)**.

Knowledge is factive: for any proposition p, if an agent knows that p, then p. To represent the factivity of knowledge, we require that the accessibility relation be *reflexive*: an agent with a reflexive accessibility relation is such that, at any world, the agent considers that world possible. An agent with a reflexive accessibility relation will be said to satisfy **(T)**.[13]

We can also formalize what are sometimes called "introspection principles" by imposing additional constraints on the accessibility relation. An agent's beliefs are positively introspective just in case, for any proposition p, if she believes that p, she believes that she believes that p. Her beliefs are negatively introspective just in case, for any proposition p, if she fails to believe that p, she believes that she fails to believe that p.[14] These constraints, which are also called **(4)** and **(5)**, correspond respectively to *transitivity* and *euclideanness* of the accessibility relation.[15] When the possibility correspondences are interpreted as representing knowledge, positive introspection, **(4)**, is the formal analogue of a simplistic "KK"-principle.

$S5$ or *partitional frames* obey both **(T)** and **(5)**.[16] (Any frame which satisfies these constraints also satisfies **(4)**.) $S4$ *frames* obey **(4)** and **(T)**, and, finally, $KD45$ frames obey **(D)**, **(4)**, and **(5)**. These names are derived from the names of normal modal logics which are sound and complete with respect to the class of frames which satisfy the relevant conditions.

Epistemic models make a number of substantial idealizations about knowledge and belief. All standard models of this kind abstract from the possibility that someone might know that Hesperus rises in the evening, but not know that Phosphorus rises in the evening (presuming this is indeed possible).[17] Abstractly, if an agent believes that p and $p = q$, then the agent believes that q. In addition, Hintikka-Kripke models make

[13]Reflexivity is the condition that: $(\forall\omega \in \Omega)\omega R_i\omega$. It is equivalent to the condition that $\forall\omega\ \omega \in P_i(\omega)$. It is also equivalent to the condition that, for all $E \subseteq \Omega$, $B_iE \subseteq E$. Reflexivity, **(T)**, is strictly stronger than seriality, **(D)**.

[14]The same terminology applies to knowledge, with "know" substituted for "believe" in the above statements.

[15]Transitivity is the condition that for all $x, y, z \in \Omega$, if xR_iy and yR_iz, then xR_iz. In our setting, it is equivalent to the condition that $\forall\omega, \omega'$ $(\omega' \in P_i(\omega) \rightarrow P_i(\omega') \subseteq P_i(\omega))$. It is also equivalent to the condition that for all $E \subseteq \Omega$, $B_iE \subseteq B_iB_iE$. Euclideanness (also called "right euclideanness") is the condition that for all $x, y, z \in \Omega$, if xR_iy and xR_iz, then yR_iz. It is equivalent to the condition that for all $\omega, \omega', \omega'' \in \Omega$, $\omega', \omega'' \in P_i(\omega) \rightarrow \omega'' \in P_i(\omega')$. It is also equivalent to the condition that for all $E \subseteq \Omega$, $\neg B_iE \subseteq B_i(\neg B_iE)$.

[16]These frames are "partitional" because $\mathcal{P}_i = \{P_i(\omega)|\omega \in \Omega\}$ partitions the state space.

[17]In this paragraph and the next I use "knowledge" in the main text. But my observations about models of knowledge apply, *mutatis mutandis*, to models of belief, and the reader should keep this in mind. (This is true generally for the remainder of the section.)

three further assumptions, of "logical omniscience" for belief and knowledge. First, by the definition of knowledge in these models, for any proposition an agent knows, she knows every proposition entailed by it. Second, if an agent knows E and knows F, she knows $E \cap F$; agents' knowledge is closed under conjunction. Finally, every agent trivially knows or believes the tautology, Ω, since every world she can see is an element of the universe. If worlds are interpreted as logically possible situations, it follows that each agent knows every logical truth.

All but a small minority of philosophers reject the claim that people's belief and knowledge exhibit logical omniscience of this kind.[18] Even computer scientists and economists, with their more robust appetite for idealization, have recognized that these assumptions are too strong.[19] But I will not contest these assumptions in what follows. Since logically omniscient agents cannot disagree on the basis of mistakes about logic, it is *harder* and not easier for them to disagree. My counterexamples to the key conditional will be more powerful because they exhibit logically omniscient agents who, in spite of their omniscience, agree to disagree.

In addition to single-agent notions of belief and knowledge, we wish to describe relationships between the knowledge and belief of a group of agents. The simplest such relationship is *mutual knowledge* (or belief). A group *mutually knows* (or, mutually knows^1) a proposition E at a world ω if and only if every agent in the group knows E at ω.[20] A group mutually knows^n a proposition E if and only if the group mutually knows that they mutually know^{n-1} E. The group *commonly knows* E if and only if, for all natural numbers n, they mutually know^n it.[21]

In the finite models we will be considering here, the definition of common knowledge as an infinite conjunction is provably equivalent to a more elegant one, in terms of what are called *self-evident propositions.*[22] A proposition E is *self-evident* to an agent i if and only if, if E obtains, i knows that it obtains.[23] When the P_i are interpreted as representing knowledge self-evident propositions are akin to what are called "luminous" propositions in Williamson (2000: 95).[24] Now call a proposition *public to a group* $G \subseteq N$ if and only

[18]The exception is Stalnaker 1999: 241-73.
[19]See Halpern and Pucella 2011 for a recent survey.
[20]In symbols: $\forall i \; i \in G \rightarrow P_i(\omega) \subseteq E$, where $G \subseteq N$.
[21]The formalism for common knowledge was developed independently by Friedell 1969, Lewis 1969 and Aumann 1976.
[22]The explicit statement is due to Monderer and Samet 1989, but the idea is present already in Friedell 1969, Aumann 1976, and in a less formal way, in Lewis 1969: 52-7.
[23]For all $\omega \in E$, $P_i(\omega) \subseteq E$. Equivalently, E is self-evident to i if and only if there are no $\omega \in E$,$\omega' \in \Omega \backslash E$ such that $\omega R_i \omega'$. Again, belief could have been used here.
[24]Williamson 2000 used "in a position to know" in place of "know".

if it is self-evident to all agents in the group.[25] Using these definitions, we can show that a proposition F is commonly known by a group G at a world ω if and only if there is some public proposition E such that, for each i, E entails that i knows F.[26]

3.2.1. Qualitative Agreement. Aumann's original theorem has been shown to be a special case of a more general theorem using "decision functions". Formally, a decision function $d : 2^{\Omega} \to A$, where A is a set of (possibly infinitely many) "actions", takes propositions as arguments, and yields abstract "actions".[27] On the standard interpretation, if $P_i(\omega) = E$, with $E \subseteq \Omega$, then at ω, the agent i does or should take the action $d(E)$. In spite of their name, "decision functions" need not have anything to do with conscious "decisions". On the standard interpretation, they simply represent some feature of the situation which is fully determined by the conjunction of all of the propositions to which a given agent stands in the relation (for example, *believes* or *knows*) represented by the possibility correspondence.

I will discuss a specific interpretation of decision functions. Consider the following thesis:

Evidential Determination: There is some relation to propositions, such that one's evidence supervenes on or is determined by the set of propositions to which one stands in this relation.[28]

Evidential Determination does *not* require that one's evidence be a proposition or set of propositions, but merely that one's evidence supervenes on or is determined by some relation to propositions.

We wish to use our models to represent a class of agents who satisfy Evidential Determination. Evidential Determination says nothing about the structure of the set of propositions to which one stands in the evidence-determining relation; this set of

[25]A proposition E is self-evident if and only if for all $i \in I$,$\omega \in E$,$P_i(\omega) \subseteq E$. Equivalently, a proposition E is self-evident if and only if there is no $i \in I$, such that for some $\omega \in E$, $\omega' \in \Omega \setminus E$, $\omega R_i \omega'$. The relationship to common knowledge is made more transparent by considering a third equivalent definition. Let R^* be the transitive closure of the union of the R_i for $i \in I$. Then a proposition E is public just in case, for all $\omega \in E$, $\omega R^* \omega' \to \omega' \in E$.

[26]Formally, F is commonly known (or: believed) at ω if and only if there exists a public proposition E, such that $E \subseteq \{\omega' \in \Omega : P_i(\omega') \subseteq F\}$. Note that, in models of knowledge, where **(T)** is assumed, if E is public, then for all $i \in N$, $E \subseteq \{\omega' \in \Omega : P_i(\omega') \subseteq F\} \leftrightarrow E \subseteq F$. So when we assume **(T)** in these models of knowledge, we can simplify the definition as follows: F is commonly known at ω if and only if there exists a public proposition E, such that $E \subseteq F$.

[27]Decision functions were introduced to the literature on Agreement by Cave 1983 and Bacharach 1985.

[28]A version of this view is explicitly endorsed by Williamson (2000: Ch.9) and Conee and Feldman 2004 Ch. 9.

propositions might even contain contradictory pairs. But in this paper, we will not countenance such unhappy sets of evidence. We will use the P_i in the Hintikka-Kripke frame to represent the evidence-determining relation. In particular, we will say that an agent's evidence at ω is determined by the propositions which include $P_i(\omega)$, or, in symbols by the set $\{E \in 2^\Omega : P_i(\omega) \subseteq E\}$. This representation means (as we saw above) that the set of propositions which determines an agent's evidence will have special properties: the propositions in it will be logically consistent (**(D)**); the set will be closed under semantic entailment and under conjunction; and the set will contain all logical truths. Although this representation of Evidential Determination is somewhat restrictive, in what follows, when I say that a model represents Evidential Determination, or that something follows from Evidential Determination, I mean this mathematical implementation of the thesis. The implementation has one very convenient feature. Since specifying a single proposition is enough to specify the whole set of propositions, we can use a single proposition as a proxy for the set.

The project is now to use decision functions to represent the conjunction of Evidential Determination with the Uniqueness Thesis. To this end, we consider modified decision functions, $d^* : 2^\Omega \times 2^\Omega \to A$, functions from ordered pairs of propositions to actions. On this interpretation, the first element of these ordered pairs is understood to be the proposition which stands proxy for a given body of evidence. The function takes as argument a pair consisting of: the proxy for a body of total evidence and a proposition—and yields an element of A, here understood as a doxastic attitude. This interpretation allows us to represent the Uniqueness Thesis. According to the Uniqueness Thesis, if one's total evidence is E, then for every proposition, there is a uniquely rational attitude which one should have to that proposition. Thus we understand the function d^* as taking evidence, proposition pairs as arguments, and delivering the uniquely rational attitude one should have to the relevant proposition, given that body of evidence.

On this interpretation, models with qualitative decision functions represent views which incorporate the conjunction of *Evidential Determination* and the *Uniqueness Thesis*. The decision function itself takes one's total evidence (or a proxy for one's total evidence) to a function which describes, for every proposition, the uniquely rational doxastic attitude one should have to that proposition. When this interpretation is in view, we call decision functions *evidential response functions.*

Without any further constraints on decision functions, we can say very little. But even weak structural constraints allow us to start proving interesting results. A decision

function d satisfies the *formal sure-thing principle* if and only if, if $d(E) = d(F) = a$ for some action a, where E and F are disjoint ($E \cap F = \emptyset$), then $d(E \cup F) = a$. Cave (1983) and Bacharach (1985) first showed that:

THEOREM (Qualitative Agreement Theorem). *If all members of a group of agents have partitional ($S5$) knowledge and use the same sure-thing decision function, then if they commonly know the value of each others' decision functions, the value of the decision functions is the same.*

To suit the application which interests us, we must modify this theorem in two minor ways. First, an evidential response function d^* satisfies the *evidential sure-thing principle* if and only if, for all $E' \subseteq \Omega$, if $d^*(E, E') = d^*(F, E') = a$ for some attitude a, where E and F are disjoint ($E \cap F = \emptyset$), then $d^*(E \cup F, E') = a$. Second, whereas the statement of Bacharach and Cave's uses knowledge, we simply require that the *evidence determining attitude* (whether it be knowledge or not) obey $S5$. We can now restate the theorem as follows

THEOREM (Modified Qualitative Agreement Theorem). *Suppose each member of a group of agents has an evidence-determining relation which induces a partition on the space of worlds, and that all members of the group use the same evidential sure-thing evidential response function. Then if they commonly know which attitude each should take to a given proposition, they should all take the same attitude to that proposition.*

The proof of Bacharach and Cave's theorem proceeds by way of a key lemma, shared by a wide range of agreement theorems. Discussion of this key lemma will illuminate the structure of this and other similar results.

A decision function is *constant* for an agent on a proposition $E \subseteq \Omega$ if, for every $\omega, \omega' \in E$, the decision function has $d(P_i(\omega)) = d(P_i(\omega'))$. The key lemma states that, if the decision function is constant for an agent on a proposition E, which is self-evident to that agent, then the value of $d(P_i(\omega))$ for any $\omega \in E$ is equal to the value of the decision function on the self-evident proposition itself, that is, $d(E)$.

In Bacharach and Cave's theorem, the key lemma is particularly transparent. Recall that a proposition E is self-evident to an agent i if and only if, for every $\omega \in E$, $P_i(\omega) \subseteq E$. In Bacharach and Cave's models, each agent's knowledge is partitional. So, for any agent, any self-evident proposition can be written as the disjoint union of elements of the agent's partition. (This follows from the fact that the models are partitional or $S5$.) Now suppose

that the decision function is constant for an agent on a self-evident proposition. The formal sure-thing principle says that if the function takes the same value on each of a pair of disjoint sets, it takes the same value at their union. So it follows by this principle that the decision function has the same value on the whole of the self-evident proposition E, as it does on each $P_i(\omega)$, for $\omega \in E$.

With this key lemma in hand, the proof of the theorem is merely a question of transforming definitions. The antecedent of the key conditional says that the value of the decision function for each agent is commonly known. Thus, by hypothesis and the definition of common knowledge, for each agent, the decision function is constant for that agent on a *public* proposition. Now (once again, by definition) public propositions are self-evident to all agents. We choose the smallest non-null public proposition F which contains the actual world.[29] The key lemma shows that, for all agents, $d(P_i(\omega)) = d(F)$, for all $\omega \in F$. Since all agents use the same decision function, and F is the same for all of them, they all take the same action.

3.2.2. Beyond $S5$. A concrete form of *Evidential Determination* is advocated by Williamson (2000: Ch. 9): that one's evidence is what one knows. On this version of *Evidential Determination*, the assumption that agents satisfy $S5$ with respect to the evidence-determining relation is implausible. Philosophers of all stripes reject the claim that knowledge satisfies the negative introspection principle (**(5)**) which is validated by partitional models (Hintikka 1962; Williamson 2000: 23-27; Stalnaker 2009: 400). Thus if knowledge (or in fact any factive attitude) underwrites the evidence-determining relation, the modified version of Bacharach and Cave's result is of limited interest for our purposes.

If we strengthen the evidential sure-thing principle, however, we can go beyond partitional knowledge, and prove a different result, using a weaker epistemic logic.[30] This follows from a more general principle: stronger constraints on decision functions allow us to weaken the epistemic logic and preserve the claim that, if agents commonly know what each will do (according to the shared decision function), then they will take the same action (as given by that decision function).

---

[29]There is guaranteed to be such a proposition because common knowledge, like each individual's knowledge, is closed under conjunction. The smallest commonly known proposition is thus the conjunction of all propositions which are commonly known. (Since the models have finitely many propositions, this is nonempty.)

[30]Even using only the sure-thing principle, we can improve the Agreement Theorem by using *local decomposability* in place of $S5$. A possibility correspondence P_i is locally decomposable with respect to a set E if and only if there is a set $F \subseteq E$ such that, for all $\omega, \omega' \in F$, $P_i(\omega) \cap P_i(\omega') = \emptyset$ and $\bigcup_{\omega \in F} P_i(\omega) = E$. A possibility correspondence is locally decomposable *simpliciter* if it is locally decomposable with respect to all self-evident sets. We can prove that all locally decomposable possibility correspondences cannot agree to disagree on a sure-thing decision function.

An example will help to make the point, and also focus attention on a key difficulty about qualitative Agreement results. A decision function satisfies the *strong sure-thing principle* if and only if, if $d(E) = d(F) = a$, and if $E \cap F = \emptyset$, or $d(E \cap F) = a$, then $d(E \cup F) = a$.[31] The strong sure-thing principle strengthens the formal sure-thing principle because it has consequences even if the relevant propositions E and F are not disjoint. Using only $S4$ for the evidence-determining attitude, we can prove an Agreement Theorem for decision functions that satisfy the strong sure-thing principle.

THEOREM 3.2.1. *If a group of $S4$ agents have the same decision function, which satisfies the strong sure-thing principle, and if each agent's action is commonly known among the group, then they all take the same decision.*[32]

Once again, we can modify the definition of the strong sure-thing principle in a straightforward way to apply to evidential response functions. An evidential response function satisfies the *strong evidential sure-thing principle* if and only if, for all $E' \subseteq \Omega$, if $d(E, E') = d(F, E') = a$, and if $E \cap F = \emptyset$, or $d(E \cap F, E') = a$, then $d(E \cup F, E') = a$. If we fix a proposition $E' \subseteq \Omega$, we can use Theorem 3.2.1 to show that the agents cannot agree to disagree *about* E'. Since E' was chosen arbitrarily, Theorem 3.2.1 has, as an immediate corollary, that $S4$ agents who obey an evidential response function which satisfies the strong evidential sure-thing principle cannot agree to disagree in their attitude to any proposition.

3.2.3. The Sure-Thing Principle: A Problem for the Argument from Qualitative Agreement. In the Introduction (Section 3.1), I stated an argument from Agreement directed against Bayesian Uniqueness. The Modified Qualitative Agreement Theorem and Theorem 3.2.1 suggest a more general argument, directed at the Uniqueness Thesis itself, and not merely the Bayesian version of the thesis. The argument runs as follows. If the rational evidential response function satisfies the strong evidential sure-thing principle, then by Theorem 3.2.1, the Uniqueness Thesis is inconsistent with rational agreeing to disagree among $S4$ agents. But, as our earlier examples of scientists and jurors showed, such disagreements seem to be possible. So the Uniqueness Thesis is false.

[31] Note for specialists: this condition is implied by *preservation under difference plus the sure-thing principle* (Rubinstein and Wolinsky 1990). It is also implied by *preservation under union*. But it does not imply either of these conditions.

[32] The proof of this theorem is given in the Appendix.

This apparently powerful argument, however, has two problems. The first of these problems is particular to the argument from Qualitative Agreement. This new argument has a new premise: that the rational evidential response function satisfies the (strong) evidential sure-thing principle.

The sure-thing principle is usually motivated by stories like the following, taken from Savage:[33]

> A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness of the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate were going to win, and again finds that he would do so. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say.

Savage concludes this example by observing that: "It is all too seldom that a decision can be arrived at on the basis of the principle used by this businessman, but, except possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance."[34]

Savage's principle, in spite of its putative "ready acceptance", is subject to counterexamples. Here is a striking one.[35] The businessman considers whether he should place a bet on the outcome of the election with a certain bookie who charges a small fee on each bet. The businessman's current subjective probabilities about the outcomes of the election match the bookie's odds exactly. So the businessman should not place a bet: the bookie's fee is greater than his nil expected gain by placing a bet, whether he bets on the Republican or on the Democrat. But the businessman has read his Savage. He considers what he would do if he knew the Republican will win. In this case, if he bet, he would bet on the Republican's victory, and would win considerably more than he would lose by paying the bookie's fee. So he would place a bet. The businessman then considers what

[33]Bacharach 1985: 181; Aumann, Hart and Perry 2005 (unpublished): 8; Samet 2010: 171.
[34]Savage 1954: 21.
[35]Thanks to Tim Williamson and an anonymous referee for this example. It bears some structural similarities to Parfit's miners example (1988 (1983), recently revived for a different purpose by Kolodny and MacFarlane 2010), though Parfit originally used his example to exhibit the difference between subjective and objective "oughts", not as a counterexample to the Sure-Thing Principle. For another related example, see Aumann, Hart and Perry 2005 (unpublished): 10.

he would do if he knew the Democrat will win. Once again, in this case, his bet would win, so he would place a bet. If he follows Savage's principle, he should place a bet, in spite of the expected loss of that action, given his current uncertainty.

It is controversial whether this example can be extended to the formal version of Savage's principle. But we need not settle this here. The point in the present context is merely to dramatize the fact that Savage's argument for his principle is not enough. If the argument from Qualitative Agreement is to be effective, the opponent of the Uniqueness Thesis will have to offer a new argument to show that the rational evidential response function satisfies (for example) the strong sure-thing principle.[36]

This was the first problem with the argument from Qualitative Agreement. The second problem—my countermodels—is in a sense the target of the remainder of the paper. The countermodels of Sections 3.4−3.7 will show that agreeing to disagree is possible among agents who update by conditionalization on a common prior. Update by conditionalization satisfies the strong evidential sure-thing principle.[37] And since agents update on a shared prior, the agents in these models have the same evidential response function. So the later countermodels will also be countermodels to the more abstract argument from Qualitative Agreement. Even if the opponent of the Uniqueness Thesis were to establish that the rational response function satisfies the strong sure-thing principle, she would still have more work to do. She would still have to rule out the four classes of counterexamples described in the remainder of the paper.

I will now present the theorem for Bayesian update, before turning to these countermodels.

---

[36]It is perhaps worth noting that a putative and much discussed problem with the formal sure-thing principle, first stated by Moses and Nachum (1990, Lemma 3.2), does not apply on our interpretation of decision functions (for recent discussion, see especially Aumann and Hart 2006 (unpublished); Aumann, Hart and Perry 2005 (unpublished); Samet 2010). The supposed problem runs as follows. In partitional frames, Qualitative Agreement theorems require an assumption that the shared decision function is defined on belief-sets or knowledge sets agents could not have. Suppose agent i has partitional knowledge. Thus for some E, the proposition that E is identical to the proposition that i knows E. Similarly, if i's partition is not the trivial one, there is some $F \neq E$, such that F is identical to the proposition that i knows F. If i knows E, she does not know F (since $E \cap F = \emptyset$; they are incompatible), and, by negative introspection, she knows that she does not know F. By the same argument, if she knows F, she does not know E, so she knows that she does not know E. The sure-thing principle says that, if i would take a given action if she knew that E, and if she would take the same action if she knew that F, then she would take the same action if she knew only that (E or F). But if i knew only that (E or F), she would know that (she knows that E or she knows that F). But, by negative introspection, she would also know that (she does not know E *and* that she does not know F), which is a contradiction. But in our interpretation of the formalism, this problem does not apply. If a decision function is the action *a given agent* would or should take, given certain knowledge, then contradiction looms. But the Uniqueness Thesis is an *interpersonal* principle. The evidential response function, on this interpretation, need not be sensitive to whether i (as opposed to j or k) could know the proposition which it takes as its argument. The response function merely gives a verdict about the doxastic attitude *someone* who had a given body of evidence should take, irrespective of which agents can or cannot have that body of evidence. The objection thus does not apply.

[37]This is proven in the appendix, as Fact 3.8.1.

3.3. Agreement on Posterior Probabilities

Aumann's original theorem concerns the specific decision function of update by conditionalizing on a common prior. To present this theorem, we must add probabilities to our frames. We expand the basic frame to include a prior μ, which is a probability measure over the finite space Ω, yielding an extended *common prior frame* $< \Omega, N, \{R_i\}_{i \in N}, \mu >$. For simplicity, we assume that μ assigns positive probability to every world in the frame; μ is *regular*. We continue to use the P_i to define a set of propositions, $\{E \in 2^\Omega : P_i(\omega) \subseteq E\}$. As before, we take this set to determine each i's evidence, in accord with Evidential Determination. Finally, the agents in these frames determine their "posterior" attitudes as Bayesian Uniqueness says they should: by conditionalizing the prior on the conjunction of the set of propositions to which they bear the evidence-determining relation. To simplify the English argumentation, when the attitude represented in the frames is factive, I interpret the frames as modeling knowledge, in line with the view of evidence in Williamson (2000: Ch. 9). I will interpret frames with a non-factive attitude as representing belief. The models, of course, do not force either of these interpretations. Those who have different views about the evidence-determining relation should substitute their preferred relation for my "know that" (if they hold that one can only bear this relation to true propositions) or "believe that" (if they think false propositions can be evidence, too).[38]

In common prior frames, an agent's "posterior" probability in a proposition $E \subseteq \Omega$ at a world $\omega \in \Omega$ can be defined as follows $\pi_i(\omega)(E) := \mu(E \cap P_i(\omega))/\mu(P_i(\omega)) = \mu(E|P_i(\omega))$.[39] It is now elementary to observe that this form of update by conditionalization obeys the strong evidential sure-thing principle (and thus, *a fortiori*, the formal sure-thing principle).[40] In Aumann's partitional frames, the Agreement Theorem for posterior probabilities obtained by conditionalizing on a common prior is thus a corollary of Bacharach and Cave's Qualitative Agreement theorem.[41] If we consider the class of

[38]Those who reject Evidential Determination in full generality should consider the set of cases (if any) where a restricted version of that thesis does hold. Any Bayesian will of course think that evidence is (at least in cases described by his or her Bayesianism) propositional, so the assumption should not be overly controversial in this context. (Thanks to an anonymous referee here.) Those who think evidence is never determined by a set of propositions in this way may have to engage in some exercises of the imagination.

[39]Since we assume that μ is regular, and that all of the $P_i(\omega)$ are nonempty, this value is always defined.

[40]This claim is proven as Fact 3.8.1 in the appendix, for completeness.

[41]In these frames, a world ω' is accessible from ω if and only if, at ω, ω' has positive probability. Probabilities march in step with the full attitudes represented by the accessibility relations. In the application under consideration (to Bayesian Uniqueness) this assumption holds by hypothesis. But in general, many deny that belief or knowledge entails full confidence or certainty. Especially when factivity fails, then, those who deny that belief requires full confidence should replace my "belief" with "full confidence" or "certainty".

$S4$ frames (and not merely the partitional ones), the Agreement Theorem for posterior probabilities in $S4$ frames is a corollary of the new Theorem 3.2.1.[42]

But this corollary is still not as general as the key conditional. The next four sections consider four further assumptions which are required to close the gap between the key conditional and the theorem.

3.4. Common Knowledge of Common Priors

In common prior frames, all agents use the same prior.[43] This assumption is often called the common prior assumption. But the name is misleading. The implementation of the common prior assumption has much stronger consequences than the mere assumption that agents have the same prior. In the class of models for which the *Agreement Theorem* is proven, agents not only have a common prior: they commonly know which prior they share, *and* commonly know that they update by conditionalizing on that prior.

These assumptions are not stated explicitly in the key conditional. Nor are they represented explicitly in Aumann's class of models, even though the models guarantee that they hold. When substantial assumptions are represented "behind the scenes" in this way, it is difficult to assess their strength. So in this section, I first introduce a class of models in which knowledge of the prior *can* be represented explicitly. (The next section turns to update by conditionalization.) I then show that, in these models, even if agents commonly know that they use the same prior, they may agree to disagree if they fail to commonly know which prior they share.[44] A group may commonly know that they share a prior, but fail to commonly know which prior they share, if some member of the group fails to know which prior she herself employs.[45]

Worlds represent the full domain of agents' uncertainty. To represent uncertainty *about* priors, we must relativize priors to worlds. A prior assignment function $p_i : \Omega \to \Delta(\Omega)$ takes each world to a probability distribution over worlds. To respect the notion that the prior is *regular*, we insist that the prior assigns positive probability to all $\omega \in \Omega$. The standard *common prior frame* is a special case of this class of models, where two

[42]In fact, we can prove a more technical result, which generalizes even this one, using a property I call "local balancedness", which is defined in the Chapter 5. Accordingly, the examples in Sections 3.6 and 3.7 will in fact give agents who not only fail to be positively introspective or have true beliefs, but do so in a way that makes them fail to be locally balanced.

[43]Thanks to Robin Hanson for correspondence about the models presented in this section.

[44]Throughout this section and the next, I use partitional frames in the examples. I thus speak of "common knowledge" and not "common belief". This is simply to hold fixed one aspect of the theorem while testing another.

[45]In a different idiom, the agents may fail to know *de re*, of a given prior, that it is the common prior. But they may still know *de dicto*, that they have the same prior. Thanks to an anonymous referee for this way of putting the point.

additional constraints are imposed: first, that for every pair of agents, i and j, $p_i = p_j$; second, that for every pair of worlds, ω,ω', $p_i(\omega) = p_i(\omega') = \mu$, where μ is the *common prior*.

To respect the view that the prior is the same for all agents, we will retain the first assumption from the common prior frames, that every player has the same prior assignment function. But to represent agents' uncertainty about the exact value of the prior, we relax the second assumption. An *uncertain common prior frame* is thus a tuple $< N, \Omega, (P_i)_{i \in N}, p >$, where p is the shared prior assignment function. As above, we stipulate that the value of agents' posterior on a proposition E at a world ω are given by $p(\omega)(E|P_i(\omega))$, in accord with Bayesian Uniqueness.[46]

This new class of models allows us to distinguish two kinds of "common knowledge of common priors": first, that agents know that they have the *same* prior (but not which one it is); second, that they know which prior they share. In uncertain prior frames, agents commonly know that they have the *same prior*, but they may fail to know which prior that is, since they may fail to know which prior they themselves have. Common knowledge of *which* prior they all share is thus strictly stronger than common knowledge that all have the same prior.

So we can use these models to test a possible response to the argument from Agreement for theorists of Bayesian Uniqueness. Proponents of Bayesian Uniqueness may well deny that properties of the prior are a matter of common belief (never mind common knowledge). For example, I do not believe the proposition that [the (evidential) prior assigns the proposition that space-time is supersymmetric probability $< .5$]. I also do not believe its negation. Perhaps I share this state with other, more rational agents. So we might wonder whether this kind of uncertainty can make room for the possibility of agreeing to disagree.

Before answering this question, we should consider how this reply might suit subjective Bayesians, as well. Without any restriction on priors, subjective Bayesianism is overly permissive to be a plausible normative theory. Many subjective Bayesians endorse a more restrictive theory, holding that cases of permissible heterogeneous priors are relatively rare, and considerably rarer than disagreements among agents who commonly know each

[46]For subjective Bayesians, these models represent a situation in which Bayesian agents commonly know that they have the same prior, whether on all events or merely on the events represented in the model. Note that, in finite frames, even if we relax the assumption that $p_i = p_j$, there will still be substantial common knowledge about which priors all agents *might* have. But if we move to infinite frames, and allow $p_i \neq p_j$, we can relax this common knowledge assumption significantly. For example, we can build models in which, for some proposition E, for any real number $x \in [0, 1]$, there is some world where agent i has a prior which has $p_i(\omega)(E) = x$. Thanks to an anonymous referee here.

others' probabilistic attitudes to a given proposition. The argument from Agreement would show that this position about the relative frequency of these occurrences is untenable. Still, like the theorist of Bayesian Uniqueness, the subjective Bayesian may believe that agents often fail to know which prior they themselves have. So once again we may ask: is this ignorance of one's own prior enough to recover the possibility of agreeing to disagree?[47]

It is. Failure to know one's own prior does provide a way out of the argument from Agreement, for theorists of Bayesian Uniqueness and for subjective Bayesians. Consider an uncertain prior frame with seven worlds ($\Omega = \{1, 2, 3, 4, 5, 6, 7\}$) and two agents, Leo and Ulrich, as depicted in Figure 3.4.1. At worlds $1, 2, 3$ and 4, both Leo and Ulrich have the prior μ_A (formally, if $\omega \in \{1, 2, 3, 4\}$, then $\mu_A = p(\omega)$). In worlds $5, 6$ and 7, however, they both have the prior μ_B (in symbols, if when $\omega \in \{5, 6, 7\}$, then $\mu_B = p(\omega)$). At every world, the agents' priors are equal to each other's, but at different worlds they have different priors from the prior they would have had in other worlds. The worlds also differ by what agents know if they are actual. Leo's partition is described by the two sets $\{1, 2, 3, 4\}, \{5, 6, 7\}$, while Ulrich's partition is given by $\{1, 2\}, \{3, 4, 5, 6, 7\}$. At every world, the agents commonly know that they have the same prior, but at every world they fail to commonly know which prior they have.

Now consider the proposition $A = \{1, 3, 7\}$. (In Figure 3.4.1, the elements of this proposition are circled.) Given his partition, Leo's posterior probability in this proposition

[47]Subjective Bayesians in philosophy often interpret priors and update by conditionalization literally, holding that priors are distributions agents have at some time, and that, at a subsequent time, they have a distribution obtained by conditionalizing on the earlier one. But subjective Bayesians may understand the combination of a common prior and update by conditionalization in other ways. For example, a number of economists have recently argued that the conjunction of these claims can be understood as a feature of the *consistency* of the beliefs of a group of agents. Yossi Feinberg (2000) has shown that under certain conditions the beliefs of a group of agents can be represented as deriving from a common prior if and only if the agents' present beliefs render them jointly immune from arbitrage by an outsider. (See also Bonanno and Nehring 1997, 1999, and Barelli 2009.) On this interpretation of the common prior, the prior is a mathematically tractable representation of the kind of doxastic consistency dramatized by immunity from arbitrage by an outside party. The vast and fast-growing literature on priors in economics includes numerous further characterizations of this assumption, which I cannot begin to do justice to here.

Since the models in the main text do not represent times, they are closely related to this interpretation of the common prior as a consistency assumption. But for subjective Bayesians who adopt a more literal approach to these assumptions, it is worth noting that my models can also be extended to directly represent their "literal" subjective Bayesian approach. Since this subjective Bayesianism takes conditionalization as form of genuine *update*, we must add a set of times to the models, T, a renaming of the natural numbers. We then define the possibility correspondences for each agent as $P_i : \Omega \times T \to 2^{\Omega \times T}$. Possibility correspondences now take world, time *pairs* as arguments, and produce subsets of the Cartesian product of the set of eternal propositions (worlds) with the set of times. We can now define the subjective Bayesian notion of a *prior* as the agent's probability distribution at time 1. A prior assignment function $p_i : \Omega \to \Delta(\Omega, t_1)$ takes each world to a distribution over histories (worlds) at the first time. We insist that the prior assigns positive probability to all pairs (ω, t_1), where $\omega \in \Omega$. In the simplest version of the models, agents are always certain of the time, though this assumption could be relaxed (earlier examples of related models can be found in Geanakoplos 1994: 1455-8, Heifetz 1996, and, from a different perspective, Hanson 2006).

(at every world) is 2/3, while Ulrich's posterior probability (again at every world, given his partition) is 1/2. Regardless of which world is actual, the agents' probabilities in A differ, in spite of the facts that the values are commonly known, that they commonly know that they have the same prior, and that they commonly know that their attitudes are given by conditionalizing the prior on what they know. If agents do not commonly know *which* prior they have, the key conditional may fail.

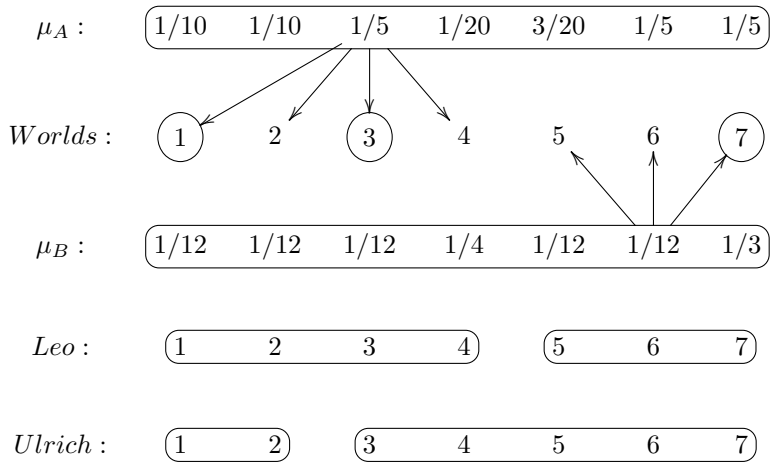


FIGURE 3.4.1. Agents Commonly Know that they Share a Prior, but still Agree to Disagree

3.5. Common Knowledge of Update by Conditionalization

In addition to common knowledge of a common prior, standard models (as well as the class of frames just introduced) assume informally that agents commonly know that they update by conditionalization. The proponent of Bayesian Uniqueness holds that rational agents form their beliefs by conditionalizing the conjunction of their evidence on the prior. But she is not committed to the claim that, for any two rational agents, they *commonly know* that each of them updates according to this rule. The subjective Bayesian may similarly believe that even in cases where a group of agents do commonly know that they began with the same probabilistic beliefs about a given subject matter, they do not commonly know that they have continued to update these beliefs correctly.

Once again, our first task is to make this common knowledge assumption explicit. We add a new set of functions $\pi_i : \Omega \to \Delta(\Omega)$, one for each $i \in N$, which describe each agent i's *posterior probabilities* at worlds. Crucially, these new posterior probabilities are no longer guaranteed to accord with what one would obtain by conditionalizing on the prior. To maintain the tight connection between probabilistic attitudes and the full attitudes of belief and knowledge, we still require that, for all i, for all $\omega, \omega' \in \Omega$, $\pi_i(\omega)(\omega') > 0$

if and only if $\omega' \in P_i(\omega)$.[48] Now we say that an agent is an *actual Bayesian* at ω if her posteriors or degrees of belief are what would be obtained by conditionalizing her prior at ω on what she knows at ω.[49] An agent i is commonly known at ω to be Bayesian if it is common knowledge at ω that i is an actual Bayesian.

These new models allow us to show that if one denies that agents commonly know that each of them satisfies the stringent demands of Bayesian Uniqueness, the argument from Agreement can be avoided. Agents who do not commonly know that they update by conditionalization may agree to disagree, even if, in the actual world, each of them does in fact update by conditionalization on a shared prior.[50] Consider the following frame, with six worlds and two agents, Arnheim and Diotima. We make the strong assumption that Arnheim and Diotima commonly know which prior they share, namely, the uniform prior (which we call μ, for the remainder of the example) over the six worlds in the model. Moreover, each agent updates on a partition. Here, Arnheim's partition is given by $\{1,2,3\},\{4,5,6\}$, while Diotima's partition is given by $\{1,2\},\{3,4,5,6\}$. Arnheim is, naturally, an actual Bayesian at all worlds. Diotima, however, is an actual Bayesian at worlds 1 and 2, but not at worlds $3,4,5$ and 6. At these worlds she assigns probability $1/2$ to world 6, and $1/6$ to each of $3,4$ and 5. Now it is commonly known at world 1 that Arnheim assigns the event $\{1,6\}$ probability $1/3$, and commonly known that Diotima assigns the same event probability $1/2$. If Diotima were a Bayesian in worlds $\{3,4,5,6\}$, then Arnheim would not even *know* (never mind commonly know) Diotima's posterior probabilities in the proposition $\{1,6\}$. But since Diotima is not a Bayesian in those worlds, Arnheim and Diotima commonly know Diotima's posterior probabilities. In spite of this common knowledge, the agents' posteriors differ. They agree to disagree, even though they are both *actual* Bayesians at world 1, and commonly know that they have the uniform distribution as a prior.

The finite examples in this section and the last have an artificial feel. When we are uncertain about our priors (or present probabilistic beliefs) in a given proposition, we are usually uncertain over an interval of probability assignments to that proposition. For example, a physicist may know that her present beliefs assign probability $> .9$ to the hypothesis that the big bang initiated the universe as we know it, but she may be uncertain over the open interval $(.9,1)$ about the probability she herself assigns to

[48]Given this condition, plus the assumption that the π_i are regular, and Ω is, as usual, finite, we can simply take the frame to be $< N, \Omega, (\pi_i)_{i\in N} >$. We can then treat the P_i as derivative, defined as follows: $P_i(\omega) := \{\omega' : \pi_i(\omega)(\omega') > 0\}$.

[49]In symbols, if $\pi_i(\omega) = p(\omega)(\cdot|P_i(\omega))$.

[50]Aumann also recognizes that his theorem requires common knowledge of Bayesianism (1998: 929 n.2).

this proposition. These infinite cases, while more familiar and more natural, require the introduction of technicalia which are irrelevant to the present, conceptual point: that a certain form of uncertainty about priors, and uncertainty about other agents' update rules can lead to countermodels to the key conditional, and a way out of the argument from Agreement for the theorist of Bayesian Uniqueness.

3.5.1. Pointed Models. The new classes of models introduced in the previous two sections allow us to make precise an informal observation made in the introduction. There, I drew the distinction between assumptions holding at the actual world, and assumptions holding at all worlds in the state space. To formalize this distinction, we can use "pointed" frames. We let an *Aumann frame* be a structure $\langle N, \Omega, (P_i)_{i\in N}, (p_i)_{i\in I}, (\pi_i)_{i\in I}\rangle$, including all of the new elements introduced in this section and the last. A pointed Aumann frame is then a tuple $\langle N, \Omega, a, (P_I)_{i\in N}, (p_i)_{i\in I}, (\pi_i)_{i\in I}\rangle$ where $a \in \Omega$. This distinguished a is interpreted as the actual world. In such a pointed model, the natural form of some of Aumann's main assumptions is:

Common Prior: For all $i, j \in N$ $p_i(a) = p_j(a)$

Update by Conditionalization: For all $i \in N \pi_i(a) = p_i(a)(\cdot|P_i(a))$

(actual-4): For all $i \in N$, $\omega \in P_i(a) \rightarrow P_i(\omega) \subseteq P_i(a)$

(actual-T): For all $i \in N$, $a \in P_i(a)$.

It is now easily checked that, given the appropriate specification of the actual world, the previous two examples satisfy all of these axioms. (In the first example, we can let a be any of the worlds in the frame; in the second, we let $a = 1$.) The examples of the next two sections will similarly satisfy all four of the "pointed" versions of the axioms. What fails in these later examples are Aumann's stronger, global versions of the axioms.

The example from this section also exhibits a theme which will recur in the next two sections. The common prior assumption is explicitly an assumption about the relationships among the beliefs of many agents, so it may be unsurprising that it leads to "social" consequences. (What is surprising is that common knowledge can do so much of the work.) But the assumption that all agents update by conditionalization does not appear to be an "interpersonal" assumption of this kind. And yet the assumption that all agents update by conditionalization is implemented in such a way that the models yield a further, stronger assumption: that agents *commonly know* that each of them updates by conditionalization. This stronger assumption does explicitly make reference to the beliefs (knowledge) of many agents. And, as our example shows, this implementation

of the assumption brings with it an important social consequence: the impossibility of agreeing to disagree.

Pointed models, and the pointed axioms just introduced, do not have this kind of consequence for what the agents in the model know. At least in part for this reason, they do not yield social consequences such as the Agreement Theorem. In the next two sections, we will see that the same point holds for the remaining two of our four target assumptions.

3.6. Inexact Knowledge: Failures of KK

The examples in the previous two sections demonstrate that natural weakenings of the assumption that agents commonly know which prior they share, or of the assumption that agents commonly know that they each live up to the stringent standards of Bayesian Uniqueness, can lead to counterexamples to the key conditional. But, as the next two sections will show, a theorist of Bayesian Uniqueness who is attracted to these assumptions has still other ways to save the phenomenon of disagreement among agents who commonly know one another's attitude to a proposition. The following example shows that certain failures of positive introspection or "KK" for knowledge can induce failures of the key conditional, even given Bayesian Uniqueness, common knowledge of the prior, and common knowledge that agents update by conditionalizing.

A subject in an experiment, Clarisse, is to be presented with a rectangular object at a distance.[51] Clarisse will attempt to assess the object's height and width by looking at it. Because of the distance, Clarisse cannot tell by looking exactly what the height and width of the object are. In particular, she cannot discriminate differences of less than .6 centimeters in height or width. Clarisse's discriminatory limitations impose limits on Clarisse's knowledge:

> **Margins For Error:** If the rectangle is n centimeters tall (or wide),
> Clarisse does not know that it is not $n \pm .6$ centimeters tall (wide).[52]

Clarisse knows that the experimenters will choose one of eight scenarios according to the roll of a fair, eight-sided die. Six of the sides of the die are labeled with the dimensions of a rectangle; if one of these sides come up, the experimenters simply display the relevant rectangle where Clarisse will see it. Clarisse knows that this rectangle will be chosen from a set of objects with dimensions varying between $10cm \times 10cm$ and $11.5cm \times 11cm$, at

[51] Thanks to Tiankai Liu and Yan Zhang for showing me a model on which this example is based.
[52] This is modeled directly on (I_i) from Williamson (2000: 115; Cf. 1992).

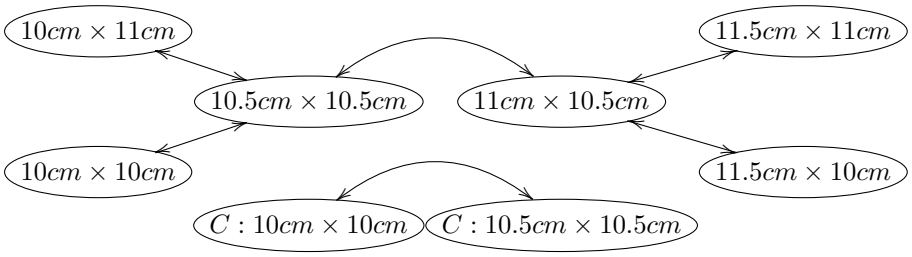


FIGURE 3.6.1. Clarisse's Possibility Correspondence in the Rectangle Experiment

increments of $.5cm$ on each dimension. But Clarisse also knows the experimenters will not choose from all twelve of these rectangles; they have left out six. In particular, they have not included $10cm \times 10.5cm$; $10.5cm \times 10cm$; $10.5cm \times 11cm$; $11cm \times 10cm$; $11cm \times 11cm$; and $11.5cm \times 10.5cm$. So they will choose only from $10cm \times 10cm$; $10cm \times 11cm$; $10.5cm \times 10.5cm$; $11cm \times 10.5cm$; $11.5cm \times 10cm$; and $11.5cm \times 11cm$. The possibilities from which they may choose are represented in the following figure.[53]

The final two sides of the die are labeled, respectively $C : 10cm \times 10cm$, and $C : 10.5cm \times 10.5cm$. If one of these two sides comes up, the experimenters will display a large card which states in clear letters that the chosen object was from the set $\{10cm \times 10cm, 10.5cm \times 10.5cm\}$. In this case, let us suppose that Clarisse's margins for error are irrelevant; in this outcome, she satisfies both positive and negative introspection for knowledge.

Suppose now that a second agent, Walter, also knows the setup of the experiment. Like Clarisse, he knows that the experimenters will roll their eight-sided die to determine what Clarisse will experience. Within these constraints, Walter is maximally ignorant about which outcome has been chosen. Perhaps he is seated with his back to the wall where Clarisse will see the rectangle or the card.

Now consider the proposition that the object which was presented was one of the extreme objects: $\{10cm \times 10cm, 10cm \times 11cm, 11.5cm \times 10cm, 11.5cm \times 11cm, C : 10cm \times 10cm\}$. Call this proposition *Extremities*. Regardless of which object is presented, Clarisse's posterior probability in *Extremities* is 1/2. Regardless of which object is presented, Walter's posterior probability in *Extremities* is 5/8.[54]

[53]Note that in this example, both agents have symmetric accessibility relations. The frame therefore satisfies **(B)**: for all i, $\forall\omega, \omega' \in \Omega$ $(\omega \in P(\omega') \leftrightarrow \omega' \in P_i(\omega))$, or, equivalently, for all i, for all $E \subseteq \Omega$, $E \subseteq B_i \neg B_i \neg E$.

[54]In the main text, we are considering probabilities as they are understood by the theorist of Bayesian Uniqueness, but little turns on this in the example. Walter and Clarisse may well have calibrated their subjective probabilities to match the distribution from which they know the rectangle will be drawn. If they update by conditionalizing on the conjunction of the propositions to which they bear the evidence-determining relation, then the argument for the difference in subjective probabilities is the same as the one given for the "objective" probabilities of the theorist of Bayesian Uniqueness.

We may suppose that Clarisse and Walter commonly know the set-up of the situation. We may also suppose that they commonly know what Clarisse will learn by seeing any given object, and that Walter will learn nothing when the object appears. Finally, there is no harm in supposing that, as the good intellectuals they are, they have performed all relevant deductions.

But Clarisse's inexact knowledge leads Walter and Clarisse to agree to disagree. At every world in the space, Clarisse assigns *Extremities* 1/2. Similarly, at every world in the space, Walter assigns *Extremities* 5/8. Since the space represents a proposition they commonly know, they commonly know what probability each assigns to *Extremities.* And yet they assign this proposition unequal probabilities.[55]

The example of Walter and Clarisse extends the example of Mr Magoo from Williamson (2000: 116-123) to two dimensions. Clarisse's correspondence satisfies **(T)**, but she fails "KK" or **(4)** for knowledge.[56]

The Margins-for-Error premise is controversial. But the exact machinery of Margins-for-Error is not required for the point of the present section. First, failures of an unrestricted, simple-minded KK principle abound for mundane reasons. Most obviously, on a natural picture of belief, we often fail to believe that we know a proposition, even if we know it. (Some agents who seem to able to know propositions may not even be able to form beliefs about knowledge.) And if, as many philosophers maintain, knowledge requires belief, then our failure to have beliefs about our knowledge will lead directly to failures of this simplistic KK principle. The Margins-for-Error premise gives us a simple structural constraint which generates failures of KK of the relevant form. But other failures of KK could have been used to generate an example which has the same structure as the example of Walter and Clarisse. And the structure of the example is all that is required to yield a countermodel to the key conditional.[57]

---

[55]Note that the two C : or "card" worlds are unnecessary to generate agreeing to disagree. If we deleted them from the frame, Clarisse would still assign *Extremities* 1/2 at every world, while Walter assigned it 2/3. The "card" worlds are only needed to provided worlds where Clarisse *actually* satisfies the pointed axioms.

[56]In fact, as noted earlier, at least one correspondence must also fail to be "locally balanced".

[57]If KK fails, how can there be common knowledge at all? It is true that, for any proposition i and j commonly know, i also knows that i knows that i knows that...that proposition. But as in the example above, the agents may commonly know one proposition but nevertheless fail to be positively introspective about another proposition. So there is no immediate inconsistency between failures of KK and the presence of common knowledge. Some motivations for denying KK do lead naturally to the denial of the possibility of common knowledge full stop (see, e.g., Greco 2014a: 169-70 cf. 179-80). For example, if one is attracted to margins-for-error principles in general (and not just for knowledge of the dimensions of objects seen at a distance), one may accept a version of the margins-for-error premise for sorites series involving knowledge itself. If, as a result, one denies the possibility of infinitely iterated knowledge of any "nontrivial" propositions, then one is committed to the impossibility of common knowledge. But this is a particular feature of a particular view about KK failures. And there seem to be a variety of consistent

In fact, failures of the KK principle are unnecessary in a further way. Suppose that the actual outcome of the above experiment is $C : 10cm \times 11cm$. In this outcome, Clarisse actually satisfies **(4)** (in fact, she also satisfies **(5)**). In the presentation just given, Clarisse and Walter can agree to disagree because of certain outcomes where Clarisse *might* have failed **(4)**. In this presentation, although Clarisse actually satisfies **(4)** at this world (and in fact both agents satisfy all of Aumann's informal assumptions there), a "real" failure of this axiom was required to produce the example.

But we could have presented a different story for our eight world model. Poor Walter might simply be a philosopher who has read Stalnaker and Williamson, and cannot make up his mind as to whether knowledge satisfies **(4)**. while Clarisse, for her part, is a stalwart economist or computer scientist. We may suppose that Clarisse is right: she does in fact satisfy the constraints of partitional models of knowledge. Still, because of Walter's uncertainty about the axioms governing knowledge, he assigns some probability to Clarisse's knowledge forming a partition, and some probability to her requiring a Margin for Error. Thus, Clarisse and Walter can agree to disagree. And this, once again, without any actual failures of Aumann's intuitive axioms.

This example shows that the theorist of Bayesian Uniqueness (and also subjective Bayesians) can avoid the argument from Agreement by rejecting positive introspection for knowledge (or the evidence-determining attitude). In fact, they may respond to it without rejecting positive introspection in the actual state, but simply allowing that agents may assign positive probability to the proposition that positive introspection fails.

This observation brings us back to the theme introduced at the close of the last section. The KK assumption itself appears to be a primarily "single-agent" principle. But in standard $S4$ models of knowledge, every agent is assumed to satisfy positive introspection at every world in the model. On this implementation of the assumption, the agents in the model commonly know that each of them satisfies KK or **(T)** and **(4)**. But, as we have seen, even if they actually satisfy positive introspection, if they merely fail to commonly know that they do, they may agree to disagree. The implicit assumption that agents commonly know that they satisfy KK thus plays a role in generating the surprising social consequence that agents cannot agree to disagree.

---

positions according to which KK failures are prevalent (for the mundane reasons given above), but it is nevertheless possible to have infinitely iterated knowledge of the kind required for common knowledge.

3.7. False Beliefs

Positive introspection or KK is one restrictive constraint which might be imposed on the knowledge of ideal agents. But the factivity of knowledge also imposes structural constraints on the models we have considered so far. The following two examples show that if agents update on false propositions, they can agree to disagree, even if they commonly believe, of a given prior μ, that that μ is their shared prior, commonly believe that they update by conditionalizing on what they believe, and commonly believe that they satisfy *both* positive *and* negative introspection for belief.

The common prior frame in figure 3.7.1 is a $KD45$ frame: each agent satisfies both positive and negative introspection at every world in the frame. But the agents agree to disagree not only on their *posterior probabilities*: they commonly know (believe) that they each believe the negation of a proposition that the other believes. In this frame, Agent 1 sees only β, regardless of which world is actual, while Agent 2 sees only α regardless of which world is actual. Any prior which gives positive probability to each world will yield a situation in which agents agree to strongly disagree. Agents 1 and 2 commonly believe that Agent 1 believes $\{\beta\}$, and Agent 2 believes $\{\alpha\} = \Omega \setminus \{\beta\}$, or, in other words, the negation of $\{\beta\}$.[58]

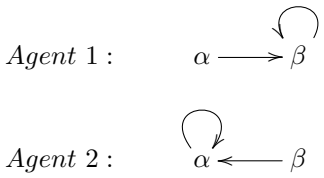


FIGURE 3.7.1. False Beliefs Allow for Common Knowledge Disagreements

In fact, we can go further. Disagreements among agents who commonly know each others' attitudes do not require that any agents have false beliefs at the actual world. Consider the frame in figure 3.7.2, again with two agents 1, and 2, and two states α and β. Once again, regardless of whether the actual world is α or β, Agent 1 sees exactly β, while Agent 2 now sees exactly the elements of $\{\alpha, \beta\}$. The result is that, for any prior that gives nonzero probability to each state, and regardless of which world is actual, it will be

[58]John Collins (1997) argues that van Fraassen's (1984) Reflection Principle can be used to rule out examples of common knowledge disagreement like the one above. The Reflection Principle says, roughly: if one is certain that one will have a certain credence at a later time, one should now have the credence that one knows one will have later. Collins argues that no rational agent could assign positive prior probability to a world in which he knows he would have a false belief. But unless the Reflection Principle is restricted to cases in which one is certain one will update on *veridical* evidence, the Reflection Principle is subject to clear counterexamples. For a recent survey see Briggs 2009: 64-6. Briggs" "qualified reflection" is restricted to cases in which one is certain one's evidence will be veridical (see (ii) on pg. 69). An earlier, similar view can be found in Elga 2007.

a matter of common belief that Agent 1 has posterior probability 1 in β, and posterior probability 0 in α. It will also be common belief that Agent 2 has positive posterior probability in both α and β. The agents thus agree to disagree about their posterior credences in the singleton proposition $\{\alpha\}$. And if β is the actual world, the agents will agree to disagree about $\{\alpha\}$, even though, in β *both* agents have *only* true beliefs. False beliefs are not required for agents to agree to disagree: the view that another *might* have a false belief suffices.

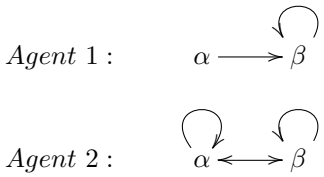


FIGURE 3.7.2. Common Knowledge Disagreement at β with only True Beliefs

This second example displays perhaps the most intuitive case for how rational, common knowledge disagreement is possible. We very often consider it possible that others are be mistaken. Even in this simple, stylized example, the fact that one agent allows for this possibility yields a situation in which agents assign different probabilities to a proposition, even though they commonly believe that their posteriors have the values they in fact have.[59] This disagreement occurs, even though both agents (commonly know that they) satisfy both negative and positive introspection. The model fails **(T)**, but not necessarily because anyone has false beliefs. It fails **(T)** because the agents do not *commonly know* that they each have only true beliefs. Formally, if β is the actual world, the model satisfies each of the "pointed" axioms in Section 3.5.1. We can even think of world α as a figment of Agent 2's imagination. As a matter of fact, in every relevant possible situation Agent 1 would have only true beliefs. It is only Agent 2's misguided view of the situation which allows for the possibility that Agent 1 has a false belief.

The theorist of Bayesian Uniqueness holds that agents whose doxastic attitudes are determined by conditionalizing the evidential prior on propositions stronger or weaker than their total evidence fail to be fully rational. A proponent of Bayesian Uniqueness who endorses a conception of evidence on which a false proposition may be part of one's evidence may accept the examples above as examples of fully rational agents. But if one believes that one's evidence must consist of only true propositions, then conditionalizing

[59]This model could be taken to represent the familiar disagreement between the skeptic and the dogmatist (Pryor 2000). The skeptic believes that the dogmatist may, in irrational exuberance, have plumped for the belief that we are in the "good case", and not in the skeptical scenario. The skeptic and the dogmatist agree to disagree.

on the set of propositions one believes, where one's beliefs may be false, may result in a failure of full rationality. Still, the proponent of Bayesian Uniqueness who endorses this factive conception of evidence may hold that such failures of full rationality are pervasive. On this version of the view, the examples above, which rely on false beliefs, would be examples of agents who fail to be fully rational, but may satisfy a weaker standard, such as reasonableness.For a similar move in a different context, see Lasonen-Aarnio 2010. When faced with disagreements among agents who commonly know each other's probabilistic attitudes, then, this theorist of Bayesian Uniqueness may wish to claim that, even though at least one of the parties to the disagreement is less than fully rational, every agent is reasonable.

The examples also show why "Savageite" *subjective* Bayesians may view Aumann's partitional, $S5$ models as uninteresting, or at least extremely restrictive. For these subjective Bayesians, false beliefs are interpreted as certainty or credence 1 in a false proposition. As a matter of sociology, subjective Bayesians have two conflicting views about the prevalence of certainty. Some subjective Bayesians are apt to claim that we rarely if ever are certain of any propositions. Subjective Bayesians may hold (in addition, or, instead) that we are *never* warranted in being certain of any contingent propositions, or any propositions whatsoever. But, on the other hand, subjective Bayesians are often attracted to the simple view that agents primarily learn or update by conditionalizing a probability function on what they learn. If standard conditionalization (not Jeffrey conditionalization) is a common mode of update, then we are often certain of propositions, namely, whenever we update. In Savage's (1954) benchmark framework, for example, certainty is prevalent, and rational agents may have false beliefs (or: certainties) without forfeiting their rationality. Subjective Bayesians inclined to this Savageite line on the prevalence of certainties may hold that, even when agents *begin* with the same priors (and satisfy the requisite common knowledge assumptions), the prevalence of rational false belief explains how rational agreeing to disagree is possible after all.

The examples also continue the elaboration of a theme we have traced now through three sections. Factivity, **(T)**, is on its face a principle about each individual's knowledge or belief. The assumption does have some social consequences on its own. For example, if all agents have only true beliefs at the actual world (and their beliefs are consistent), then it cannot be that one agent believes the negation of a proposition that another believes. But when we impose **(T)** in interactive frames, we generally assume that all agents satisfy this constraint at all worlds. We thus make the further assumption that

the agents in the frame commonly believe that they each have only true beliefs. But this stronger assumption has a more surprising social consequence. If Aumann's other assumptions are held fixed, it eliminates the possibility of agreeing to disagree. Pointed models, and the pointed axioms introduced in Section 3.5.1 help us to see exactly how this global method of imposing the axioms leads, via unintended consequences about common knowledge, to the impossibility of agreeing to disagree.

3.8. Conclusion

Sections 3.4−3.7 have focused on four assumptions which are implicit in Aumann's original models: (1) that agents commonly know which prior they share; (2) that agents commonly know that they update by conditionalization; (3) that agents commonly know that they satisfy positive introspection; (4) that agents commonly know that they update on true propositions. Natural ways of relaxing these assumptions have been shown to lead to countermodels to the key conditional. As noted in the Introduction (Section 3.1), some have argued that observed violations of the Agreement Theorem, both for belief and knowledge, imply that humans have heterogeneous priors. The four main examples in this paper show that this argument is invalid. Disagreements among agents who commonly know the probability each assigns to a proposition may stem from factors other than heterogeneous priors.

Throughout the paper, we have seen that that special care must be taken when imposing constraints on epistemic models, especially models with many agents. Any constraint which is imposed at all worlds in epistemic models of the kind used here will be a matter of common knowledge among the agents in the models. These unintended assumptions about common knowledge can cause the models to have surprising and even counterintuitive validities. But unlike the surprising deliverances of deep mathematics, these validities simply reveal that the models contain assumptions which we did not intend to make. They are symptoms of errors in our attempt to represent agents' knowledge and beliefs, not profound discoveries about the structure of our social lives. The Agreement Theorems themselves are one of these symptoms.

Appendix

Recall the claim to be proven.

THEOREM. *If a group of $S4$ agents have the same decision function, which satisfies the strong sure-thing principle, and if each agent's action is commonly known among the group, then they all take the same decision.*

PROOF. Say that a decision function d is constant on a set E if and only if $d(P_i(\omega)) = d(P_i(\omega'))$ for all $\omega, \omega' \in E$. We first show that any self-evident set E for an $S4$ correspondence is such that, if a decision function satisfies the strong sure-thing principle, and is constant on E with some arbitrary value a, then $d(E) = a$.

Define the set $\mathcal{P}_E = \{P_i(\omega) | \omega \in E\}$. The proof is by induction on the cardinality of $\mathcal{P}_E$. If $|\mathcal{P}_E| = 1$, the result follows trivially, since, by **(T)**, $P_i(\omega) = E$ for all $\omega \in E$. Now suppose the result holds for $|\mathcal{P}_E| \leqslant n$. We use this hypothesis to show that it holds for $|\mathcal{P}_E| = n + 1$. We consider two cases. **(A)** If, for some $\omega \in E$, $P_i(\omega) = E$, the result follows trivially. **(B)** If there is no $\omega \in E$ with $P_i(\omega) = E$, then by **(4)** and **(T)**, there exist at least two worlds $\alpha, \beta \in E$ such that there is no world $\omega \in E$ with $P_i(\alpha) \subsetneq P_i(\omega)$, and, similarly, no $\omega \in E$ with $P_i(\beta) \subsetneq P_i(\omega)$. We divide further into two subcases. **(B1)** If $P_i(\alpha) \cap P_i(\beta) = \emptyset$, then by the strong sure-thing principle, we know that $d(P_i(\alpha) \cup P_i(\beta)) = d(P_i(\alpha))$. We define a new possibility correspondence P_i^* so that $P_i^*(\omega) = P_i(\alpha) \cup P_i(\beta)$ for all ω with $P_i(\omega) = P_i(\alpha)$ or $P_i(\omega) = P_i(\beta)$, and $P_i^*(\omega) = P_i(\omega)$ otherwise. The set E is self-evident for the new $S4$ correspondence P_i^*, and for all $\omega \in E$, $d(P_i^*(\omega)) = d(P_i(\omega))$. Moreover, $\mathcal{P}_E^* = \{P_i^*(\omega) | \omega \in E\}$ has cardinality n, so, by the induction hypothesis, $d(E) = P_i(\omega)$, for all $\omega \in E$. **(B2)** If $P_i(\alpha) \cap P_i(\beta) \neq \emptyset$, we know, by **(4)**, that every $\omega \in P_i(\alpha) \cap P_i(\beta)$ has $P_i(\omega) \subseteq P_i(\alpha) \cap P_i(\beta)$. (Suppose for contradiction, there was a world in $\omega \in P_i(\alpha) \cap P_i(\beta)$ with $\omega' \in P_i(\omega)$ and $\omega' \notin P_i(\alpha) \cap P_i(\beta)$. Transitivity would be violated, since either α or β would be unable to "see" ω'.) It follows that $P_i(\alpha) \cap P_i(\beta)$ is itself a self-evident set, call it F, with $|\mathcal{P}_F|$ at most equal to $n - 1$. By the induction hypothesis it follows that for all $\omega \in F$, $d(F) = d(P_i(\omega))$. We thus have $d(P_i(\alpha) = d(P_i(\beta)) = d(P_i(\alpha) \cap P_i(\beta))$. It follows by strong union consistency that $d(P_i(\alpha) \cup P_i(\beta)) = a$. We define P_i^* as in B1, and use the same reasoning to conclude that $d(P_i(E)) = a$.

The value of a decision function is commonly known to a group if and only if the decision function is constant on a proposition which is public for the group. Every public proposition is self-evident for each agent, and we have shown that any self-evident set E, with d constant on E at value a has $d(E) = a$. We simply consider a public E so

that $(\forall i)(\exists k_i)(\forall \omega \in E)(d(P_i(\omega)) = k_i)$. But now since $(\forall i)(d(E) = k_i)$, it follows that $(\forall i)(\forall j)(k_i = k_j)$. □

FACT 3.8.1. *Let (Ω, Σ, μ) be a probability space where Ω is finite and μ is regular. For $E, A, B \in \Sigma$, if (1) $\mu(E|A) = \mu(E|B) = a$ and $A \cap B = \emptyset$, or if (2) $\mu(E|A) = \mu(E|B) = \mu(E|A \cap B) = a$, then $\mu(E|A \cup B) = a$.*

PROOF. We first show that the conclusion follows given (1). (This establishes satisfaction of the evidential sure-thing principle.) Suppose that $A \cap B = \emptyset$, and $\mu(E \mid A) = \mu(E \mid B) = a$, for $A, B, E \in \Sigma$. We can rewrite this equation as $\mu(E \cap A)/\mu(A) = \mu(E \cap B)/\mu(B) = a$. Taking each equation individually, we have: $\mu(E \cap A) = a\mu(A)$, and $\mu(E \cap B) = a\mu(B)$. Adding these two equations, and dividing by $\mu(A) + \mu(B)$, we have:

$$\frac{\mu(E \cap A) + \mu(E \cap B)}{\mu(A) + \mu(B)} = a.$$

But by the disjointness of A and B, we know that $\mu(A \cup B) = \mu(A) + \mu(B)$, and, moreover, that $\mu(X \cap (A \cup B)) = \mu(A \cap X) + \mu(B \cap X)$. So the above equation is equivalent to $\mu(E|A \cup B) = a$, as required.

We now show (2). We first establish that, if $\mu(E|A) = \mu(E|C) = a$, where $A, C, E \in \mathscr{P}(\Omega)$, and $C \subsetneq A$, then $\mu(E|A \setminus C) = a$. (Note that if $C = A$, the claim is trivial. Since μ is regular and $C \subsetneq A$, $\mu(A) - \mu(C) > 0$.) By assumption, rearranging somewhat, we have: $\mu(E \cap A) = \mu(E \cap C)\mu(A)/\mu(C)$. Since $C \subsetneq A$, $\mu(A) = \mu(C) + \mu(A \setminus C)$, and similarly for $E \cap A$, $E \cap C$ and $E \cap A \setminus C$. So we have:

$$\mu(E|A \setminus C) = \frac{\mu(E \cap A) - \mu(E \cap C)}{\mu(A) - \mu(C)}$$

By assumption, and rearranging again, we have: $\mu(E \cap A) = \mu(E \cap C)\mu(A)/\mu(C)$. We use this equation to get:

$$\frac{\mu(E \cap A) - \mu(E \cap C)}{\mu(A) - \mu(C)} = \frac{\mu(E \cap C)\mu(A)/\mu(C) - \mu(E \cap C)}{\mu(A) - \mu(C)} \tag{3.8.1}$$

We multiply $\mu(E \cap C)$, by $\mu(C)/\mu(C)$, to make this clearer. Then (3.8.1) becomes:

$$\frac{\mu(E \cap C)\mu(A)/\mu(C) - \mu(E \cap C)\mu(C)/\mu(C)}{\mu(A) - \mu(C)}$$

Rearranging, we have

$$\frac{(\mu(A) - \mu(C))\mu(E \cap C)/\mu(C)}{\mu(A) - \mu(C)}$$

And so $\mu(E|A \setminus C) = \mu(E|C) = a$ as desired.

Now suppose we have that $\mu(E|A) = \mu(E|B) = \mu(E|A \cap B) = a$. Since $A \cap B \subseteq A$ and $A \cap B \subseteq B$, repeated application of the claim we have just shown gives us that $\mu(E|A \setminus B) = \mu(E|B \setminus A) = a$. Now, since $(A \cap B) \cap A \setminus B = \emptyset$, $(A \cap B) \cap B \setminus A = \emptyset$, and $A \setminus B \cap B \setminus A = \emptyset$, the evidential sure-thing principle gives us directly that $\mu(E|A \cup B) = a$. □

CHAPTER 4

State Space Models do *Not* Preclude Unawareness (with Peter Fritz)

Abstract. Dekel, Lipman and Rustichini (1998) claim that "standard" state space models preclude unawareness. We show that their results depend on the implicit assumption that axioms governing unawareness are common knowledge. This assumption is problematic, since it entails that the agents know facts about events of which they are unaware. We use "pointed" state space models to show that without this problematic assumption, the triviality results disappear.

Dekel, Lipman and Rustichini (1998, hereafter, DLR) claim that "standard state space models preclude unawareness". They support this claim with three formal triviality results. In each case, they show that certain axioms on unawareness lead to triviality, in the presence of one or another natural assumption about knowledge.

To model unawareness, DLR use *unawareness models*, structures (Ω, K, U) where K and U are functions on "events", understood as subsets of the "state space", Ω. K and U represent the agent's knowledge and unawareness, respectively: K maps each event E to the event $K(E)$ of the agent knowing E; U similarly takes each E to $U(E)$, the event of the agent being unaware of E.

In such "standard state-space models", if E is an event, then the event that it does not occur is $\Omega \backslash E$, abbreviated as $\neg E$. Similarly, for any two events E and F, the event of their both occurring is $E \cap F$. Thus $\neg(E \cap \neg F)$ is the event that if E occurs F also occurs, abbreviated $E \to F$.

Fixing $M = (\Omega, K, U)$, DLR's axioms on unawareness are then:

Plausibility:	$U(E) \to (\neg K(E) \cap \neg K \neg K(E)) = \Omega$
KU-Introspection:	$\neg KU(E) = \Omega$
AU-Introspection:	$U(E) \to UU(E) = \Omega$

Their triviality results concern the following natural axioms on knowledge:

Necessitation:	$K(\Omega) = \Omega$
Monotonicity:	If $E \subseteq F$ then $K(E) \subseteq K(F)$
Weak Necessitation:	$K(E) \subseteq K(\Omega)$

Their main results are then:

THEOREM 4.0.2. *[DLR] Let $M = (\Omega, K, U)$ be an unawareness model satisfying plausibility, KU-introspection and AU-introspection.*

1(i): *If M satisfies necessitation, then for every event E, $U(E) = \emptyset$.*

1(ii): *If M satisfies monotonicity, then for all events E and F, $U(E) \subseteq \neg K(F)$.*

2: *If M satisfies weak necessitation, then for all events E, F and G, $U(E) = U(F) \subseteq \neg K(G)$.*

DLR lay the blame for these results on state space models. But on inspection, DLR's formalization of the intuitive ideas behind their axioms leads to unintended and undesirable consequences about what the agents in the model know. These consequences suggest that the blame rests with DLR's axioms, and not with state space models. And in fact, as we will see, this suggestion turns out to be correct.

We start with the unintended consequences about agents' knowledge. Suppose an unawareness model satisfies AU-introspection and necessitation. Thus, for any E, $U(E) \rightarrow UU(E) = \Omega$. By necessitation, then, for any E, the agent *knows* $U(E) \rightarrow UU(E)$. In fact, by necessitation, $K(U(E) \rightarrow UU(E))$ is itself identical to Ω, so the agent knows this as well. And the point iterates indefinitely. For any event E, let $K^1(E) = K(E)$ and $K^{n+1}(E) = KK^n(E)$. Then, as usual, we define the event of the agent *commonly knowing* E by $CK(E) = \bigcap_{1 \leq n < \omega} K^n(E)$.[1] In the presence of necessitation, AU-introspection entails that for every E, the agent does not just *know* $U(E) \rightarrow UU(E)$, she also *commonly* knows this event. If we replace necessitation with monotonicity or with weak necessitation, the point is only slightly weaker: if the agent commonly knows anything, then, for any E, she commonly knows $U(E) \rightarrow UU(E)$. AU-introspection is not a special case. Similar points could also be made using plausibility or KU-introspection instead.

These facts about agents' knowledge conflict with the motivations DLR provide for their axioms. The axioms governing unawareness are "based on the idea that unawareness of a possibility corresponds to a complete lack of positive knowledge regarding it" (p. 160). But given this motivation, if an agent is unaware of E, the agent must not be able to (commonly) know events such as $U(E) \rightarrow UU(E)$. This, if anything, would amount to substantial "positive knowledge" about E. But, as we have seen, necessitation and AU-introspection lead to the *requirement* that even agents who are unaware of E know and commonly know the event $U(E) \rightarrow UU(E)$.

[1] Since we are working with models for a single agent, this is a degenerate case of common knowledge. But everything we say below carries over straightforwardly to the multi-agent case.

What went wrong? The intuitive idea behind AU-introspection was that if an agent is unaware of an event, she is unaware of being unaware of it. But AU-introspection says something much stronger than this: it says not just that $U(E) \rightarrow UU(E)$ holds at the actual state, but that this event is equal to the universe Ω.

An analogy may help to illustrate this distinction. Suppose we wish to evaluate a novel regulatory strategy which the government has chosen in secret from the public. If we use a state space in which the government adopts the same regulatory strategy at every state, then investors in the model, unlike the ones in the real world, know which policy the government has chosen. To model the real investors' behavior more accurately, we must represent their limited information by including some states at which the government adopts different regulatory strategies than the one it has actually adopted.

We can formalize the notion of an "actual" state by using "pointed" state space models. A *pointed unawareness model* is a tuple $M = (\Omega, a, K, U)$ such that (Ω, K, U) is an unawareness model and $a \in \Omega$. The state a is interpreted as the actual state of the world; the other states are required to represent the agent's limited information about this state. For a pointed unawareness model, the natural versions of DLR's axioms are then:

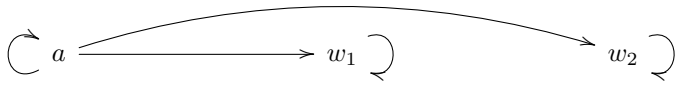
a-Plausibility:	$a \in U(E) \rightarrow (\neg K(E) \cap \neg K \neg K(E))$
a-KU-Introspection:	$a \in \neg KU(E)$
a-AU-Introspection:	$a \in U(E) \rightarrow UU(E)$

These axioms are the ones supported by DLR's informal motivations. And once we use these more accurate formalizations, we can verify the suggestion mentioned earlier: the blame for DLR's impossibility results rests not on state space models, but rather on DLR's unduly strong formal implementation of these intuitive ideas. In fact, we can exhibit non-trivial unawareness in a model which not only satisfies these a-axioms on unawareness, but also simultaneously satisfies all three of the natural assumptions about knowledge introduced earlier:

PROPOSITION 4.0.3. *There is a pointed unawareness model satisfying monotonicity, necessitation, a-plausibility, a-KU-introspection and a-AU-introspection in which there is an event E such that $a \in U(E)$.*

PROOF. The following figure describes a three-state model M witnessing the claim:

$$P(a) = \Omega \qquad P(w_1) = \{w_1\} \qquad P(w_2) = \{w_2\}$$



$$N(a) = N(w_1) \qquad N(w_1) = \{\{w_1\}, \{a, w_1\}\} \qquad N(w_2) = \emptyset$$

Here, we use possibility sets $P(w)$ to give a simple representation of the agent's knowledge: $K(E) = \{w : P(w) \subseteq E\}$. The directed graph depicts P in the sense that $P(w)$ is the out-neighborhood of a given vertex w. Unawareness is slightly subtler. Each world is associated with $N(w) \subseteq 2^\Omega$, and we set $U(E) = \{w : E \in N(w)\}$. It is easily seen that M satisfies the five conditions, and that $a \in \{a, w_1\} = U(\{a, w_1\})$.[2] □

The proposition shows that DLR's Theorem 1(i) cannot be extended to the present setting. In fact, any model $M = (\Omega, a, K, U)$ which satisfies the five conditions in Proposition 1 can also be used to show that DLR's other two results cannot be extended either. For 1(ii), note that by necessitation, $a \in K(\Omega)$, so $a \in U(E)$ despite $a \in K(\Omega)$. For 2, note first that necessitation entails weak necessitation; further, by necessitation and a-plausibility, $a \notin U(\Omega)$, but as before $a \in U(E)$ and $a \in K(\Omega)$.

The model in Proposition 4.0.3 satisfies a number of additional attractive constraints on knowledge, demonstrating that non-trivial unawareness is compatible with a strong logic of knowledge. In particular, it satisfies:

Distribution:	$K(E) \cap K(F) = K(E \cap F)$
Anti-Necessitation:	$K(\emptyset) = \emptyset$
Reflexivity:	$K(E) \to E = \Omega$
Positive Introspection:	$K(E) \subseteq KK(E)$

Moreover, the model in Proposition 4.0.3 is not an isolated case. On the very mild assumption that the agent doesn't know the trivially false event $\emptyset$, we can characterize the conditions under which a given model for the knowledge of an agent can be extended to a pointed unawareness model according to which the agent is unaware of a given event E. Let a pointed unawareness model M *extend* a *pointed knowledge model* $M = (\Omega, a, K)$ if and only if there is a function $U : 2^\Omega \to 2^\Omega$ such that $M = (\Omega, a, K, U)$.

[2]One might consider an extension of a-plausibility along the following lines: Let M be a-ω-plausible just in case $a \in U(E) \to (\neg K)^\omega(E)$ for all events E, where $(\neg K)^\omega(E) = \bigcap_{1 \leq n < \omega} (\neg K)^n(E)$, and $(\neg K)^n(E)$ is defined inductively by the two clauses: $(\neg K)^1(E) = \neg K(E)$ and $(\neg K)^{n+1}(E) = \neg K(\neg K)^n(E)$. The model used in the proof of Proposition 4.0.3 is also a-ω-plausible.

PROPOSITION 4.0.4. *Let $M = (\Omega, a, K)$ be a pointed knowledge model such that $a \notin K(\emptyset)$ and let E be an event. M has an extension satisfying a-plausibility, a-KU-introspection and a-AU-introspection such that $a \in U(E)$ if and only if*

(i) $a \in \neg K(E) \cap \neg K\neg K(E)$, and

(ii) there is an event F such that $a \in F \cap \neg K(F) \cap \neg K\neg K(F)$.[3]

PROOF. Assume first that (i) and (ii). Let $U : 2^\Omega \to 2^\Omega$ be defined so that $U(E) = U(F) = F$, and $U(G) = \emptyset$ for all other events G. It is routine to verify that $M = (\Omega, a, K, U)$ satisfies a-plausibility, a-KU-introspection and a-AU-introspection, and $a \in U(E)$, using the assumption that $a \notin K(\emptyset)$ in the proof of a-KU-introspection.

For the converse, note that (i) follows from a-Plausibility. For (ii), let $F = U(E)$. Then $a \in F$, by a-AU-introspection, $a \in U(F)$, and by a-plausibility, $a \in \neg K(F) \cap \neg K\neg K(F)$. □

DLR's informal motivations suggest that if an agent is unaware of an event E, we should not require that the agent (commonly) know, for example, $U(E) \to UU(E)$. But we might wonder whether, if the agent is *not* unaware of E, the agent can then (commonly) know events of this kind, while still being unaware of other events F. That is, we might wonder whether we can impose the following constraints on a pointed unawareness model (Ω, a, K, U) without running into a triviality result:

a-CK-Plausibility: $a \notin U(E) \Rightarrow a \in CK(U(E) \to (\neg K(E) \cap \neg K\neg K(E)))$

a-CK-KU-Introspection: $a \notin U(E) \Rightarrow a \in CK(\neg KU(E))$

a-CK-AU-Introspection: $a \notin U(E) \Rightarrow a \in CK(U(E) \to UU(E))$

In fact, we can: the example in Proposition 4.0.3 satisfies these three conditions as well as the conditions stated explicitly there. More generally, Proposition 4.0.4 can be extended straightforwardly to these three additional conditions, assuming we impose the plausible constraints of necessitation and anti-necessitation:

PROPOSITION 4.0.5. *Let $M = (\Omega, a, K)$ be a pointed knowledge model satisfying necessitation and anti-necessitation and let E be an event. Then M has a non-trivial extension satisfying a-plausibility, a-KU-introspection, a-AU-introspection, a-CK-plausibility, a-CK-KU-introspection and a-CK-AU-introspection such that $a \in U(E)$ just in case*

(i) $a \in \neg K(E) \cap \neg K\neg K(E)$, and

(ii) there is an event F such that $a \in F \cap \neg K(F) \cap \neg K\neg K(F)$.[4]

[3] This result also holds if we replace a-plausibility by a-ω-plausibility, (i) by $a \in (\neg K)^\omega(E)$, and (ii) by the claim that there is an event F such that $a \in F \cap (\neg K)^\omega(F)$.

[4] Again, we can extend this result to a-ω-plausibility analogously to the extension in the previous footnote.

PROOF. We establish (i) and (ii) as in the proof of Proposition 4.0.4. Assuming (i) and (ii), we define U as in the proof of Proposition 4.0.4, which we have there seen to satisfy a-plausibility, a-KU-introspection and a-AU-introspection, and $a \in U(E)$. For the CK-conditions, consider any event G such that $a \notin U(G)$. Then by construction of U, $U(G) = \emptyset$. Therefore $U(G) \to H = \Omega$ for any event H, and so $a \in CK(U(G) \to H)$ by necessitation, which establishes a-CK-plausibility and a-CK-AU-introspection. For a-CK-KU-introspection, note that by anti-necessitation, $KU(G) = \emptyset$, so $\neg KU(G) = \Omega$, from which $a \in CK(\neg KU(G))$ follows again by necessitation. □

In this note, we have focused on DLR's results. But pointed models provide a completely general strategy for eliminating common knowledge assumptions from epistemic models. We often wish to impose conditions on a model while preserving necessitation, but without assuming that the agents in the model know or commonly know the conditions we are imposing. Pointed models allow us to do this without abandoning the simplicity of state-space models.

CHAPTER 5

Agreement and Equilibrium with Minimal Introspection

Abstract. Standard models in epistemic game theory make strong assumptions about agents' knowledge of their own beliefs. Agents are typically assumed to be *introspectively omniscient*: if an agent believes an event with probability p, she is certain that she believes it with probability p. This paper investigates the extent to which this assumption can be relaxed while preserving some standard epistemic results. Geanakoplos (1989) claims to provide an Agreement Theorem using the "truth" axiom, together with the property of *balancedness*, a significant relaxation of introspective omniscience. I provide an example which shows that Geanakoplos's statement is incorrect. I then introduce a new property, *local balancedness*, which allows us both to correct Geanakoplos's result, and to extend it to cases where the truth axiom may fail. I exploit this general Agreement Theorem to provide novel epistemic conditions for correlated and Nash equilibrium, both of which relax the assumption of introspective omniscience. In all three cases, the results are also extended to infinite state spaces.

JEL Classification Number: C70, C72; D82

Keywords: Agreement Theorem, Nash Equilibrium, Correlated Equilibrium, common prior assumption, epistemic game theory, interactive epistemology, bounded rationality.

5.1. Introduction

5.1.1. Motivation. People often don't know what they want.[1] They don't know what they want for dinner, they don't know who they want to win the game, they don't know where they'd prefer to go on vacation, and so on. The simplest way of modeling this uncertainty is to take it at face value: agents are uncertain about their preferences.

This kind of uncertainty is not just pervasive; in many cases it seems intuitively rational. Learning one's own preferences can require considerable time, effort, and cost. People who try many varieties of wine, for example, expend effort and resources to learn not just about wine, but also about their own preferences for different wines. Similarly,

[1]Thanks to Frank Arntzenius, Amanda Friedenberg, Peter Fritz, Stephen Morris, Sujoy Mukerji and Timothy Williamson for discussion of aspects of this material.

it takes investors a great deal of experience to learn their own levels of risk tolerance. In a range of areas, "you can't know what you like until you try it". But if the cost of determining one's own preferences over a given choice set is greater than the expected value derived from knowing those preferences, it may be rational to remain uncertain of them.

In many cases, it is natural to model agents' uncertainty about their own preferences as uncertainty about their utility function, and this has been the usual approach in economics. But in other cases, agents' uncertainty about their preferences seems better represented as uncertainty about their beliefs. Agents are often unable to draw fine distinctions in their hypothetical choice behavior. For example, most people don't know at what odds they would bet on Hilary Clinton's election to the presidency. While this fact can be modeled by agents' uncertainty of their own utilities, a more natural strategy is to represent these people as uncertain about which probability distribution represents their preferences. As before, being uncertain of this distribution can be intuitively rational. Since it can take a great deal of time and effort to imagine bets with sufficient clarity to predict one's own choice behavior, there may be cases where it is efficient to remain uncertain of one's own probabilistic beliefs.

This uncertainty about one's own preferences—and the resulting uncertainty about one's own beliefs—can be rational in a formal sense as well. Agents who satisfy Savage's axioms on preferences may still be uncertain of their own preferences, and thus of their own beliefs.[2] As a matter of fact, even agents who are rational in Savage's sense may still be uncertain whether they satisfy Savage's axioms at all. In an important paper, Stephen Morris (1996) states additional axioms which can be added to Savage's to ensure that agents *are* certain of their beliefs. But Morris himself concludes that the new axioms on preferences are unpersuasive from a normative standpoint.[3] Even rational agents may fail to know their own beliefs.

Standard epistemic models abstract from the phenomenon of agents' uncertainty about their own beliefs. In Harsanyi type structures, for example, agents are generally assumed to "know" their own type. Since each type is associated with a single probability distribution, by knowing his or her own type, each agent knows his or her own distribution. Let an agent be *introspectively omniscient* just in case, if the agent believes an event

[2]For this formal observation, see Morris 1996 and Epstein and Wang 1996: 1351. Savage himself also made some interesting remarks on the conceptual side of the issue (1967: 308).
[3]Morris 1996: 22.

with probability p, she is certain of believing that event with probability p.[4] In standard Harsanyi type structures, then, agents are introspectively omniscient. Moreover, since every type of every player satisfies this condition, the players in the model have common belief that each of them is introspectively omniscient.

Introspective omniscience implies two more familiar "introspection axioms" on certainty. An agent is *positively introspective* if, if she is certain of an event, she is certain of being certain of that event. She is *negatively introspective* if, if she is not certain of an event, she is certain of not being certain of it. Negative and positive introspection together are still weaker than introspective omniscience: an agent may satisfy both of these conditions but still fail to be certain what probability p (where p is less than 1) she assigns to a given event.

Uncertainty about preferences can lead to failures of both positive and negative introspection (and thus of introspective omniscience). First, consider an example which merely reports an empirical fact. I am uncertain whether my preferences would be represented as assigning the event of there being a banking crisis in 2008 probability 1 or merely very high probability. If in fact I assign this event probability less than 1, then I fail to be negatively introspective; if in fact I assign it probability 1, I fail to be positively introspective.

But less brute, subtler examples can also be given. As mentioned earlier, even agents who satisfy Savage's axioms on preferences may be uncertain whether they are rational in Savage's sense at all. Someone might reasonably assign positive probability to the event that she herself is ambiguity averse, and so be uncertain whether she is rational in Savage's sense. Such a person would be disposed to bet at some odds that there is no single distribution which represents her preferences. If in fact there is a distribution which represents her preferences, she fails to be positively introspective: she may be certain of some event, but since she is uncertain whether she is represented by a single probability distribution at all, she is not certain of being certain of this event.

These observations raise the question: do standard epistemic analyses survive without the usual, idealized assumptions about introspection? Our first example will show that the simple answer is: "not in full generality". This paper then takes up the task of finding weak sufficient conditions which do guarantee that these standard results hold. As a test case, I will restrict attention to one class of prominent models, in which it is assumed

[4]I will use "is certain of E" and "E is a certainty" to mean an agent "assigns E probability 1" and "E is an event the agent assigns probability 1", respectively.

that agents' interim beliefs are consistent with a "common prior". In this common prior setting, I provide a new agreement theorem and new epistemic conditions for correlated equilibrium and Nash equilibrium.

An earlier literature explored the extent to which the Agreement Theorem survives relaxing some of these axioms, and, in particular, negative introspection.[5] But this work was carried out in the presence of the "truth" axiom—that if an agent is certain of an event, that event obtains. The truth axiom is now recognized to be overly strong: rational agents, too, may be wrong about the state of the world. In this paper, I will be exploring not just failures of introspection axioms, but failures of introspection axioms for agents who also may be wrong about the state of the world.[6]

5.1.2. Example. Aumann (1987) showed that if introspectively omniscient agents update on a common prior, commonly know the game, and are commonly known to be rational, then the prior forms a correlated equilibrium distribution for the game in question. We do not yet have a formal definition of the game, rationality or the common prior, but I will present a simple example somewhat informally, to give a feel for the topic of the paper.

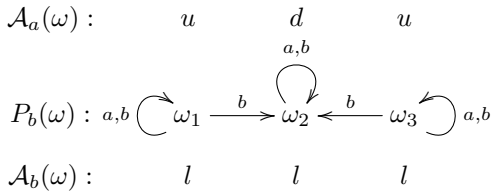
EXAMPLE 5.1.1. In this example, there are three states of the world ($\Omega = \{\omega_1, \omega_2, \omega_3\}$), two players, a and b, and two actions for each player, $A_a = \{u, d\}$; $A_b(\{l, r\}$. At each world, player a considers only that world possible; a always knows the exact state of the world. (Formally, $P_a(\omega_1) = \{\omega_1\}$; $P_a(\omega_2) = \{\omega_2\}$ $P_a(\omega_3) = \{\omega_3\}$.) Player b is not so lucky. If the state of the world is ω_2, then b knows that this is the true state. But if the state of the world is ω_1, player b is uncertain whether the state is ω_1 or ω_2. And if the state is ω_3, player b is uncertain whether the true state is ω_2 or ω_3. (Formally, $P_b(\omega_1) = \{\omega_1, \omega_2\}$; $P_b(\omega_2) = \{\omega_2\}$; $P_b(\omega_3) = \{\omega_2, \omega_3\}$.) The players assign probabilities to events as if they were updating on a common prior μ which assigns $\frac{1}{3}$ to each possible state. So, for example, in ω_1, a assigns $\{\omega_1\}$ probability $1 = \mu(\{\omega_1\}|P_a(\omega_1))$, while b assigns the same event $\frac{1}{2} = \mu(\{\omega_1\}|P_b(\omega_1))$.

[5]Bacharach 1985, Cave 1983, Geanakoplos 1989, Monderer and Samet 1989, Rubinstein and Wolinsky 1990, Samet 1990, Shin 1993, Brandenburger et al. 1992. Two further motivations for studying failures of negative introspection are well known. First, the condition of "plausibility" on unawareness structures (see Dekel et al. 1998) requires a failure of negative introspection. Second, a plausible constraint on the idealization of a rational agent is that the set of sentences of which the agent is certain at least be recursively enumerable. Moreover, it seems plausible that a rational agent could be certain of the theorems of arithmetic. But if such an agent in addition satisfies negative introspection, then it can be shown that the agent is certain of a contradiction: his or her theory of the world is logically inconsistent (McGee 1991 n. 7, Shin and Williamson 1994, Proposition 1).

[6]Agreement Theorems have also been proven with introspective omniscience, but without truth: see, e.g., Bonanno and Nehring 1999 and Hellman 2012. I discuss these results in Section 3.4.

The players' payoffs are the same regardless of the actual state; they are given by the matrix in Figure 5.1.1 where a picks rows and b picks columns.

Players' actions at each state are also represented in Figure 5.1.1 (along with their possibility correspondences). As is easily checked, at each state, each player's actions maximizes subjective expected utility relative to their probabilistic beliefs at that state. (As in this figure, I will use directed graphs throughout the paper to represent simple models. In these figures, directed edges are labelled to indicate that the edge is part of a given player's graph. The *out-neighborhood* of a vertex $\omega \in \Omega$ in a player i's graph is the event i considers possible at ω: in symbols, $P_i(\omega) = N^+_{G_i}(\omega)$.)



	l	r
u	$3,0$	$2,2$
d	$3,3$	$0,0$

The game

FIGURE 5.1.1. Example of Failure of Correlated Equilibrium

The players, as usual, commonly believe the universe $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Thus, since the players' beliefs are consistent with the prior at every world, and they maximize expected utility at every world, they commonly believe that they are rational and that their beliefs are consistent with the prior. By the same reasoning, they also commonly believe that their payoffs are the ones depicted in the matrix.

But the prior induces a distribution over action profiles such that (u, l) is played with probability $\frac{2}{3}$, and (d, l) is played with probability $\frac{1}{3}$. This distribution is not a correlated equilibrium distribution: given a's play, the correlated distribution would require that b play r at least as often as a plays u.

As later definitions will make clear, this result fails even though the players obey the "truth axiom" and also positive introspection.

5.1.3. Interpretation and Outline. As this example illustrates, standard common prior-based results cannot be extended if we allow even slight weakenings of the usual introspection assumptions. The bulk of this paper is dedicated to providing weak conditions on introspection which are sufficient to "resurrect" these standard results. The possibility of providing necessary conditions is discussed in the conclusion.

But these results are best understood negatively. They exhibit the way in which standard common prior-based results rely on introspection assumptions of one form or

another. The fact that these common prior-based analyses depend on introspection assumptions can be seen as a conceptual difficulty with these analyses, and with the common prior assumption itself.

The plan of the paper is then as follows. Section 5.2 lays out the basic model. Geanakoplos (1989) "Theorem 6" provides what would be, if it were correct, the most general known Agreement Theorem for interim beliefs. I present a counterexample to Geanakoplos's statement, which motivates the search for a correct, general Agreement Theorem. Section 5.3 introduces the property of local balancedness, an extension of the property used by Geanakoplos (1989). I then provide a new result which corrects Geanakoplos's, extends it to cases where the truth axiom may fail, and also extends it to infinite state spaces. This is the main result of the paper. The rest of the paper exploits this agreement theorem to provide epistemic conditions for correlated equilibrium and Nash equilibrium. Section 5.4 presents two results for correlated equilibrium: one extends the result of Brandenburger et al. (1992); the other has a new form. Section 5.5 presents epistemic conditions for Nash equilibrium which improve those of Bach and Tsakas (2014) by allowing failures of introspection. In each case, the details of the epistemic conditions are somewhat involved. The main payoff is rather that we *can* use the new agreement theorem to give these epistemic conditions.

Section 5.6 concludes; discussion of related work is given at the close of each section.

5.2. The Basic Model: Earlier Use of Balancedness

5.2.1. The Model. A pure epistemic model will be a tuple $\langle I, (\Omega, \mathcal{T}), (p_i)_{i \in I} \rangle$ including a finite set of players $I = 1, 2, ...n$; and a topological space $(\Omega, \mathcal{T})$, which is separable and completely metrizable (it is "Polish"). The Borel algebra generated by the topology is then used to give the measurable state space $(\Omega, \mathcal{B})$, where elements of $\mathcal{B}$ are *events*. In the sequel, I will often omit reference to $\mathcal{B}$ and $\mathcal{T}$; when the algebra is clear from context I write simply Ω. In Sections 2 and 3, "model" will mean a pure epistemic model; in Sections 4 and 5, a "model" will be a game model.

For a Polish space such as Ω, let $\Delta(\Omega)$ be the set of probability measures on Ω, where $\Delta(\Omega)$ itself is endowed with the weak topology. The *belief maps* $p_i : \Omega \to \Delta(\Omega)$ represent the players' probabilistic beliefs at elements of Ω; these too are required to be measurable (where the algebra on $\Delta(\Omega)$ is taken to be the Borel algebra of the weak topology). A player i is said to be certain of an event $E \in \mathcal{B}$ at a state ω if and only if $p_i(\omega)(E) = 1$.

Now for a measure μ on Ω, let $supp(\mu)$ denote the support of μ, $\{\omega : \omega \in N_\omega \in \mathcal{T} \Rightarrow \mu(N_\omega > 0)\}$, that is, the set of states such that every open neighborhood of every state has positive measure. We use this idea to define *possibility correspondences* from the belief maps, as follows: $P_i(\omega) = supp(p_i(\omega))$. A player i is then said to *believe* an event $E \in \mathcal{B}$ at a state $\omega \in \Omega$ if and only if $P_i(\omega) \subseteq E$. It will be useful also to have *belief functions* $B_i : \mathcal{B} \to \mathcal{B}$, defined in terms of the possibility correspondences: $B_i(E) := \{\omega : P_i(\omega) \subseteq E\}$. A player i's belief function takes each event E to the event of i's believing E. Note that these definitions allow that a player may be certain of an event which he or she does not believe.

Later, we will describe events in terms of elements of the model. For example, for some $p^* \in \Delta(\Omega)$, we may wish to speak of the event that i's probabilistic beliefs are equal to p^*. To do this, I will use square brackets as follows: $[p_i = p^*]$ will denote the event of i's probabilistic beliefs being equal to the distribution p^*, that is, the event $\{\omega : p_i(\omega) = p^*\}$. Similarly, if E is an event and $k \in [0, 1]$, I will write $[p_i(E) = k]$ for the event that i assigns E probability k: $\{\omega : p_i(\omega)(E) = k\}$.

Introspective omniscience is defined as follows:

Introspective Omniscience: $\forall\omega \in \Omega$, $P_i(\omega) \subseteq \{\omega' : p_i(\omega') = p_i(\omega) \wedge P_i(\omega') = P_i(\omega)\}$

As noted in the introduction, introspective omniscience implies the following two conditions:

Positive Introspection: $\forall\omega, \omega'\ \omega' \in P_i(\omega) \to (P_i(\omega') \subseteq P_i(\omega))$ (also called (4))
Negative Introspection: $\forall\omega, \omega'\ \omega' \in P_i(\omega) \to P_i(\omega) \subseteq P_i(\omega')$ (also called (5))

These axioms are more familiar when stated using the notion of belief. Positive introspection says that if an agent believes an event, then she believes that she believes it ($B_i(E) \subseteq B_iB_i(E)$). Negative introspection says that if she does not believe an event, she believes that she does not believe it (using $\neg$ for relative complement within Ω, $\neg B_i(E) \subseteq B_i(\neg B_i(E))$). (Note that introspective omniscience also implies the analogous conditions for certainty, but in our setting it is more natural to consider the ones for belief.)

Aumann's (1976) original models were "partitional". Every agent in these models satisfies (T), (4) and (5), where (T) is:

Truth: $\forall\omega\ \omega \in P_i(\omega)$ (also called (T))

(in fact, the conjunction of (T) and (5) entails (4)).

We wish to describe not just one player's beliefs, but the relationship between players' beliefs. Accordingly, we introduce some key interactive notions.

DEFINITION 5.2.1. Fix a model $\mathcal{M}$. An event E is *self-evident to* i if and only if for all $\omega \in E$, $P_i(\omega) \subseteq E$. An event is *public* to a group $G \subseteq I$ if it is self-evident to all $i \in G$.

A group $G \subseteq I$ *commonly believes* an event F at ω just in case there is some event E with E public to G, $\omega \in E$, and, for all $i \in G$, $\forall \omega \in E, P_i(\omega) \subseteq F$.[7] Just as with belief, we define a function $CB_G(F)$ which takes events to the event of their being commonly believed by G.

This paper studies cases in which agents' interim beliefs are consistent with a common prior. We define a prior as follows:

DEFINITION 5.2.2. Fix a model $\mathcal{M}$ and a group $G \subseteq I$. An event $E \in \mathcal{B}$ is *common prior consistent* for a group $G \subseteq I$ if there exists a measure $\mu \in \Delta(\Omega)$ such that

(i) $\mu(E) = 1$;

for every $i \in G$, for every $\omega \in E$,

(iia)$\mu(P_i(\omega)) > 0$; and

(iib) $p_i(\omega) = \mu(\cdot|P_i(\omega))$.

μ is then a *common prior for* G *over* E. If $G = I$, μ is *a common prior over* E.

Note that this definition is consistent with more recent justifications of the common prior assumption, as a form of consistency among interim beliefs (Bonanno and Nehring 1999; Feinberg 2000). The common prior is not assumed to be unique or to be a "genuine" representation of agents' beliefs at some prior time.

5.2.2. Earlier Definitions of Balancedness; Geanakoplos's Theorem. Geanakoplos (1989) defines his property of "balancedness" for a restricted class of models.

DEFINITION 5.2.3. A model $\mathcal{M}$ is *regular* if:

(i) $|\Omega|$ is finite;

(ii) Ω is common prior consistent for the group I; and,

(iii) for some common prior μ over Ω, for all $\omega \in \Omega$, $\mu(\{\omega\}) > 0$.

[7]Monderer and Samet (1989) prove that in finite spaces, the characterization in terms of *public* events (suggested originally by Aumann 1976, and informally by Lewis 1969: 52-7; cf. also Friedell 1969) is equivalent to the characterization as an infinite intersection: letting $B_G^1(E) = \cap_{i \in G} B_i(E)$ and $B_G^n(E) = \cap_{i \in G} B_i B_G^{n-1}(E) \cap B_B^{n-1}(E)$, $CB_G(E) = \cap_{n \in \mathbb{N}} B_G^n(E)$. Their result holds in our "regular models" (see below, Definition 5.2.3). Kajii and Morris (1997, Theorem 9) prove an analogous result for measure spaces in general, using equivalence classes of measurable sets defined by symmetric difference under a prior μ. Their result applies to a well-behaved subset of the models considered here. But since I am taking the P_i as primitive, and not defined from certainties, I do not need this machinery.

(Note that condition (iii) implies that the algebra $\mathcal{B}$ must contain all singletons.) In regular models, we can replace the p_i with the specification of a prior μ, since together with the P_i, this is enough to determine players' interim probabilistic beliefs.

Balancedness comes in two varieties. In the first instance it is defined with respect to a specific event. The second, general definition is then given with respect to all *self-evident* events. Here is Geanakoplos's definition:

DEFINITION 5.2.4. Fix a regular model $\mathcal{M}$. A possibility correspondence P_i is balanced *with respect to* an event $E \subseteq \Omega$ if and only if there exists some function $\lambda : \{P_i(\omega)|\omega \in \Omega\} \to \mathbb{R}$ such that for every $\omega \in \Omega$, $\sum_{C \in \{P_i(\omega')|\omega' \in \Omega\}} \lambda(C)\chi_C(\omega) = \chi_E(\omega)$.

P_i is then *balanced* if and only if, for every self-evident event E, it is balanced with respect to E.

Note that in the unqualified definition of balancedness, *different* functions λ may be used to "balance" each self-evident event: all that is required is that for each self-evident event, there is *some* such λ.

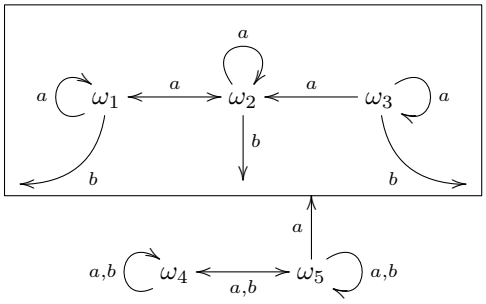
Geanakoplos (1989) claims that agents in regular models agents who are balanced and satisfy (T) cannot agree to disagree ("Theorem 6"):

CLAIM 5.2.5. Let $\mathcal{M}$ be a regular model and $G \subseteq I$ a group where for each $i \in G$, P_i is balanced and satisfies (T). If there is a state $\omega \in \Omega$, an event E and for each $i \in G$, a $k_i \in [0, 1]$ such that it is commonly known that i assigns E probability k_i, then for each i, j, $k_i = k_j$.

The following example shows that this claim is incorrect.

EXAMPLE 5.2.6. Let $\mathcal{M}$ be a model with: $I = \{a, b\}$, $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$.

The model is depicted in Figure 5.2.1. (Note that the directed graph is altered for the sake of simplicity; arrows pointing to the box indicate that an agent considers the worlds in the box possible.) The table describes what players consider possible at each world. Their probabilities at each world ω are consistent with the uniform prior μ, defined so that $\forall\omega \in \Omega$, $\mu(\omega) = \frac{1}{5}$. For this μ, for all $\omega \in \Omega$ and $i \in I$, $p_i(\omega) = \mu(\cdot|P_i(\omega))$. Thus the model is regular.



	$P_a(\omega_n)$	$P_b(\omega_n)$
ω_1	$\{\omega_1, \omega_2\}$	$\{\omega_1, \omega_2, \omega_3\}$
ω_2	$\{\omega_1, \omega_2\}$	$\{\omega_1, \omega_2, \omega_3\}$
ω_3	$\{\omega_2, \omega_3\}$	$\{\omega_1, \omega_2, \omega_3\}$
ω_4	$\{\omega_4, \omega_5\}$	$\{\omega_4, \omega_5\}$
ω_5	Ω	$\{\omega_4, \omega_5\}$

FIGURE 5.2.1. Countermodel to Geanakoplos's Agreement Theorem

It is easily seen that both agents in the example are *balanced.* For b, this is trivial. The only self-evident events are $\{\omega_1, \omega_2, \omega_3\}$, $\{\omega_4, \omega_5\}$ and the universe Ω. The first is balanced by $\lambda(P_b(\omega_1)) = 1$, and for all other admissible E, $\lambda(E) = 0$. The second is balanced by $\lambda(P_b(\omega_4)) = 1$ and for all other admissible E, $\lambda(E) = 0$. For Ω, we let λ be 1 on both $P_b(\omega_1)$ and on $P_b(\omega_4)$.

For a, balancedness is similarly trivial for the self-evident events $\{\omega_1\omega_2\}$, $\{\omega_2, \omega_3\}$ $\{\omega_4, \omega_5\}$ and Ω. The only mildly tricky case is the event $\{\omega_1, \omega_2, \omega_3\}$. Here we set $\lambda(P_a(\omega_5)) = 1$, $\lambda(P_a(\omega_4) = -1$, and for all other admissible E, $\lambda(E) = 0$.

The event $E = \{\omega_1, \omega_2, \omega_3\}$ is public, and thus a matter of common belief. Consider now the event $F = \{\omega_1, \omega_3\}$. Given the prior μ, it is easy to see that for all $\omega \in E$, $p_a(\omega)(F) = \mu(F|P_a(\omega)) = \frac{1}{2}$. But it is equally easy to see that $\forall\omega \in E$, $P_b(\omega)(F) = \mu(F|E) = \frac{2}{3}$. So it is both common belief that $p_a(F) = \frac{1}{2}$ and $p_b(F) = \frac{2}{3}$. The agents "agree to disagree".

The problem in this example arises because a's correspondence can be balanced only by using sets "outside" of the relevant self-evident event. Geanakoplos's "balancing" sum allows this correspondence to count as balanced because it ranges over the whole set $\{P_i(\omega)|\omega \in \Omega\}$. We need some restriction to rule out the counterexample.

Brandenburger et al. (1992) restrict this sum. They also call their property balancedness, but I will call it *subset balancedness*:

DEFINITION 5.2.7. Fix a regular model $\mathcal{M}$. A possibility correspondence is subset balanced *with respect to* an event $E \subseteq \Omega$ if and only if there exists some function $\lambda : \{P_i(\omega)|\omega \in \Omega\} \to \mathbb{R}$ such that $\forall\omega \in \Omega$, $\sum_{C \in \{P_i(\omega')|P_i(\omega') \subseteq E\}} \lambda(C)\chi_C(\omega) = \chi_E(\omega)$.

As usual it is subset balanced if it is subset balanced with respect to all self-evident events.

Note here that the sum ranges only over elements of $\{P_i(\omega)|P_i(\omega) \subseteq E\}$.

If subset balancedness were used instead of balancedness plus truth (T), then we could prove a version of the Agreement Theorem. But this is only because we require that the correspondence be subset balanced on the whole universe. Later, however, we will want to relax the truth axiom, and to use the Agreement Theorem "locally" within a specific self-evident event, without requiring any "global" properties on the whole universe. But under these conditions, subset balancedness is ill-behaved. To see this, consider a modification of Example 5.2.6 so that $P_a(\omega_5) = \{\omega_1, \omega_2, \omega_3\}$. P_a is now subset balanced with respect to the self-evident event $E = \{\omega_1, \omega_2, \omega_3\}$. Since this event is still self-evident, and since for all $\omega \in E$ the possibility correspondences are unchanged, the same argument as above shows that a and b can agree to disagree in their probabilities for F.[8] The property of local balancedness, introduced in the next section, will avoid this kind of problem.

5.3. Local Balancedness: The Agreement Theorem

5.3.1. Local Balancedness. To give the intuition for local balancedness, I will first define it in the finite case, providing two results which relate it to better-known properties from epistemic models.

Subset balancedness uses a particular restriction on the sum in the definition of balancedness. Local balancedness is a variation on this theme. It will be convenient to have the following notation. The set $\mathcal{C}_E$ will denote the events the agent *could* believe, if E obtains: $\mathcal{C}_E = \{C | \exists \omega(\omega \in E \wedge P_i(\omega) = C)\}$. The elements of $\mathcal{C}_E$ are events "local" to E in the sense that, if E occurs, they might represent the information of the agent in question.

Definition 5.3.1. Fix a regular model $\mathcal{M}$. A possibility correspondence P_i is *locally balanced with respect to* an event $E \subseteq \Omega$ if and only if there exists a function $\lambda : \{P_i(\omega) | \omega \in \Omega\} \to \mathbb{R}$ such that for all $\omega \in \Omega$, $\sum_{C \in \mathcal{C}_E} \lambda(C)\chi_C(\omega) = \chi_E(\omega)$.

A possibility correspondence is locally balanced *within* Ω if it is locally balanced with respect to every self-evident event.

This is not our official definition, since it is restricted to regular models. But the definition for regular models allows us to prove some preliminary results which help to give a sense for the property.

[8] To see that this correspondence is not subset balanced *globally*, consider the self-evident event $\{\omega_1, \omega_2, \omega_3, \omega_5\}$.

First, let $\mathcal{M}$ be a regular model, $i \in I$ a player so that P_i satisfies (T), and an E which is self-evident to i. Then it is easy to see that P_i is subset balanced with respect E if and only if it is locally balanced with respect to E. The two properties (for self-evident events) differ only once we have relaxed the truth axiom.

Next, we can show, following Geanakoplos:

PROPOSITION 5.3.2. *(Geanakoplos) Let $\mathcal{M}$ be a regular model and i a player. If P_i satisfies* (4) *and* (T)*, P_i is locally balanced.*

Geanakoplos proves this proposition for his property of balancedness, but the argument is the same for local balancedness.

PROOF. By induction on $|\Omega|$. The base case is trivial. Now we suppose the result holds for all $|\Omega| < n$, and consider $|\Omega| = n$. Find some ω^* such that $|P_i(\omega^*)|$ is minimal among $\{P_i(\omega)|\omega \in \Omega\}$. By (4), every $\omega' \in P_i(\omega^*)$, has $P_i(\omega') \subseteq P_i(\omega^*)$. Since $|P_i(\omega^*)|$ is minimal, it must be that $P_i(\omega') = P_i(\omega^*)$. Now construct a new model, $\mathcal{F}'$, where $\Omega' = \Omega \setminus P_i(\omega^*)$, and $P_i^{'}(\omega) = P_i(\omega) \setminus P_i(\omega^*)$. The new P_i' itself satisfies (4) and (T), and $|\Omega'| < n$, so P_i' is locally balanced, by the induction hypothesis. Now consider an E which is self evident for the original P_i. If $E \cap P_i(\omega^*) = \emptyset$, then E is self-evident to the new P_i' as well, so we use the λ which witnesses the local balancedness of P_i'. If $E = P_i(\omega^*)$, we simply use $\lambda(P_i(\omega^*)) = 1$, and $\lambda(P_i(\omega)) = 0$ for all other sets $P_i(\omega)$. Finally if $P_i(\omega^*) \subsetneq E$, we first consider $F = E \setminus P_i(\omega^*)$, which must be self-evident for P_i'. We use the λ which witnesses the local balancedness of P' on F. This leaves $P_i(\omega^*)$ unbalanced, but we can simply set $\lambda(P_i(\omega^*))$ to whatever value is needed. □

It is an immediate corollary of this proposition that every partitional correspondence is locally balanced (since partitional correspondences obey both (T) and (4)). Thus local balancedness affords more generality than the standard partitions of Aumann 1976. Moreover, it is known that in finite spaces, the Agreement Theorem holds in the presence of positive introspection and truth (Bacharach 1985, Samet 1990). If we can provide an Agreement Theorem using local balancedness, it will be more general than those results.

We will also extend these results to infinite spaces, where local balancedness will prove to be more flexible than standard conditions. Samet 1992 shows that, in infinite spaces, positive introspection (4) plus truth (T) is not enough to ensure that the Agreement Theorem holds. Local balancedness not only generalizes the finite case; it also provides the needed condition in infinite spaces.

Before we define local balancedness in infinite spaces, we extend the notion of a characteristic function, using intersection instead of membership.

DEFINITION 5.3.3. Let $\chi_C : \mathscr{P}(\Omega) \to \{0,1\}$ be the function with $\chi_C(F) = 1$ if $C \cap F \neq \emptyset$, and $\chi_C(F) = 0$ otherwise.

Instead of using individual states in the definition of local balancedness, we now use cells of finite partitions. Recall that a partition $\mathcal{P}$ of a set E is *at least as fine* as a partition $\mathcal{Q}$ of E if for every $P \in \mathcal{P}$, there is some $Q \in \mathcal{Q}$ such that $P \subseteq Q$.

In addition to using finite partitions, we have to require finiteness in a further respect. Note that when we move from regular models to infinite spaces, the sum in the definition of local balancedness is no longer guaranteed to be well defined. For a given i and E, $|\mathcal{C}_E|$ is no longer be guaranteed to be finite, so even in the countable case we would need to sum relative to a sequence. In the official definition of local balancedness, we avoid this difficulty by simply stipulating that the balancing function λ assign non-zero value to at most finitely many elements of $\mathcal{C}_E$.

DEFINITION 5.3.4. Fix a model $\mathcal{M}$. A possibility correspondence P_i is *locally balanced with respect to an event* $E \in \mathcal{B}$ if and only if there is some partition $\mathcal{Q}$ of Ω which

(i) induces a finite partition of E; and

(ii) for all finite partitions $\mathcal{P}$ of Ω which are at least as fine as $\mathcal{Q}$, there exists a function $\lambda : \{P_i(\omega) | \omega \in \Omega\} \to \mathbb{R}$ so that

(a) $|\{C \in \mathcal{C}_E : \lambda(C) \neq 0\}| \in \mathbb{N}$; and

(b) for every $P \in \mathcal{P}$, $\sum_{C \in \mathcal{C}_E} \lambda(C)\chi_C(P) = \chi_E(P)$.

A correspondence is locally balanced within E if it is locally balanced with respect to every self-evident $E' \subseteq E$.

The definition of balancedness *within* E is not the global definition we used earlier for balancedness and subset balancedness. Those definitions use the special case where a correspondence is (locally) balanced within Ω.

Since $\mathcal{P}$ is itself finite, (iia) and (iib) together have the consequence that there is a finite cover of E contained in $\mathcal{C}_E$. In other words, there is an event F such that $|F|$ is finite and $\{P_i(\omega) | \omega \in F\}$ covers E.

We will be particularly interested in the following class of events:

DEFINITION 5.3.5. Fix a model $\mathcal{M}$ and a group $G \subseteq I$. A common prior consistent event E is *consistently balanced for* G if for every $i \in G$, P_i is locally balanced within E.

E is *consistently balanced* if it is consistently balanced for I.

5.3.2. Agreement Theorem. The next lemma is key to the Agreement Theorem. But in a way it is more flexible than the Agreement Theorem, and will also provide the basis for the later epistemic conditions. The lemma gives conditions under which an agent's probabilistic beliefs throughout a self-evident event agree with the distribution obtained by conditionalizing a prior on that event.

LEMMA 5.3.6. *Let $\mathcal{M}$ be a model, i a player, E an event which is consistently balanced for i, and $E^* \subseteq E$ an event which is self-evident for i. If there is an event F and a $k \in [0,1]$ such that $E^* \subseteq [p_i(F) = k]$ then for any common prior μ over E, $k = \mu(F|E^*)$.*

PROOF. By hypothesis, there is a k such that $E^* \subseteq [p_i(F) = k]$. Since μ is a common prior, we know that $\mu(C) > 0$ for all $C \in \{C|\exists\omega(\omega \in E^* \wedge P_i(\omega) = C)\} = \mathcal{C}_{E^*}$, and moreover $\dfrac{\mu(F \cap C)}{\mu(C)} = k$ for all such C.

For a given $C \in \mathcal{C}_E$, we rewrite as $\mu(F \cap C) = k\mu(C)$. Choose a finite partition $\mathcal{Q}$ of E^* which is sufficiently fine that, for all $C \in \mathcal{C}_E$ which are given non-zero value by a λ which balances E^*, $\mathcal{Q}$ induces a partition of $F \cap C$, and $\mathcal{Q}$ itself witnesses local balancedness with respect to E^*. (It's clear that such a partition exists by the definition of local balancedness.) So for a given C we have:

$$\sum_{Q \in \mathcal{Q}} \chi_C(Q)\chi_F(Q)\mu(Q) = \sum_{Q \in \mathcal{Q}} k\chi_C(Q)\mu(Q)$$

Since this is true for each $C \in \mathcal{C}_{E^*}$ to which λ assigns nonzero value, we have

$$\sum_{C \in \mathcal{C}_{E^*}} \lambda(C) \sum_{Q \in \mathcal{Q}} \chi_C(Q)\chi_F(Q)\mu(Q) = \sum_{C \in \mathcal{C}_{E^*}} \lambda(C) \sum_{Q \in \mathcal{Q}} k\chi_C(Q)\mu(Q) \tag{5.3.1}$$

for any function λ. By local balancedness with respect to E^*, and by our choice of $\mathcal{Q}$, for any cell $Q \in \mathcal{Q}$, we have, for an appropriately chosen λ:

$$\chi_{E^*}(Q) = \sum_{C \in \mathcal{C}_{E^*}} \lambda(C)\chi_C(Q).$$

Using this λ, equation 5.3.1 becomes:

$$\sum_{Q \in \mathcal{Q}} \chi_{E^*}(Q)\chi_F(Q)\mu(Q) = \sum_{Q \in \mathcal{Q}} k\chi_{E^*}(Q)\mu(Q)$$

Or, equivalently: $\mu(F|E^*) = k$, as required. □

With this lemma in hand, the Agreement Theorem follows easily:

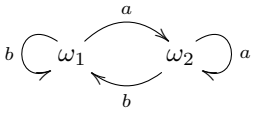
PROPOSITION 5.3.7. *(Agreement Theorem) Let $\mathcal{M}$ be a model, $G \subseteq I$ a group, and E an event which is consistently balanced for G. Suppose there is an $\omega \in E$, and an event F so that for each $i \in G$, there is a $k_i \in [0,1]$ such that $\omega \in CB_G([p_i(F) = k_i])$. Then for all $i, j \in G$, $k_i = k_j$.*

PROOF. For each i, we know there is a public event $E_i \subseteq [p_i(F) = k_i]$. For each such E_i, $\omega \in E_i$, so choosing one for each player, $\bigcap_{i \in G} E_i$ is nonempty and thus also public; we denote this intersection E^*. Now pick an arbitrary $i \in G$. For any $C \in \mathcal{C}_{E^*}$, we write, as above: $\dfrac{\mu(F \cap C)}{\mu(C)} = k_i$. By the hypothesis that each agent is locally balanced with respect to E^*, it follows by Lemma 5.3.6, that, for each i, $k_i = \mu(F|E^*)$. Since μ, E^* and F are the same for all agents in G, $k_i = k_j$ for all $i, j \in G$. □

An extensive earlier literature examined failures of introspective omniscience in connection with the Agreement Theorem and in the presence of the truth axiom (Bacharach 1985, Geanakoplos 1989, Rubinstein and Wolinsky 1990, Samet 1990, Shin 1993). Geanakoplos's theorem is the most general one presented in this initial body of work; Proposition 5.3.7 corrects this result, and extends it both to infinite spaces, and possibility correspondences which do not satisfy (T).

5.3.3. Conditioning Events and "False" Beliefs. More recently, Bonanno and Nehring (1999) proved an Agreement Theorem for finite $KD45$ models (where, in addition to (4) and (5), we have, as here: (D) $\forall \omega \in \Omega$ $P_i(\omega) \neq \emptyset$). But there is an important conceptual difference between their setting and the one used here. In defining consistency with a common prior, Bonanno and Nehring take the agent's "posterior" at ω (thought of as "interim beliefs") to be defined as $\mu(\cdot|\{\omega'|P_i(\omega') = P_i(\omega)\})$. In words, the agent's interim beliefs are as if the agent had conditioned on the event of her having the beliefs she has. This is conceptually difficult to understand; intuitively, agents should conditionalize (or be consistent with regard to) their information, not the event of their having the information they in fact have.

An example may help to illustrate the point.



$\mu(\omega_n) =$ .5 .5

FIGURE 5.3.1. Conditioning Events

EXAMPLE 5.3.8. Let $\mathcal{M}$ be a model with $\Omega = \{\omega_1, \omega_2\}$, and $I = \{a, b\}$. The possibility correspondences are as follows: $\forall \omega \in \Omega$, $P_a(\omega) = \{\omega_2\}$ while $P_b(\omega) = \{\omega_1\}$ (this is depicted in the figure). Finally, $p_a(\omega_1) = 0$, $p_a(\omega_2) = 1$, $p_b(\omega_1) = 1$ and $p_b(\omega_2) = 0$. Intuitively, these probabilistic beliefs are consistent with the uniform prior μ, defined so that $\forall \omega \in \Omega$, $\mu(\omega) = .5$. Player a conditionalizes this prior on what she believes; b conditionalizes the prior on what he believes. But according to Bonanno and Nehring's definition, as is easily checked, these beliefs are not consistent with a common prior.

This is the conceptual difficulty just noted. The P_i are supposed to represent what the agent believes. So the agent should conditionalize on $P_i(\omega)$ itself—what she believes—not on the event that she has the beliefs she in fact has. Even if we do not interpret the relationship between the "prior" and "posterior" as given by the agents' "update", it is unclear how to interpret the resulting property of "consistency".

This definition of conditioning events is not specific to Bonanno and Nehring. In results we will discuss later, Aumann and Brandenburger (1995), Barelli (2009) and Bach and Tsakas (2014), all similarly define the conditioning events as the set of all worlds where agents have a given type. As a formal matter, this means that if any agent has a false belief at a state, the only admissible prior will be one which assigns 0 to this state. But if the prior assigns a state 0 , then *all* agents' interim beliefs must assign that state 0. So although the models nominally relax the truth axiom, the common prior assumption plus this definition of conditioning events effectively restores it. From the standpoint of the probabilities, the agents in the models are *de facto* partitional.

Hellman (2012) also proves an Agreement Theorem for $KD45$ models. His theorem applies to $KD45$ models not included in the hypothesis of Proposition 5.3.7, but Proposition 5.3.7 describes many models in which the P_i are not $KD45$.

5.4. Correlated Equilibrium

Example 5.1.1 showed that standard epistemic conditions for correlated equilibrium may fail if agents are not introspectively omniscient. In this section, I will provide two different sets of epistemic conditions for correlated equilibrium. Both results replace the usual assumption of introspective omniscience with properties related to local balancedness.

5.4.1. Game Models, Conjectures, Rationality, and Games. To study games, we must first extend our pure epistemic models. Let a game model be a tuple

$$\mathcal{M} = \langle I, (\Omega, \mathcal{T}), (p_i)_{i\in I}, A, (\mathcal{A}_i)_{i\in I}, (u_i)_{i\in I}\rangle.$$

The new elements of the model are:

- a finite set of action profiles : $A = A_1 \times A_2, ...A_n$, where each A_i is also finite (and is assumed to be endowed with the discrete algebra);
- for each i, a measurable function $\mathcal{A}_i : \Omega \to A_i$ mapping states to actions; and
- for each i, a state-dependent utility function $u_i : \Omega \times A \to \mathbb{R}$, which for all ω, is integrable with respect to $(\Omega, \mathcal{B}, p_i(\omega))$.

In this section and Section 5, when I write "model", I will mean a game model. Since game models are extensions of epistemic models, the definitions in terms of pure epistemic models can be carried over straightforwardly to this new class of models.

As before we will want to consider certain events, such as the event that a player takes a given action. For a given $a_i \in A_i$, we use $[a_i] = \{\omega | \mathcal{A}_i(\omega) = a_i\}$ to denote the event that i plays action a_i. (This notation is slightly different than the one for conjectures and utilities.) As usual, we will use A_{-i} for $A_1 \times A_2 ... A_{i-1} \times A_{i+1} ... A_n$, and a_{-i} for an element of this set (recall that I use n for $|I|$). The event $[a_{-i}]$ is then defined in the obvious way.

In earlier results, we considered groups of players, subsets of I. When dealing with actions and utilities it becomes clumsy to achieve this kind of generality. So the main results in the following sections will use I as the group in question. This loss of generality is very slight, and it could be remedied easily, by requiring that some players' utilities be insensitive to the actions of others, and defining games relative to a group. But these extensions are both cumbersome and mathematically trivial, so I won't pursue them here.

We will consider one restriction on the relationship between actions and beliefs.

Definition 5.4.1. An event E is *action evident* for i if for all $\omega \in E$, $\exists a_i \in A_i$ such that $P_i(\omega) \subseteq [a_i] \cap E$.

This definition is motivated by the plausible idea that if players are playing a game (E obtains), they know their own actions. Formally, action evidence means more than this: each player's actions are self-evident to him or her within E.

Beliefs over actions will be particularly important. A player i's *conjecture* at a world ω, $\phi_i(\omega) \in \Delta(A_{-i})$, is a distribution over the action profiles of *other* players, defined

as follows: $\phi_i(\omega)(a_{-i}) := p_i(\omega)([a_{-i}])$. i's *conjecture about* j at ω, $\phi_i^j(\omega)$ is defined by $\phi_i^j(\omega)(a_j) = p_i(\omega)([a_j])$.

A player i is *rational* at ω^* if

$$\int_\Omega u_i(\omega, \mathcal{A}_i(\omega^*), \mathcal{A}_{-i}(\omega))dp_i(\omega^*) \geqslant \int_\Omega u_i(\omega, b_i, \mathcal{A}_{-i}(\omega))dp_i(\omega^*)$$

for all $b_i \in A_i$.

I will use Rat_i for the event that i is rational (the set of states ω such that i is rational at ω), and Rat_I for $\cap_{i\in I} Rat_i$.

For fixed $\omega \in \Omega$, $u_i(\omega, \cdot)$ denotes a state *in*dependent utility function. A *game* Γ is then a tuple of pure actions and state independent payoff functions $\Gamma = ((A)_{i\in I}, (u_i(\omega_i^*, \cdot))_{i\in I})$. Note again that the game is not defined for an arbitrary group, but only for I itself.

Fixing a game $\Gamma = ((A)_{i\in I}, (u_i(\omega_i^*, \cdot))_{i\in I})$, the event of a player i *believing in* Γ, denoted $B_i(\Gamma)$, is the event of i believing that he has the utilities he would have in Γ, that is, $B_i(\Gamma) = B_i[u_i = u_i(\omega^*, \cdot)]$. Note that for i to believe in the game, he doesn't have to have any beliefs about *others'* utilities; i believes in the game in the very limited sense that i believes his or her own utilities are as they would be in Γ.

Earlier, we used consistency with a common prior, where the prior must agree with interim beliefs on all events. When considering games we won't make this strong requirement. Instead, we'll use priors *over actions*. To do this, we extend the notion of action-consistency introduced by Barelli (2009).

Definition 5.4.2. Fix a model $\mathcal{M}$. An event $E \in \mathcal{B}$ is action consistent among a group $G \subseteq I$ if and only if there exists a probability distribution $\mu \in \Delta(\Omega)$ such that

(i) $\mu(E) = 1$;

for every $i \in G$, and every $\omega \in E$,

(iia) $\mu(P_i(\omega)) > 0$; and

(iib) for every $a_{-i} \in A_{-i}$ $\phi_i(\omega)([a_{-i}]) = \mu([a_{-i}]|P_i(\omega))$.

Call any such μ an *action prior* over E. If $G = I$, we call μ simply an action prior over E.

Recall that our definition of conditioning events is different from the one used by Barelli and others, as noted above in Section 3.4. Thus the above definition is slightly different than Barelli's.

Every common prior is an action prior, but the converse does not hold in general. Action priors behave like a common prior *over actions* in E, by ensuring that players'

conjectures (not necessarily the full complement of their beliefs) relate to each other as if they were derived from a common prior.

5.4.2. Positive Local Balancedness and Brandenburger et al. (1992). Brandenburger et al. (1992) use a version of subset balancedness to provide epistemic conditions for correlated equilibrium. Here, I will show that we can provide a related result using local balancedness.

First, we modify the definition of local balancedness by restricting the sign of the "balancing" function λ:

DEFINITION 5.4.3. Fix a model $\mathcal{M}$. A possibility correspondence P_i is *positively locally balanced with respect to* an event $E \in \mathcal{B}$ if and only if it is locally balanced with respect to E by a function $\lambda : \{P_i(\omega) | \omega \in \Omega\} \to \mathbb{R}_{>0}$.

It is positively locally balanced within E if it is positively locally balanced with respect to every self-evident $E^* \subseteq E$.

Instead of requiring that the agent be locally balanced with respect to *every* self-evident event, we need only the following weaker property:

DEFINITION 5.4.4. Fix a model $\mathcal{M}$. An agent $i \in I$ is *balanced in action* within E if, for every $a_i \in A_i$, P_i is locally balanced with respect to $[a_i] \cap E$.

i is *positively balanced in action* within E if, for every $a_i \in A_i$, i is positively locally balanced with respect to $[a_i] \cap E$.

If E itself is action-evident then for each a_i, $[a_i] \cap E$ is self-evident. So if a correspondence P_i is locally balanced within an action-evident event E, it is balanced in action within E. But since not every self-evident event need be of this form, a correspondence P_i may be balanced in action within E, but not locally balanced within E.

Now, by analogy to the case of consistently balanced events, we use action consistency and balancedness in action to define:

DEFINITION 5.4.5. Fix a model $\mathcal{M}$ and a group $G \subseteq I$. An action consistent event E is *(positively) consistently action balanced* for G if for every $i \in G$, P_i is (positively) balanced in action within E.

An action consistent event E is (positively) consistently action balanced if for every $i \in I$, P_i is (positively) balanced in action within E.

The general definition (not restricted by "positively") will be used later, in Section 5. The next proposition—our first epistemic condition for correlated equilibrium—uses the notion of a positively consistently action balanced event:

PROPOSITION 5.4.6. *(Brandenburger et al. 1992) Let $\mathcal{M}$ be a model, Γ a game, and E an action evident, positively consistently action balanced event. If $E \subseteq \bigcap_{i \in I}(B_i(\Gamma) \cap Rat_i)$, then the distribution $\hat{\mu} \in \Delta(A)$, given by $\hat{\mu}(a) := \mu([a])$, is an objective correlated equilibrium distribution for Γ.*

Brandenburger et al. (1992) provide the basis of this proposition. The above statement generalizes their result in three ways. (i) It allows for false beliefs (whereas they require truth), (ii) it uses action-consistency in place of a common prior, and, finally (iii) it extends the result to infinite spaces. Note also that we use local balancedness in place of subset balancedness**.**

The proof, however, is a straightforward extension of theirs, so I simply recite it in an appendix for completeness.

5.4.3. Local Action Balancedness: A Second Result. Now we turn to a new kind of epistemic condition for correlated equilibrium, which will use our earlier results about agreement. The basic strategy will be to show that the agents each agree with an action prior conditioned on their own action, in spite of not being introspectively omniscient.

First, we want to ensure that our earlier Lemma 5.3.6 about agreement carries over to action priors:

LEMMA 5.4.7. *Let $\mathcal{M}$ be a model, i a player, and E an event which is consistently balanced for i. Suppose there is an event $E^* \subseteq E$ which is action evident for i so that for some $A^*_{-i} \subseteq A_{-i}$, and some $k \in [0,1]$, $E^* \subseteq [p_i(\{\omega | \mathcal{A}_{-i}(\omega) \in A^*_{-i}\}) = k]$. Then for any action prior μ for i over E, $k = \mu(F|E^*)$.*

PROOF. Observe that Lemma 5.3.6 used only the fact that the prior μ is a prior for P_i *for the relevant event* F, that is, for all $\omega \in E$, $\mu(F|P_i(\omega)) = p_i(\omega)(F)$. The requirement that F have a fixed projection on A_{-i} suffices to ensure that the action prior μ is a prior for F, and the proof goes through as above. □

With this Lemma in hand, we state epistemic conditions for correlated equilibrium:

PROPOSITION 5.4.8. *Let $\mathcal{M}$ be a model, Γ a game, and E an action consistent event. Suppose there is an action-evident $E^* \subseteq E$ such that for every $i \in I$*

(a) $E^* \subseteq B_i(\Gamma) \cap Rat_i$;

(b) i is balanced in action within E^*;

(c) for each $a_i \in A_i$, $\exists \phi_i^* \in \Delta(A_{-i})$ such that $E^* \cap [a_i] \subseteq [\phi_i = \phi_i^*]$;

Then for any action prior μ over E, the distribution $\hat{\mu} \in \Delta(A)$ given by $\hat{\mu}(a) := \mu([a]|E^*)$ is an objective correlated equilibrium distribution for Γ.

This proposition is independent of the one given by Brandenburger et al. (1992). Some correspondences are positively locally balanced within E but fail to be balanced in action within E. Conversely, some correspondences are balanced in action within a given E, and have constant conjectures in their actions, as above, but fail to be positively balanced within E.

PROOF. Since E^* is action-evident, for each i and each a_i, $E \cap [a_i]$ is self-evident. Now for each i, P_i is balanced in action and ϕ_i is constant in each $E^* \cap [a_i]$. So by Lemma 5.4.7, for every $\omega \in E \cap [a_i]$, and $a_{-i} \in A_{-i}$, $\phi_i(\omega)([a_{-i}]) = \hat{\mu}([a_{-i}]|[a_i]))$. Since each player is rational with respect to each of these conjectures in the game Γ, it follows immediately that $\hat{\mu}$ is a correlated equilibrium distribution of Γ. □

5.5. Epistemic Conditions for Nash

Aumann and Brandenburger (1995) provided epistemic conditions for games with more than 2 players *via* the agreement theorem. Example 5.1.1 shows that the Agreement Theorem cannot be extended to generalized possibility correspondences. With slight alteration, the example could be used to show that we also cannot extend Aumann and Brandenburger's result to this more general setting.

The question then arises: can we use the Agreement Theorem of Section 5.3 to provide epistemic conditions for Nash along the lines of Aumann and Brandenburger's original? Is the new Agreement Theorem also robust to more recent extensions of Aumann and Brandenburger's result? In this Section, I answer the question in the affirmative. The details are somewhat involved, and the new result does not give a particularly clear form to Aumann and Brandenburger's core insights. But still it is of some interest that such a result can be given.

5.5.1. Nash for Interim Beliefs without a Prior. We start by re-stating some standard facts about Nash equilibrium—not in the common prior setting. These results

are very clear conceptually, and they also form the basis of the more involved, technical proposition which follows.

To state these results, we need one more definition. A player i's conjectures about j and k are *independent* if, for all $a_j \in A_j$ and $a_k \in A_k$, $\phi_i^j(\omega)([a_j])\phi_i^k(\omega)([a_k]) = \phi_i^{j,k}(\omega)([a_j] \cap [a_k])$. A conjecture $\phi_i(\omega)$ is independent if it is independent for all $j, k \in I \setminus \{i\}$. A player's conjecture being independent is itself an event, which I will denote by $Ind(\phi_i)$.

In general, for two measures μ_1 and μ_2, I will write the product measure as $\mu_1 \otimes \mu_2$. If more than two distributions are involved, I will use the symbol for product; for example, $\phi_i(\omega) = \prod_{j \in I \setminus \{i\}} marg_{A_j}\phi_i(\omega) = marg_{A_1}\phi_i(\omega) \otimes marg_{A_2}\phi_i(\omega) \otimes ...marg_{A_{i-1}}\phi_i(\omega) \otimes marg_{A_{i+1}} \otimes ...marg_{A_b}\phi_i(\omega)$ expresses the statement that i's conjecture at ω is independent.

To set the stage for the common-prior based result, we first develop simple, minimal epistemic conditions for Nash. The following Lemma is now standard:

LEMMA 5.5.1. *Let $\mathcal{M}$ be a model and $\Gamma = ((A)_{i \in I}, (u_i(\omega_i^*, \cdot))_{i \in I})$ a game. Suppose that for some ω, for some $i, j \in I$ with $i \neq j$,*

(a) for some $\phi_i^* \in \Delta(A_{-i})$, $\omega \in B_j[\phi_i = \phi_i^*]$,

(b) $\omega \in B_j B_i \Gamma$ and

(c) $\omega \in B_j(Rat_i)$.

Then for any $a_i \in A_i$ such that $\phi_j(\omega)([a_i]) > 0$, $\sum_{a_{-i} \in A_{-i}} \phi_i^*([a_{-i}])u_i(\omega^*, a_i, a_{-i}) \geq \sum_{a_{-i} \in A_{-i}} \phi_i^*([a_{-i}])u_i(\omega^*, b_i, a_{-i})$, for all $b_i \in A_i$.

Under conditions (a)-(c) any action which has positive support in j's conjecture about i is rational for i with respect to the conjecture j believes i has.

PROOF. By hypothesis, at ω, j believes three events: $[\phi_i = \phi_i^*]$, $B_i\Gamma = \{\omega' \in \Omega | P_i(\omega') \subseteq [u_i(\omega^*, \cdot)]\}$ and Rat_i. It follows that $p_j(\omega)$ assigns 1 to these three events. Since $\phi_j(\omega)([a_i]) > 0$, $p_j(\omega)$ assigns positive probability also to $[\phi_i = \phi_i^*] \cap B_i\Gamma \cap Rat_i \cap [a_i]$. So since this intersection is nonempty, it must be that a_i maximizes $u_i(\omega^*, \cdot)$, with respect to the conjecture ϕ_i^*. □

We then have:

PROPOSITION 5.5.2. *Let $\mathcal{M}$ be a model and $\Gamma = ((A)_{i \in I}, (u_i(\omega_i^*, \cdot))_{i \in I})$ a game. Suppose that for some world $\omega \in \Omega$, for every player $i \in I$, there is a $\sigma_i \in \Delta(A_i)$ so that*

for every $j, k \in I \setminus \{i\}$, $\phi_j^i(\omega) = \phi_k^i(\omega) = \sigma_i$. *Suppose further that, for every* $i \in I$, *there is some player* $j \in I \setminus \{i\}$, *such that*

(a) for every $k \in I \setminus \{i\}$, $\omega \in B_j[\phi_i^k = \phi_i^k(\omega)]$;

(b) $\omega \in B_j B_i(\Gamma)$,

(c) $\omega \in B_j(Rat_i)$,

(d) $\omega \in B_j(Ind(\phi_i))$.

Then $(\sigma_1, \sigma_2 ... \sigma_I)$ *is a Nash Equilibrium of* Γ.

Notice that (a) in this Proposition is different from (a) in the previous Lemma. Here, it is not just that there is *some* conjecture j believes i has; the believed conjecture must (a) agree with i's actual marginal conjectures (i's marginal conjectures at ω) *and* (d) represent others' actions as independent.

Proof. Consider an arbitrary player i. By hypothesis, there is some $j \neq i$ such that j believes that, for every $k \in I \setminus \{i\}$, $i's$ conjecture about k is ϕ_i^{k*}, where in fact this is correct and $\phi_i^{k*} = \phi_i^k(\omega)$. Since every player agrees in their marginal conjectures at ω, then for all $k \in I \setminus \{i\}$, this $\phi_i^k(\omega) = \sigma_k$. Moreover, at ω, j believes that i's full conjecture is $\prod_{k \in I \setminus \{i\}} \phi_i^k(\omega) = \prod_{k \in I \setminus \{i\}} \sigma_k$, j believes that i is rational, and believes that i believes i has utility $u_i(\omega_i^*, \cdot)$. By Lemma 5.5.1, it follows that every a_i assigned positive probability by $\phi_j^i(\omega) = \sigma_i$ is a best response to $\prod_{k \in I \setminus \{i\}} \sigma_k$, when i's payoffs are given by $u_i(\omega_i^*, \cdot)$. Since i was chosen arbitrarily, this holds for every player, and thus $(\sigma_1, \sigma_2 ... \sigma_n)$ is a Nash Equilibrium of Γ. □

Versions of this fact are given in Brandenburger and Dekel (1987: Proposition 4.1), Aumann and Brandenburger (1995) Remark 7.1, Liu (2010 (unpublished)), Dekel and Siniscalchi (2014: 21-2) and Battigalli et al. (2014). As usual, this result is more general than the standard ones, since I have made no assumptions whatsoever about introspection. In fact, this proposition is very general, since it does not even require local balancedness.

5.5.2. Independence. As Proposition 5.5.2 illustrates, two claims must be established to give the epistemic conditions for Nash: agreement of marginal conjectures, and independence of believed conjectures. For agreement, we will simply use the earlier Lemma 5.4.7. Independence will require two new lemmas.

In later results we will be using events E which are both action evident and consistently action balanced (see Definition 5.4.5). Our first lemma will show that if E satisfies these properties, it turns out that each P_i is also locally balanced with respect to E itself.

LEMMA 5.5.3. *Let $\mathcal{M}$ be a model, i a player, and E an event which is action consistent and action-evident for i. If P_i is balanced in action within E, then P_i is locally balanced with respect to E.*

PROOF. $\mathcal{P} = \{\{[a_i] \cap E\} | a_i \in A_j\}$ is a finite partition of E. Moreover, since E is action-evident, each of these events is self-evident to i, so for any distinct $a_i, \hat{a}_i \in A_i$ of $\mathcal{C}_{E\cap[a_i]} \cap \mathcal{C}_{E\cap[\hat{a}_i]} = \emptyset$. Fix an enumeration of the $a_i \in A_i$, and denote the nth action as a_i^n. For each such a_i^n, by hypothesis, there exists $\mathcal{Q}_n$ a finite partition of $E \cap [a_i^n]$, such that, for $\mathcal{Q}_n$, and any finer finite partition of $E \cap [a_i^n]$, there is some λ_n which witnesses the local balancedness of P_i with respect to this $E \cap [a_i^n]$. For each a_i^n pick such a $\mathcal{Q}_n$, and now consider the union of all these partitions $\bigcup_{1 \leqslant n \leqslant k} \mathcal{Q}_n = \mathcal{Q}$. This new $\mathcal{Q}$ partitions E, and is finite. We now construct λ so that, if $\omega \in P_n$, $\lambda(P_i(\omega)) = \lambda_n(P_i(\omega))$. (Given any finer but still finite partition, we can construct λ in the same way.) These λ witness the local balancedness of P_i with respect to E. □

Now we turn to independence directly.

LEMMA 5.5.4. *Let $\mathcal{M}$ be a model, i a player, and E an event which is action evident, and consistently action balanced for i. If for some $\phi_i^* \in \Delta(A_{-i})$, $E \subseteq [\phi_i = \phi_i^*]$, then for any action prior μ for i over E, $marg_A \mu = marg_{A_i} \otimes marg_{A_{-i}}$, and for every $\omega \in E$, $\phi_i(\omega) = marg_{A_{-i}} \mu$.*

PROOF. By constancy of i's conjectures within the action-evident E, and i's local balancedness with respect to $E \cap [a_i]$, for each $a_i \in A_i$, it follows from Lemma 5.4.7 that, for $\omega \in E$ with $\mathcal{A}_i(\omega) = a_i$,

$$\phi_i(\omega)([a_{-i}]) = \mu([a_i, a_{-i}] | [a_i]) \tag{5.5.1}$$

for all $a_{-i} \in A_{-i}$. Rearranging, for every $\omega \in E$ with $\mathcal{A}_i(\omega) = a_i$, we have $\mu([a_i, a_{-i}]) = \phi_i(\omega)([a_{-i}])\mu([a_i])$, for all $a_{-i} \in A_{-i}$. Since this holds for each $a_i \in A_i$, it follows that $marg_A \mu = marg_{A_i} \mu \otimes marg_{A_{-i}} \mu$. But then for each a_i, it's clear that $marg_A(\mu(\cdot|[a_i]) = marg_{A_{-i}} \mu$. Since P_i is action balanced within the action evident E, and $\exists \phi^*$ such that $E \subseteq [\phi_i = \phi_i^*]$, it follows from Lemma 5.4.7 that, for any $\omega \in E$, $\phi_i(\omega) = marg_A \mu(\cdot|[\mathcal{A}_i(\omega)]) = marg_{A_{-i}} \mu$. □

5.5.3. Pairwise Epistemic Conditions. Following Bach and Tsakas (2014), we can give epistemic conditions for Nash equilibrium using a notion of *pairwise* general action consistency, defined relative to a graph on the set of players. Our strategy will be

to build up agreement and independence pair-by-pair along the edges of the graph, to show that the conjectures of the whole group agree, and that each conjecture treats all others as acting independently.

First, we recall some graph-theoretic definitions. An *undirected graph* $G = (N, \mathcal{E})$ consists of a set of vertices and a set of edges described by the binary symmetric relation $\mathcal{E} \subseteq N \times N$ (where $(i,j) \in \mathcal{E}$ describes the same edge as $(j,i) \in \mathcal{E}$). A sequence $(i_k)_{k=1}^m$ of players is a *path* if and only if $(i_k, i_{k+1}) \in \mathcal{E}$ for all $k \in \{1, ..., m-1\}$. A graph is *connected* if for every $i, j \in N$, there is a path $(i_k)_{k=1}^m$ with $i = i_1$ and $j = i_m$. A *cut vertex* of a connected graph is a node $i \in N$ such that, if it is deleted (and any edges containing it are removed), the resulting graph is not connected. A graph is *biconnected* if it has no cut vertices.

Definition 5.5.5. Fix a model $\mathcal{M}$. Let a state $\omega \in \Omega$ be pairwise action consistent if and only if there exists a biconnected graph $H = (I, \mathcal{E})$ where for each $(j,k) \in \mathcal{E}$, there exists an action evident, consistently action balanced event E_{jk} for the group $G_{jk} = \{j,k\}$ such that:

(i) $\omega \in E_{j,k}$; and

(ii) for some $\phi_j^* \in \Delta(A_{-j})$, $\phi_k^* \in \Delta(A_{-k})$, $E_{jk} \subseteq [\phi_j = \phi_j^*] \cap [\phi_k = \phi_k^*]$.

This definition sets the stage for our final proposition:

Proposition 5.5.6. *Let $\mathcal{M}$ be a model, Γ a game, and $\omega \in \Omega$ a pairwise action consistent state.*

Suppose that for every $i \in I$, there exists some $j \in I$ with $j \neq i$, such that:

(a) $\omega \in B_j[\phi_i = \phi_i(\omega)]$

(b) $\omega \in B_j B_i(\Gamma)$

(c) $\omega \in B_j(Rat_i)$.

Then all $j \in I \setminus \{i\}$ share the same conjecture σ_i about i, and the vector of these conjectures about individuals, $(\sigma_1, ..., \sigma_I)$ is a Nash Equilibrium for the game Γ.

Proof. We break the proof into three stages: (1) agreement; (2) independence; (3) Nash.

Agreement. Pick an arbitrary $i \in I$, and an arbitrary $j \in I \setminus \{i\}$. We first show, by induction on path length in H, that $\phi_j^i(\omega) = \phi_k^i(\omega)$ for all $k \in I \setminus \{i,j\}$. Suppose, for the base case, that $(j,k) \in \mathcal{E}$: k is reachable from j in a single step. By Lemma 5.5.3, Definition 5.5.5 $(i) - (ii)$ imply j and k's local balancedness with respect to E_{jk}. We now

use Lemma 5.4.7, and Definition 5.5.5 $(i) - (ii)$, to obtain $\phi_j(\omega)([a_i]) = \phi_k(\omega)([a_i])$, for all $a_i \in A_i$.

Now we suppose this holds for any k_n reachable in n steps on the graph H, and show that it holds for any k_{n+1} reachable in $n+1$ steps on H. Any such k_{n+1} is reachable in one step from some k_n with $(k_n, k_{n+1}) \in \mathcal{E}$. So Definition 5.5.5 $(i) - (ii)$ hold for $G_{k_n,k_{n+1}} = \{k_n, k_{n+1}\}$, where $k_n \in I$ is some player reachable in n steps from this initial j. Now another application of Lemma 5.4.7 gives us that $\phi_{k_n}(\omega)([a_i]) = \phi_{k_{n+1}}(\omega)([a_i])$, for all $a_i \in A_i$. But, by the induction hypothesis, $\phi_{k_n}(\omega)([a_i]) = \phi_j(\omega)([a_i])$, for all $a_i \in A_i$. Since the graph H is biconnected, and I is finite, we can reach every element of $I \setminus \{i\}$ in a finite number of steps, without passing through i,. Moreover, since the initial i was chosen arbitrarily, it follows that for any $i \in I$, all $j, k \in I \setminus \{i\}$ agree in their conjecture about i.

Independence. We now use agreement in marginal conjectures to show that all players' conjectures are independent. First we show that, for every $i, j, k, l \in I$, with $(i, j), (k, l) \in \mathcal{E}$, $marg_A\mu_{ij} = marg_A\mu_{kl}$, by showing that an arbitrary subpath $(a, b), (b, c) \in \mathcal{E}$ preserves equality of $marg_A\mu_{ab} = marg_A\mu_{bc}$. By b's local action balancedness relative to E_{ab} and Lemma 5.5.4, we have $marg_A\mu_{ab} = marg_{A_b}\mu \otimes \phi_b(\omega)$. By Lemma 5.4.7, a's constancy in E_{ab}, and a's local balancedness with respect to E_{ab} (given, once again by using Lemma 5.5.3, and Definition 5.5.5$(i) - (ii)$), we have $marg_{A_b}\mu = \phi_a^b = \sigma_b$, where σ_b is the marginal conjecture shared among all players other than b about b. Thus $marg_A\mu_{ab} = \sigma_b \otimes \phi_b(\omega)$. The same argument shows that $\mu_{bc} = \sigma_b \otimes \phi_b(\omega)$, and thus, that $\mu_{ab} = \mu_{bc}$. Since a, b, c were chosen arbitrarily, any such path preserves equality. By the (bi)connectedness of H, there is such a path from any i to any l. Every step in any such path preserves equality of the relevant action-priors, and thus, for all $i, j, k, l \in I$, with $(i, j), (k, l) \in \mathcal{E}$, $marg_A\mu_{ij} = marg_A\mu_{kl}$, as required. From now on we denote this distribution on actions simply by μ.

Now we show that $\mu = \prod_{i \in I} marg_{A_i}\mu$. Pick an arbitrary $i \in I$. By connectedness of H, there is some j with $(i, j) \in \mathcal{E}$. By the argument just given $marg_A\mu_{ij} = \mu$. Moreover, by i's local action balancedness with respect to E_{ij}, we have $marg_A\mu_{ij} = \mu = marg_{A_i}\mu \otimes marg_{A_{-i}}\mu$. Since this holds for all $i \in I$, we have, by Aumann and Brandenburger (1995) Lemma 4.6, that $\mu = \prod_{i \in I} marg_{A_i}\mu$. (The full argument is by induction on the cardinality of I, and may be found in their paper.)

To complete the proof of independence, it remains to show only that, for every i, $\phi_i = marg_{A_{-i}}\mu$. By the connectedness of H, every i has at least one neighbor, j, and for

every $(i,j) \in \mathcal{E}$, there is some E_{ij}, with μ_{ij} an action prior on E_{ij}, and i locally balanced with respect to E_{ij}. By the constancy of i's conjecture in E_{ij}, we can use Lemma 5.4.7, to derive that $\phi_i(\omega) = marg_{A_{-i}}\mu_{ij}$. But, by the earlier argument $marg_{A_{-i}}\mu_{ij} = marg_{A_{-i}}\mu$, as required.

Nash. With agreement and independence in hand, Nash follows by Lemma 5.5.1. □

Proposition 5.5.6 generalizes Bach and Tsakas (2014) in two ways. First, it allows belief in rationality to float free from the graph structure. In the above statement, belief in rationality may fail to form even a connected graph.[9] Second, local action balancedness is weaker than their assumption that agents are introspectively omniscient. Note once again that the definition of conditioning events used here is different from theirs as well.

Polak (1999) observed that, in Aumann and Brandenburger's Theorem B, common knowledge of conjectures plus common knowledge of payoffs implies common knowledge of rationality. As a result, in perfect information games, where payoffs are assumed to be commonly known, satisfaction of Aumann and Brandenburger's weak conditions still implies common knowledge of rationality. Barelli (2009) and Bach and Tsakas (2014) have shown that this result need not follow, by giving weaker epistemic conditions for Nash Equilibrium. Proposition 5.5.6 and Proposition 5.5.2 share this feature, though the more general and transparent Lemma 5.5.1 has a more striking version of it.

5.6. Conclusion: Interpretation of Local Balancedness

Without introspective omniscience, some common prior-based results fail. Local balancedness is considerably weaker than introspective omniscience, and it allows us to preserve some standard results.

But this leaves us with two questions. First, are these the weakest constraints which can be given which validate these results? Second, and somewhat relatedly, does local balancedness have any interesting interpretation? Given positive introspection, for example, the Agreement Theorem holds. Positive introspection has a simple interpretation, and some might be happy maintaining positive introspection in this setting. But in others, e.g. for correlated equilibrium, positive introspection is not enough. (Recall that the agents in Example 5.1.1 were both positively introspective—they obeyed (4)—but the prior there failed to form a correlated equilibrium.) For local balancedness, we don't seem to have an interpretation which is as clear as the one for positive introspection.

[9]For example, with four players, a, b, c, d: if the other conditions are satisfied, it suffices that a believe that b is rational and b believe that a is rational, while c believe that d is rational, and d believes that c is rational. a and b may each believe c and d not to be rational, and c and d may return the favor.

Ideally we would like to find some way of understanding local balancedness in terms of the constraints it imposes on players' beliefs and knowledge.

The first question can be answered partly in the affirmative. As I have noted, the main result of the paper is the Agreement Theorem. Geanakoplos provided a partial converse to his version of this result. Recall that we can specify regular models just by specifying a prior over the state space in the model. Then if $\mathcal{M} = \langle \Omega, I, (P)_{i \in I}, \mu \rangle$ is a regular model, and $i \in I$ is a player, then a regular model $\mathcal{M}' == \langle \Omega, I', (P')_{i \in I}, \mu' \rangle$ is *i-equivalent* if $i \in I'$, and $P_i = P'_i$.

Then we can state Geanakoplos's converse, using local balancedness, as follows:

PROPOSITION 5.6.1. *(Geanakoplos 1989) Let $\mathcal{M}$ be a regular model, and $i \in I$ a player. If P_i is not locally balanced, there is an i-equivalent $\mathcal{M}'$ and a $j \in I'$ such that P_j is locally balanced within Ω, and i and j agree to disagree.*

This converse applies only to regular models. It also requires quantifying i-equivalent models, that is, over possible priors and possible balanced possibility correspondences. It says nothing about agents on a specific occasion, given the beliefs they in fact have. But unfortunately, stronger converses are not available. If two agents have possibility correspondences which are ill-behaved in exactly the same ways—that is, $\forall \omega \ P_i(\omega) = P_j(\omega)$ and $p_i(\omega) = p_j(\omega)$—then they cannot agree to disagree, because they cannot disagree at all. This converse may be the best we can hope for.[10]

How should we understand local balancedness? Geanakoplos suggested that "Balancedness is a condition under which one can say that every $\omega \in E$ is equally scrutinized by the information correspondence E. Every element $C \in \mathcal{P}$ has an intensity $\lambda(C)$, and the sum of the intensities with which each $\omega \in E$ is considered possible by P is the same, namely 1." (1989: 18) The following model, however, is both balanced and locally balanced, although it does not appear to exhibit "equal scrutiny" in any intuitive sense.

[10]Hellman 2012 gives more exact necessary conditions, but his theorem is of an essentially different kind than the ones studied in this paper. The statement of Hellman's conditions on agents' beliefs are essentially interactive: a given possibility correspondence does not satisfy them (or fail to satisfy them) if a group has not been specified. This makes his result essentially different from the standard theorems, where interactive results are derived from at least *prima facie* single-agent constraints.

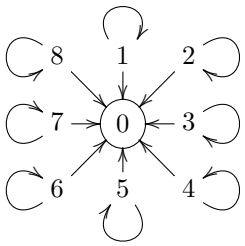


FIGURE 5.6.1. Local Balancedness and Truth do Not Imply Equal Scrutiny (0 can see itself)

World 0 is seen from 9 worlds, whereas all other 8 worlds in the model are seen only from one, namely, themselves. The example could be extended to yield any finite "degree of scrutiny" for one world, while every other world is seen by only itself.

So we are still in need of an intuitive characterization. In the case of positive introspection, we give an intuitive characterization of a model-theoretic property by considering the propositional modal logic which characterizes the class of models which satisfy this property (or, in the logical terminology the class of "frames"). One might hope that, as with positive introspection, we can similarly use a propositional modal logic to give a transparent characterization of which models are locally balanced. Formally, this would mean finding a logic of belief which is valid exactly on the class of models $\mathcal{M}$ in which agents are (positively) locally balanced within the universe Ω of $\mathcal{M}$. If a class of models can be characterized in this way, we say that the class of models is *definable*. Unfortunately, the following proposition shows that such an equivalence cannot be given; it thus suggests that properties related to local balancedness may not admit an intuitive characterization in terms of belief.

PROPOSITION 5.6.2. *The class of (positively) locally balanced models is undefinable by a normal modal logic.*

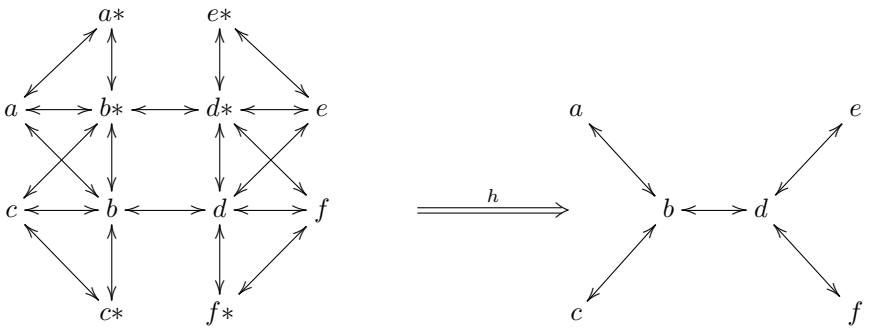


FIGURE 5.6.2. A Locally Balanced model which has a Bounded Morphic Image that is not Locally Balanced

Recall that for conditions such as $(T)+(4)$ on possibility correspondences, the class of models is definable by a normal modal logic, in this case, the logic $S4$.

PROOF. By the Goldblatt-Thomason theorem, a class of models is definable by a normal modal logic only if it is closed under bounded morphic images.[11] In figure 5.6.2 the left-hand model is positively locally balanced (and therefore locally balanced). To see this, note that the only self-evident event is the universe, and that $\{P_i(a^*), P_i(e^*), P_i(c^*), P_i(f^*)\}$ partitions the universe. The right-hand model is not locally balanced. To see this, consider the uniform prior μ defined so that: $\forall\omega \in \Omega$, $\mu(\omega) = \frac{1}{6}$. While $\mu(\{a,c,e,f\}) = \frac{2}{3}$, $\mu(\{a,c,e,f\}|P_i(\omega)) = \frac{1}{2}$ for all $\omega \in \Omega$. Since $\mu(\{a,c,e,f\}|P_i(\omega))$ is constant within Ω, but does not agree with $\mu(\{a,c,e,f\}|\Omega)$, it follows by contraposition of Lemma 5.3.6 that the correspondence is not locally balanced. But the model on the right-hand side of Figure 5.6.2 is a bounded morphic image of the one on the left. Consider the function h which takes every unasterisked world to itself, but takes the asterisked worlds to their unasterisked counterparts. Call the accessibility relation (the relation of directed adjacency) of the left-hand model 'R', and the accessibility relation of the right hand model 'R''. It is straightforward to check that (i) wRv implies $h(w)R'h(v)$; (ii) $h(w)R'v'$ implies that there is some v such that wRv and $h(v) = v'$; (iii) h is surjective. □

Appendix

PROPOSITION. *(Brandenburger et al. 1992) Let $\mathcal{M}$ be a model, Γ a game, and E an action evident, positively consistently action balanced event. If $E \subseteq \bigcap_{i\in I}(B_i(\Gamma) \cap Rat_i)$, then the distribution $\hat{\mu} \in \Delta(A)$ given by $\hat{\mu}(a) := \mu([a])$ is an objective correlated equilibrium distribution for Γ.*

PROOF. (Brandenburger et al. 1992) Following Brandenburger et al, fixing the game $\Gamma = ((A)_{i\in I}, (u_i(\omega_i^*, \cdot))_{i\in I})$ and given a distribution $\nu \in \Delta(A)$, with $\nu(\{a_i\} \times A_{-i})) > 0$, let

$$Q_\nu(a_i) = \{q \in \Delta(A) : supp(q) \subseteq supp(\nu(\cdot|a_i)),$$

$$\sum_{a_{-i}\in A_{-i}} q(a_{-i})u_i(\omega_i^*, a_i, a_{-i}) \geqslant \sum_{a_{-i}\in A_{-i}} q(a_{-i})u_i(\omega_i^*, b_i, a_{-i}) \text{ for all } b_i \in A_i.\}$$

(We use *supp*, as usual, to denote the smallest closed set with measure 1, and use the discrete topology on the finite set of action profiles.) Observe that $Q_\nu(a_i)$ is a compact,

[11] For a standard reference, and proof of the Goldblatt-Thomason theorem, see Blackburn et al. 2001: Sections 3.8 and 5.4.

convex subset of $\Delta(A_{-i})$. We now show that for any action-prior μ on E, $\hat{\mu}$, defined as above, is a correlated equilibrium distribution. This claim is equivalent to the claim that, for each i, for every $a_i \in A_i$ such that $\mu([\{a_i\} \times A_{-i}]) > 0$, $marg_{A_{-i}}\hat{\mu}(\cdot|[a_i]) \in Q_{\hat{\mu}}(a_i)$.

For every $i \in I$, $a_i \in A_i$ and $\omega \in E \cap [a_i]$, it is straightforward to show that $\phi_i(\omega) \in Q_{\hat{\mu}}(a_i)$. Since μ is an action prior, for every $i \in I$, and $\omega \in E$, $supp(\phi_i(\omega)) \subseteq supp(marg_{A_{-i}}(\hat{\mu}))$. Moreover, by the fact that $E \subseteq \bigcap_{i\in I}(B_i(\Gamma) \cap Rat_i)$, we have that for every $i \in I$, $a_i \in A_i$ and $\omega \in E \cap [a_i]$, a_i is optimal with respect to $\phi_i(\omega)$, and $u_i(\omega_i^*, \cdot)$. Thus for every $i \in I$, $a_i \in A_i$ and $\omega \in E \cap [a_i]$, $\phi_i(\omega) \in Q_{\hat{\mu}}(a_i)$.

By the convexity of $Q_{marg_A\mu}(a_i)$, it now suffices to show that, for every $i \in I$, for every $a_i \in A_i$, $marg_{A_{-i}}\hat{\mu}(\cdot|[a_i]) \in conv\{\phi_i(\omega)|\omega \in E \cap [a_i]\}$, where $conv$ denotes the convex hull of the set of all such conjectures. By action consistency, we have $\phi_i(\omega)([a_{-i}]) = \mu([a_{-i}]|P_i(\omega))$, for every $a_{-i} \in A_{-i}$. Recall that given an event E and a player i, we use $\mathcal{C}_E$ to denote the set $\mathcal{C}_E = \{C|\exists\omega(\omega \in E \wedge P_i(\omega) = C)\}$. Consider an arbitrary $i \in I$, $a_i \in A_i$, and $a_{-i} \in A_{-i}$, so that $\mu([a_{-i}]|E \cap [a_i]) > 0$. Now choose a finite partition $\mathcal{Q}$, but fine enough to partition $E \cap [a_i, a_{-i}]$, and fine enough to witness the positive local balancedness of i with respect to $E \cap [a_i]$. Since i is positively locally balanced in action within E,

$$\mu([a_{-i}]|[a_i]) = \frac{1}{\mu([a_i])} \sum_{Q\in\mathcal{Q}} \chi_{[a_i]\cap E}(Q)\chi_{[a_{-i}]}(Q)\mu(Q)$$

$$= \frac{1}{\mu([a_i])} \sum_{Q\in\mathcal{Q}} \sum_{C\in\mathcal{C}_{E\cap[a_i]}} \lambda(C)\chi_C(Q)\chi_{[a_{-i}]}(Q)\mu(Q)$$

for an appropriately chosen λ. It follows that:

$$\mu([a_{-i}]|[a_i]) = \frac{1}{\mu([a_i])} \sum_{C\in\mathcal{C}_{E\cap[a_i]}} \lambda(C)\mu(C)\frac{1}{\mu(C)} \sum_{Q\in\mathcal{Q}} \chi_C(Q)\chi_{[a_{-i}]}(Q)\mu(Q)$$

$$= \frac{1}{\mu([a_i])} \sum_{C\in\mathcal{C}_{E\cap[a_i]}} \lambda(C)\mu(C)\mu([a_{-i}]|C). \tag{5..1}$$

But local positive balancedness gives us that $\frac{1}{\mu([a_i])} \sum_{C\in\mathcal{C}_{E\cap[a_i]}} \lambda(C)\mu(C) = 1$, and, moreover, that every value in the sum is at least 0. So 5..1 gives us

$$\mu([a_{-i}]|[a_i]) \in conv\{\phi_i(\omega)([a_{-i}])|\omega \in E \cap [a_i]\} \tag{5..2}$$

This holds for all $a_{-i} \in A_{-i}$, so $marg_{A_{-i}}\mu(\cdot|[a_i])$ lies in the convex hull of the $\phi_i(\omega)$ such that $\omega \in E \cap [a_i]$, as required. Since $Q_{marg_A\mu}(a_i)$ is convex, it follows by 5..2 that $marg_{A_{-i}}\mu$ lies in $Q_{marg_A\mu}(a_i)$ as well, for every a_i such that $marg_A\mu(\{a_i\} \times A_{-i})) > 0$. So $marg_A\mu = \hat{\mu}$ is a correlated equilibrium for Γ. □

Bibliography

Aarnio, Maria Lasonen. 2010. Unreasonable knowledge. *Philosophical Perspectives*, **24**(1), 1–21. 123

Aaronson, Scott. 2005. The Complexity of Agreement. *Symposium on the Theory of Computing*, Extended Abstract. Full Version at www.scottaaronson.com/papers/agree–econ.pdf. 96

Akkoyunlu, Eralp A., Ekanadham, Kattamuri, & Huber, R. V. 1975. Some Constraints and Tradeoffs in the Design of Network Communications. *SIGOPS Oper. Syst. Rev.*, **9**(5), 67–74. 13

Almotahari, Mahrad, & Glick, Ephraim. 2010. Context, content, and epistemic transparency. *Mind*, **119**(476), 1067–1086. 65

Arad, Ayala, & Rubinstein, Ariel. 2012. The 11-20 Money Request Game: A Level-k Reasoning Study. *American Economic Review*, **102**(7), 3561–73. 17

Aumann, Robert J. 1976. Agreeing to Disagree. *The Annals of Statistics*, **4**(6), 1236–1239. 18, 95, 102, 140, 141, 145

Aumann, Robert J. 1985. On the non-transferable utility value: A comment on the Roth-Shafer examples. *Econometrica: Journal of the Econometric Society*, 667–677. 27

Aumann, Robert J. 1987. Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, **55**(1), 1–18. 137

Aumann, Robert J. 1998. Common Priors: A Reply to Gul. *Econometrica*, **66**(4), 929–938. 115

Aumann, Robert J., & Brandenburger, Adam. 1995. Epistemic Conditions for Nash Equilibrium. *Econometrica*, **63**(5), 1161–1180. 25, 149, 154, 156, 159

Aumann, Robert J., & Hart, Sergiu. 2006 (unpublished). *Agreeing on decisions.* The Hebrew University of Jerusalem, Center for the Study of Rationality. 109

Aumann, Robert J., Hart, Sergiu, & Perry, Motty. 2005 (unpublished). *Conditioning and the sure-thing principle.* Center for the Study of Rationality. 108, 109

Bach, Christian W., & Tsakas, Elias. 2014. Pairwise epistemic conditions for Nash equilibrium. *Games and Economic Behavior*, **85**(0), 48 – 59. 25, 139, 149, 157, 160

Bacharach, Michael. 1985. Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, **37**(1), 167 – 190. 103, 105, 108, 137, 145, 148

Barelli, Paolo. 2009. Consistency of Beliefs and Epistemic Conditions for Nash and Correlated Equilibria. *Games and Economic Behavior*, **67**, 363–375. 25, 113, 149, 151, 160

Barwise, Jon. 1988. Three views of common knowledge. *Pages 365–379 of: Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*. Morgan Kaufmann Publishers Inc. 42

Battigalli, Pierpaolo, Brandenburger, Adam, Friedenberg, Amanda, & Siniscalchi, Marciano. 2014. *Strategic Uncertainty: The Epistemic Approach to Game Theory*. 25, 156

Bernheim, B. Douglas. 1984. Rationalizable Strategic Behavior. *Econometrica*, **52**(4), 1007–1028. 27, 35

Binmore, Ken, & Samuelson, Larry. 2001. Coordinated action in the electronic mail game. *Games and Economic Behavior*, **35**(1), 6–30. 25

Blackburn, Patrick, de Rijke, Maarten, & Venema, Yde. 2001. *Modal Logic*. Cambridge University Press. 163

Bonanno, Giacomo, & Nehring, Klaus. 1997 (September). *Agreeing to Disagree: A Survey*. http://www.econ.ucdavis.edu/faculty/bonanno/PDF/agree.pdf. 113

Bonanno, Giacomo, & Nehring, Klaus. 1999. How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, **28**(3), 409–434. 113, 137, 141, 148

Bonanno, Giacomo, & Nehring, Klaus. 2000. Common Belief with the Logic of Individual Belief. *Mathematical Logic Quarterly*, **46**(1), 49–52. 93

Brandenburger, Adam, & Dekel, Eddie. 1987. Rationalizability and correlated equilibria. *Econometrica*, **55**(6), 1391–1402. 27, 35, 156

Brandenburger, Adam, & Keisler, Jerome H. 2006. An Impossibility Theorem on Beliefs in Games. *Studia Logica*, **84**, 211–240. 83

Brandenburger, Adam, Dekel, Eddie, & Geanakoplos, John. 1992. Correlated Equilibrium with Generalized Information Structures. *Games and Economic Behavior*, **4**, 182–201. 137, 139, 143, 152, 153, 154, 163

Briggs, Rachael. 2009. Distorted Reflection. *The Philosophical Review*, **118**(1), 59–85. 121

Camerer, Colin. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press. 22

Cave, Jonathan A.K. 1983. Learning to agree. *Economics Letters*, **12**(2), 147 – 152. 103, 105, 137

Chwe, Michael Suk-Young. 2001. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press. 36

Clark, Herbert H. 1996. *Using language*. Cambridge University Press. 44

Clark, HH, & Marshall, CR. 1981. Definite reference and mutual knowledge. 10–64. 44

Cohen, Stewart, & Comesaña, Juan. 2013. Williamson on Gettier Cases and Epistemic Logic. *Inquiry*, **56**(1), 15–29. 78

Collins, John. 1997. *How We Can Agree to Disagree*. http://collins.philo.columbia.edu/disagree.pdf. 121

Conee, Earl, & Feldman, Richard. 2004. *Evidentialism: Essays in Epistemology*. Oxford University Press. 103

Cowen, Tyler, & Hanson, Robin. 2014 (forthcoming). Are disagreements honest? *Journal of Economic Methodology*. 96

Dekel, Eddie, & Siniscalchi, Marciano. 2014 (14 January). *Epistemic Game Theory*. 25, 156

Dekel, Eddie, Lipman, Bart L, & Rustichini, Aldo. 1998. Standard state-space models preclude unawareness. *Econometrica*, **66**(1), 159–73. 7, 128, 137

Easwaran, Kenny. 2014. Regularity and Infinitesimals. *Philosophical Review*, **123**(1), 1–41. 97

Elga, Adam. 2007. Reflection and Disagreement. *Noûs*, **41**(3), 478–502. 121

Epstein, Larry G, & Wang, Tan. 1996. " Beliefs about Beliefs" without Probabilities. *Econometrica: Journal of the Econometric Society*, 1343–1373. 135

Fagin, Ronald, Halpern, Joseph Y., Moses, Yoram, & Vardi, Moshe Y. 1995. *Reasoning about Knowledge*. MIT Press. 13, 18, 100

Feinberg, Yossi. 2000. Characterizing Common Priors in the Form of Posteriors. *Journal of Economic Theory*, **91**(2), 127 – 179. 113, 141

Feldman, Richard. 2007. Reasonable Religious Disagreements. *Chap. 17, pages 194–214 of:* Antony, Louise (ed), *Philosophers without Gods: Meditations on Atheism and the Secular*. Oxford University Press. 97

Friedell, Morris F. 1969. On the structure of shared awareness. *Behavioral Science*, **14**(1), 28–39. 18, 102, 141

Geanakoplos, John. 1989 (May). *Game Theory Without Partitions, and Applications to Speculation and Consensus.* Cowles Foundation Discussion Papers 914. Cowles Foundation for Research in Economics, Yale University. 7, 134, 137, 139, 141, 142, 148, 161

Geanakoplos, John. 1994. Common Knowledge. *Chap. 40, pages 1437–1496 of:* Aumann, Robert J., & Hart, Sergiu (eds), *Handbook of Game Theory with Economic Applications*, vol. 2. North Holland. 113

Gilbert, Margaret. 1989. *On social facts.* Princeton University Press. 41

Goodman, Jeremy. 2013. Inexact knowledge without improbable knowing. *Inquiry*, **56**(1), 30–53. 30

Gray, Jim. 1978. Notes on Data Base Operating Systems. *Pages 393–481 of: Operating Systems, An Advanced Course.* London, UK: Springer-Verlag. 13

Greco, Daniel. 2014a. Could KK be OK? *Journal of Philosophy*, **111**(4), 169–197. 10, 119

Greco, Daniel. 2014b. Iteration and Fragmentation. *Philosophy and Phenomenological Research*, n/a–n/a. 10, 78

Grimm, V., & Mengel, F. 2012. An experiment on learning in a multiple games environment. *Journal of Economic Theory*, **147**(6), 2220–2259. 28

Groenendijk, Jeroen, & Stokhof, Martin. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, **14**(1), 39–100. 44

Hájek, Alan. 2010. *Staying regular.* Unpublished Manuscript. 97

Halpern, Joseph Y, & Pucella, Riccardo. 2011. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial intelligence*, **175**(1), 220–235. 102

Hanson, Robin. 2006. Uncommon Priors Require Origin Disputes. *Theory and Decision*, **61**(4), 318–328. 113

Hawthorne, John, & Magidor, Ofra. 2009. Assertion, Context, and Epistemic Accessibility. *Mind*, **118**(470), 377–397. 65, 86, 87

Hawthorne, John, & Magidor, Ofra. 2011. Assertion and Epistemic Opacity. *Mind*, **119**(476), 1087–1105. 65

Heal, Jane. 1978. Common knowledge. *The Philosophical Quarterly*, **28**(111), 116–131. 15, 41

Heifetz, Aviad. 1996. Comment on Consensus without Common Knowledge. *Journal of Economic Theory*, **70**(1), 273–77. 113

Heim, Irene. 1983. On the projection problem for presuppositions. *Pages 114–125 of: West Coast Conference in Linguistics*, vol. 2. 44

Heinemann, Frank, Nagel, Rosemarie, & Ockenfels, Peter. 2004. The Theory of Global Games on Test: Experimental Analysis of Coordination Games with Public and Private Information. *Econometrica*, **72**(5), 1583–1599. 22

Hellman, Ziv. 2012 (June). *Deludedly Agreeing to Agree*. Abstract in TARK 2013; full paper unpublished: http://www.coll.mpg.de/sites/www.coll.mpg.de/files/workshop/DeludedPartitions21.pdf. 137, 149, 161

Hintikka, Jaako. 1962. *Knowledge and Belief*. Cornell University Press. 100, 106

Kajii, Atsushi, & Morris, Stephen. 1997. Common p-Belief: The General Case. *Games and Economic Behavior*, **18**, 73–82. 141

Kamp, Hans, Van Genabith, Josef, & Reyle, Uwe. 2011. Discourse representation theory. *Pages 125–394 of: Handbook of Philosophical Logic*. Springer. 44

Karttunen, Lauri. 1974. Presupposition and linguistic context. *Theoretical linguistics*, **1**(1), 181–194. 43

Kelly, Thomas. 2013. How to be an Epistemic Permissivist. *In:* Greco, John, Steup, Matthias, & Turri, John (eds), *Contemporary Debates in Epistemology*. Blackwell Publishing Ltd. 97

Kinderman, Peter, Dunbar, Robin, & Bentall, Richard P. 1998. Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, **89**(2), 191–204. 45

Kolodny, Niko, & MacFarlane, John. 2010. Ifs and oughts. *The Journal of Philosophy*, **107**(3), 115–143. 108

Kripke, Saul A. 1979. A Puzzle about Belief. *Pages 239–283 of:* Margalit, Avishai (ed), *Meaning and Use*. Dordrecht. 42

Lewis, David. 1975. Languages and language. *Minnesota Studies in the Philosophy of Science*. 66

Lewis, David K. 1969. *Convention: A philosophical study*. Harvard University Press. 18, 102, 141

Lin, Shuhong, Keysar, Boaz, & Epley, Nicholas. 2010. Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, **46**(3), 551–556. 45

Liu, Qinming. 2010 (unpublished). *Higher-Order Beliefs and Epistemic Conditions for Nash Equilibrium.* 25, 156

McGee, Vann. 1991. We turing machines aren't expected-utility maximizers (even ideally). *Philosophical Studies*, **64**(1), 115–123. 137

Mengel, Friederike. 2012. Learning across games. *Games and Economic Behavior*, **74**(2), 601 – 619. 28

Monderer, Dov, & Samet, Dov. 1989. Approximating Common Knowledge with Common Beliefs. *Games and Economic Behavior*, **1**(2), 170–190. 18, 102, 137, 141

Morris, Stephen. 1995. The Common Prior Assumption in Economic Theory. *Economics and Philosophy*, **11**(2), 227–253. 96

Morris, Stephen. 1996. The logic of belief and belief change: A decision theoretic approach. *Journal of Economic Theory*, **69**(1), 1–23. 135

Moses, Yoram, & Nachum, Gal. 1990. Agreeing to disagree after all. *Pages 151–168 of: Proceedings of the 3rd conference on Theoretical Aspects of Reasoning and Knowledge.* 97, 109

Nagel, Rosemarie. 1995. Unraveling in Guessing Games: An Experimental Study. *American Economic Review*, **85**(5), 1313–26. 17

Parfit, Derek. 1988 (1983). *What we together do.* Unpublished Manuscript. 108

Peacocke, Christopher. 2005. *Joint attention: Its nature, reflexivity, and relation to common knowledge.* Oxford: Oxford University Press. 41, 42

Pearce, David G. 1984. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica*, **52**(4), 1029–1050. 27, 35

Pinker, Steven, Nowak, Martin A., & Lee, James J. 2008. The logic of indirect speech. *Proceedings of the National Academy of Sciences*, **105**(3), 833–838. 37

Polak, Ben. 1999. Epistemic Conditions for Nash Equilibrium and Common Knowledge of Rationality. *Econometrica*, **67**, 673–676. 25, 160

Portner, Paul. 2007. Imperatives and modals. *Natural Language Semantics*, **15**(4), 351–383. 44

Pryor, James. 2000. The Skeptic and the Dogmatist. *Noûs*, **34**(4), 517–549. 122

Rubinstein, Ariel. 1989. The Electronic Mail Game: Strategic Behavior Under "Almost Common Knowledge". *The American Economic Review*, **79**(3), 385–391. 21, 22, 24

Rubinstein, Ariel, & Wolinsky, Asher. 1990. On the logic of 'agreeing to disagree' type results. *Journal of Economic Theory*, **51**(1), 184 – 193. 107, 137, 148

Samet, Dov. 1990. Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory*, **52**(1), 190–207. 137, 145, 148

Samet, Dov. 1992. Agreeing to Disagree in Infinite Information Structures. *International Journal of Game Theory*, **21**(3), 213–218. 145

Samet, Dov. 2010. Agreeing to disagree: The non-probabilistic case. *Games and Economic Behavior*, **69**(1), 169 – 174. 108, 109

Savage, Leonard J. 1954. *The Foundations of Statistics*. John Wiley and Sons. 108, 123

Savage, Leonard J. 1967. Difficulties in the Theory of Personal Probability. *Philosophy of Science*, **34**(4), 305–10. 135

Schiffer, Stephen R. 1972. *Meaning*. Oxford: Oxford University Press. 11

Shin, Hyun Song. 1993. Logical structure of common knowledge. *Journal of Economic Theory*, **60**(1), 1–13. 137, 148

Shin, Hyun Song, & Williamson, Timothy. 1994. Representing the knowledge of turing machines. *Theory and Decision*, **37**(1), 125–146. 137

Srinivasan, Amia. 2013. Are We Luminous? *Philosophy and Phenomenological Research*, n/a–n/a. 39

Stahl, Dale, & Wilson, Paul W. 1994. Experimental evidence on players' models of other players. *Journal of Economic Behavior & Organization*, **25**(3), 309–327. 17

Stahl, Dale, & Wilson, Paul W. 1995. On Players' Models of Other Players: Theory and Experimental Evidence. *Games and Economic Behavior*, **10**(1), 218–254. 17

Stalnaker, Robert. 1998. On the representation of context. *Journal of Logic, Language and Information*, **7**(1), 3–19. 43, 53

Stalnaker, Robert. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press. 86, 102

Stalnaker, Robert. 2002. Common ground. *Linguistics and Philosophy*, **25**(5), 701–721. 43, 53, 57

Stalnaker, Robert. 2006. On Logics of Knowledge and Belief. *Philosophical Studies*, **128**, 169–199. 31

Stalnaker, Robert C. 1974. Pragmatic presuppositions. New York: New York University Press. 43, 57

Stalnaker, Robert C. 1978. Assertion. *Pages 315–32 of:* Cole, P. (ed), *Syntax and Semantics*, vol. 9. New York Academic Press. 43, 86

Stalnaker, Robert C. 2009. On Hawthorne and Magidor on Assertion, Context, and Epistemic Accessibility. *Mind*, **118**(470), 399–409. 12, 30, 65, 86, 92, 106

Stalnaker, Robert C. 2014. *Context*. Oxford: Oxford University Press. 43, 53, 57, 60, 61, 93

Stiller, James, & Dunbar, Robin IM. 2007. Perspective-taking and memory capacity predict social network size. *Social Networks*, **29**(1), 93–104. 45

Strzalecki, Tomasz. 2010. *Depth of reasoning and higher order beliefs*. Harvard Institute of Economic Research, Harvard University. 22, 26

Van Eijck, Jan, & Kamp, Hans. 2011. *Representing discourse in context*. Elsevier. Chap. 3, pages 181–252. 44

Van Eijck, Jan, & Visser, Albert. 2012. Dynamic Semantics. *In:* Zalta, Edward N. (ed), *The Stanford Encyclopedia of Philosophy*, winter 2012 edn. 44

van Fraassen, Bas. 1984. Belief and the Will. *Journal of Philosophy*, **81**, 235–56. 121

Veltman, Frank. 1996. Defaults in update semantics. *Journal of Philosophical Logic*, **25**(3), 221–261. 44

von Fintel, Kai, & Gillies, Anthony. 2011. *Might Made Right*. Oxford University Press. 66

White, Roger. 2005. Epistemic Permissiveness. *Philosophical Perspectives*, **19**, 445–459. 97

White, Roger. 2013. *In:* Greco, John, Steup, Matthias, & Turri, John (eds), *Contemporary Debates in Epistemology*. Blackwell Publishing Ltd. 97

Williamson, Timothy. 1992. Inexact knowledge. *Mind*, **101**(402), 217–242. 117

Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford: Oxford University Press. 31, 38, 78, 98, 102, 103, 106, 110, 117, 119

Williamson, Timothy. 2007. How probable is an infinite sequence of heads? *Analysis*, **67**(3), 173–180. 97

Williamson, Timothy. 2013. Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier Cases in Epistemic Logic. *Inquiry*, **56**(1), 77–96. 76

Yalcin, Seth. 2007. Epistemic modals. *Mind*, **116**(464), 983–1026. 44, 48, 53, 57