PEOPLE WITH COMMON PRIORS CAN AGREE TO DISAGREE

HARVEY LEDERMAN
New York University

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" (Aumann, 1976, p. 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.

§1. Introduction. Robert Aumann presents 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" (Aumann, 1976, p. 1236). According to this *key conditional*, if two agents commonly know what probability they each assign to a proposition, their probabilities agree.[1] 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:

> "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."[2]

Received: September 1, 2013.

[1] I will say that agents "commonly know" that p or "commonly believe" that p to abbreviate "it is common knowledge among the agents" that p, and "it is common belief among the agents" that p.

[2] Morris (1995, pp. 227–228).

doi:10.1017/S1755020314000380

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."[3]

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

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

(2-Global Bayesianism) that agents commonly know that each of them updates by conditionalization;

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

(4-Global Truth) that agents commonly know that they update only on true propositions.[4]

I describe a class of models in which each of these assumptions can be made explicit, and then show that natural weakenings of any one of them allows for countermodels to the key conditional. Six 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 probability. The examples show that the *reductio* articulated by Morris and implicit in Aaronson's remarks requires these further, strong assumptions.

In fact, the examples show something subtler as well. Consider the following four claims:

(A1-Shared Prior) there is a common prior;

(A2-Bayesianism) agents update by conditionalization;

(A3-Positive Introspection) if an agent knows a proposition, she knows that she knows it;

(A4-Truth) each agent updates only on true propositions.

In my examples the common knowledge assumptions (1)–(4) will be relaxed, but all of (A1)–(A4) will be preserved. To separate these two kinds of assumptions, I will make extensive use of "pointed" models, which include the specification of a single, designated or "actual" state. In these models, all four of (A1)–(A4) hold at the actual state, while the agents still agree to disagree at that state.

To illustrate the conceptual import of this observation, consider Shared Prior (A1). When Aumann and others have motivated this "common prior assumption", they have argued that if rational agents have the same information or evidence, they have the same beliefs.[5] When one formalizes the common prior assumption, however, one typically considers a unique common prior defined "globally", at every state in a given model. This formal implementation has the consequence that agents do not just *have* a common prior, they commonly know, of some prior μ, that it is the common prior. (That is, it yields Global Uniqueness, (1), above.) But this new common knowledge assumption is not warranted by the original motivation; the informal idea – that if rational agents have the same evidence, they have the same beliefs – motivates only the claim that rational agents *in fact* update on the common prior, not that they know or commonly know what this prior is.

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

[4] When I present the formal models, I will state slightly different formalizations of these assumptions. For discussion of this difference, see the paragraph immediately following n. 21.

[5] E.g., Aumann (1976, p. 1237; 1987, pp. 7, 14; 1998, pp. 932, 937). The following passage is typical: "As we have said, the CPA expresses the idea that differences in probabilities should reflect differences in information only." (Aumann, 1998, p. 932).

A similar point holds for each of the four target assumptions (1)–(4). In each case, the usual informal motivations suggest at most that a given assumption holds *at the actual state*, not that the assumption is common knowledge, nor that it obtains at every state in the model. Using pointed models allows us to implement (A1)–(A4) in our models, without also making the unwanted assumptions (1)–(4). And when we do so, the paradoxical result of the Agreement Theorem disappears.

1.1. Epistemological motivation. A main aim of this paper is to make problems related to the Agreement Theorem more accessible to philosophers. To this end, 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, and readers primarily interested in the formal observations may wish to focus on Section 3, after acquainting themselves with the models in Section 2.[6]

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.[7]

In this paper, I will focus on a Bayesian elaboration of the Uniqueness Thesis, which corresponds to the conjunction of Shared Prior (A1) and Bayesianism (A2):

Bayesian Uniqueness: Necessarily, there is some prior μ, such that for any agent i and 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]

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

[6] Sections 4 through 6 may also be skimmed with an eye to the "pointed" interpretation of the models presented there. Countermodels based on relaxing Positive Introspection have been implicit in the literature since Cave (1983) and Bacharach (1985): cf. e.g. Geanakoplos (1989), Samet (1990), and Rubinstein & Wolinsky (1990). Countermodels based on relaxing the Truth axiom have been equally well-known: see Bonanno and Nehring (1998), Collins (1997), and now Hellman (2012). For specialists, the new observation of these sections will be that it is consistent with agreeing to disagree that (A1)–(A4) hold. The discussion of Moses and Nachum's (1990) problem may also be of interest, and is found in §8, n. 37.

[7] White (2014), cited in Kelly (2014). An earlier version is found in White (2005, p. 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). I interpret the thesis as prefixed with "necessarily" since otherwise if the conditional is material, then if no one is fully rational in taking a given doxastic attitude, the principle is true, but vacuously so.

[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 nonempty $E \in \mathscr{P}(\Omega)$. Thus I will be abstracting from substantial issues about undefined conditional probabilities. For some discussion, see Williamson (2007), Hájek (2010), and Easwaran (2014).

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]

Aumann's key conditional can be used as a premise in a *reductio* of Bayesian Uniqueness which is directly analogous to Morris's argument, quoted above. 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, it is claimed, 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 the jurors may have different opinions about the defendant's guilt. But these disagreements do not appear to give us grounds to fault the scientists' or jurors' rationality.

The argument from Agreement threatens more than just Bayesian Uniqueness. For one thing, it is not restricted to Bayesians. §8 discusses in detail how a version of the argument could be given for alternative views of the kinds of attitudes rational agents should have, given their evidence (and even for alternative views of the nature of evidence).

But the argument also applies to much more conservative Bayesian theories than Bayesian Uniqueness. Some will reject Bayesian Uniqueness in its most general form, but still hold that for specific groups of agents, and when confined to certain subject matters, a version of Bayesian Uniqueness holds. Many subjective Bayesians, for example, hold that although *some* agents' priors *sometimes* differ, very often or at least when restricted to

[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 an "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 I will call "the agent's representer on P". Thus:

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 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. The Agreement Theorem thus applies straightforwardly to Mushy Uniqueness.) I will not explore these 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.

propositions in certain subject matters, agents do have the same priors. These subjective Bayesians would hold that, for a wide range of specific sets of agents N and specific subject matters $\mathcal{P}$:[10]

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

It's plausible that the scientists in our earlier example form such an N, while their area of research forms such a subject matter $\mathcal{P}$. Supposing these scientists were educated similarly, they would have formed the same priors by the end of their training. Their views about their area of research should then be formed by conditionalizing the scientific evidence on this prior, whether the prior is understood as a probability function which represents the attitudes they had at this earlier time, or simply as a representation of their current dispositions to revise their beliefs in light of evidence they might receive. But since such scientists often "agree to disagree", the argument from Agreement shows either that they are irrational, or that, contrary to what one might have thought, their parallel education did not leave them with the same priors. If the argument is sound, then it imposes a surprising and severe restriction on the prevalence of shared priors. It shows that common priors are only as common as rational agreeing to disagree is rare.

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 Moderate Bayesian Uniqueness applies only to domains in which rational agents do not agree to disagree. The countermodels to the key conditional will then offer the theorist of (Moderate) 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 two responses to the argument from Agreement. The first is simply to accept the conclusion of the argument: some of the scientists and jurors in my examples *are* irrational, and this irrationality becomes evident in their public disagreements. This reply has the cost of greatly limiting the applicability of the target notion of rationality. The scientists and jurors in these examples are intended to represent the greatest efforts humans can make to assess their evidence. If these agents are not doxastically or epistemically rational, it is unclear who is. This cost may or may not be a deep problem for the reply under consideration, but in any case I will simply set the reply aside, and continue to speak as if rational agreeing to disagree *is* possible, and that an epistemological theory should seek to accommodate this datum.

According to a second response, the scientists or jurors in the above examples may disagree not because they are irrational, but because they do not in fact commonly know one another's opinions. This response is more promising than the one just mentioned; I too believe common knowledge occurs more rarely than is often claimed. But still, I won't consider this line of response further in the paper. Many of those working in related areas

[10] More fully, a subject matter $\mathcal{P}$ is the algebra of propositions generated by a set of propositions which are taken as atomic in a given domain, so that the subject matter itself is guaranteed to form a complete Boolean algebra.

believe that common knowledge is extremely common. If denying that these scientists or jurors have achieved common knowledge represents the only resort of the Uniqueness Theorist, Aumann's Agreement Theorem would already have provided a surprising result: that the Uniqueness Thesis is inconsistent with standard views about the prevalence and attainability of common knowledge.

1.2. Outline. Section 2 sets up the basic model, and introduces Aumann's result. Sections 3–6 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.

Appendix A (Section 8) uses 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). The Appendix also states a qualitative Agreement Theorem, which uses only $S4$ (not $S5$) for knowledge. The proofs are in Appendix B (Section 9).

§2. The Agreement Theorem.

2.1. Aumann frames. An *agreement frame* will be a structure

$$\mathcal{F} = \langle \Omega, N, (P_i)_{i\in N}, (\pi_i)_{i\in N}, (p_i)_{i\in N} \rangle$$

containing:

- a finite set Ω of "states" or "worlds";
- a countable set, N, of "agents";
- a set of "possibility correspondences" (one for each agent) $P_i : \Omega \to \mathscr{P}(\Omega)$;
- a set of "prior" assignment functions (one for each agent) $\pi_i : \Omega \to \Delta(\Omega)$ (where $\Delta(\Omega)$ is the set of probability measures on $(\Omega, \mathscr{P}(\Omega))$);
- a set of "posterior" assignment functions (one for each agent) $p_i : \Omega \to \Delta(\Omega)$.

I will restrict attention to frames in which the set of states, Ω, is finite; this ensures that my countermodels will not turn on irrelevant paradoxes of infinity. Informally, in this finite setting, we think of worlds as coarse-grainings of the space of maximally specific situations, by equivalence classes taken with respect to what is relevant in the case we are modeling.[11]

[11] The assumption that the state space is finite is in fact an important one mathematically. For example, Samet (1992) shows that infinite state spaces allow countermodels to the key conditional in 4-Aumann models (defined below) Since it is natural to think that epistemic state spaces will in general be infinite, one might conclude that considering only finite state spaces closes off an important reply for the proponent of Bayesian Uniqueness, namely, to invoke counterexamples in infinite spaces à la Samet. But I think matters are not so clear. First, we typically think of worlds in these models as what Savage called "small worlds"; these small worlds represent only those aspects of the situation which are relevant to the case at hand. This space of *small worlds* may well be finite; the distinctions we do draw in any given situation is not obviously as large as the full space of epistemic possibilities we are capable of considering. Second, counterexamples such as Samet's require a very specific pattern in the possible knowledge states of an agent – an infinite descending chain of ever more precise knowledge. While one can concoct interpretations to match the formalism, it is difficult to imagine that such situations are sufficiently prevalent in real life to explain the prevalence of public disagreements. So while the assumption of finiteness is not entirely innocent, I will not discuss it further.

The possibility correspondences P_i are used to describe each agent's evidential state at each world: evidence is thought of as a set of propositions, where a *proposition* is itself a set of states, $E \subseteq \Omega$.[12] A proposition belongs to an agent's evidence at a state if and only if it is a superset of the value of her possibility correspondence at that state ($P_i(\omega) \subseteq E$).[13] In the finite frames we will be considering, $P_i(\omega)$ itself is the conjunction of the propositions which belong to the agent's evidence at the state ω; when we turn to probabilities, a Bayesian rational agent i will update by conditionalizing his or her prior at ω on the proposition $P_i(\omega)$, in accordance with one of the main demands of Bayesian Uniqueness.

Four axioms on agents' evidence will be crucial in what follows. For a given agreement frame $\mathcal{F}$, a given state $a \in \Omega$ satisfies

Consistency iff $(\forall i \in N)(P_i(a) \neq \emptyset)$

Truth iff $(\forall i \in N)(a \in P_i(a))$

Positive Introspection iff $(\forall i \in N)(\forall \omega \in \Omega)(\omega \in P_i(a) \Rightarrow P_i(\omega) \subseteq P_i(a))$

Negative Introspection iff $(\forall i \in N)(\forall \omega \in \Omega)(\omega \in P_i(a) \Rightarrow P_i(a) \subseteq P_i(\omega))$

Consistency says that, at a, each agent's evidence is consistent. Truth says that if a proposition belongs to an agent's evidence at a, it is true; Truth entails Consistency. Positive Introspection says that if an agent's evidence contains a proposition E at a, then the proposition that the agent's evidence contains E is also part of the agent's evidence at a. If an agent's evidence is what she knows, then this condition is the Positive Introspection condition discussed in the introduction: if an agent knows a proposition, she knows that she knows it. This principle is thus a simple-minded version of the so-called "KK" principle. Finally, Negative Introspection says that if an agent's evidence does not contain a proposition E, the proposition that the agent's evidence does not contain E is part of the agent's evidence.

For each of these "pointwise" definitions, we can also give "global" analogs. An agreement frame $\mathcal{F}$ satisfies Global Consistency just in case every state satisfies Consistency; similarly a frame satisfies Global Truth, Global Positive Introspection or Global Negative Introspection if and only if every world satisfies Truth, Positive Introspection or Negative Introspection, respectively. In this paper, I will not need a language or a logic, so I won't introduce one formally. But for those with a background in epistemic logic, it is worth observing that these "global" conditions correspond to familiar logical ones: given the usual interpretation of modal languages into such frames and the usual definition of validity, frames which satisfy Global Consistency will validate the modal axiom **(D)** for each operator; those which satisfy Global Truth will validate the modal axiom **(T)**, and those which satisfy Global Positive and Global Negative Introspection will, respectively,

[12] 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 in this way. Nevertheless, I adopt this simplifying, idealized assumption throughout, in accordance with the results I will be discussing.

[13] Philosophers will be more familiar with epistemic or doxastic accessibility relations in place of possibility correspondences. The two formalisms are equivalent: we can define accessibilty relations R_i in terms of possibility correspondences as follows: $\omega R_i \omega' := \omega' \in P_i(\omega)$. Alternatively, we could have taken the R_i as primitive, and used them to define possibility correspondences: $P_i(\omega) := \{\omega' \in \Omega \mid \omega R_i \omega'\}$.

validate the modal axioms **(4)** and **(5)**.[14] From here on, I will take it as part of the definition of an agreement frame that it satisfies Global Consistency.

Officially, we interpret the possibility correspondences abstractly as representing the agents' evidence in the way just described. Unofficially, however, it will be useful to have a more concrete understanding of this idea. Thus, when a frame satisfies Global Truth, I will interpret it as representing the agents' knowledge. In particular, I will say that agent i knows E at a just in case $P_i(a) \subseteq E$. On this interpretation the frames represent the view of evidence in Williamson (2000, chap. 9), where a proposition belongs to an agent's evidence just in case the agent knows it. On the other hand, if the frames do not satisfy Global Truth, I will interpret the models as representing belief, and once again say that i believes E at a just in case $P_i(a) \subseteq E$. It deserves emphasizing, however, that the models themselves do not force us to either of these interpretations. Thus, although I will discuss the models using "know" and "believe", my official view will still be that the possibility correspondences represent facts about agents' evidence. Those who have a particular view about the nature of evidence should substitute their preferred ideology 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).[15]

Epistemic models make a number of substantial idealizations about knowledge and belief. Models of the kind used here make three strong assumptions in particular. (For the remainder of this paragraph and the next, I will speak only of knowledge, but what I say carries over straightforwardly to belief.) 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, at every state, every agent knows the universe, Ω.

If the states in our space are interpreted as "logically possible" worlds, the third of these assumptions corresponds to an assumption about agents' "logical omniscience" – every agent knows the tautology, and hence knows every logical truth. All but a small minority of those working in epistemic logic reject the claim that people's knowledge exhibits logical omniscience of this last kind.[16] There are in the literature a variety of well-known ways of responding to this problem, each of which merits considerable discussion in its own right. But for the purposes of this paper, I will simply assume that agents are in fact logically omniscient, since making this assumption will strengthen and not diminish the force of my counterexamples. Logically omniscient agents cannot disagree about questions of logic; my counterexamples to the key conditional will thus be more powerful because they exhibit logically omniscient agents who, in spite of their omniscience, agree to disagree.

[14] See e.g. Fagin *et al.* (1995) for the details. Frames which satisfy Global Consistency, Global Positive Introspection and Global Negative Introspection are often called $KD45$ in honor of the modal logic which would be valid on this class of frames, given the standard definition of validity and the standard rules for interpreting modal logics in them. For a similar reason, frames which satisfy Global Truth and Global Positive Introspection are often called $S4$; those which satisfy Global Truth and Global Negative Introspection (which entails that they satisfy Global Positive Introspection) are often called $S5$ or "partitional" (since in such frames $\mathcal{P}_i = \{P_i(\omega) | \omega \in \Omega\}$ partitions the state space).

[15] Those who do not think evidence is propositional at all should refer to Appendix A (Section 8) for discussion of how related models can be used to represent their views.

[16] The most notable exception is Stalnaker (1999, pp. 241–273). See Halpern and Pucella (2011) for a recent survey of technical work in the area.

Turning now to the prior assignment functions, an agreement frame $\mathcal{F}$ satisfies

Regularity iff $(\forall i \in N)(\forall \omega, \omega' \in \Omega)(\pi_i(\omega)(\omega') > 0)$.[17]

In what follows, I will again take it as part of the definition of an agreement frame that it satisfies Regularity. Even when other assumptions are stated explicitly, this (along with Global Consistency) will always be assumed.

Now we consider two different assumptions related to the notion of the common prior. Given an agreement frame $\mathcal{F}$, a state $a \in \Omega$ satisfies

Shared Prior iff $(\forall i, j \in N)(\pi_i(a) = \pi_j(a))$.

A frame $\mathcal{F}$ satisfies Global Shared Prior just in case every state satisfies Shared Prior.

A second, independent assumption governs the relationship between an agent's prior at different worlds. An agreement frame $\mathcal{F}$ satisfies

Constant Prior iff $(\forall i \in N)(\forall \omega, \omega' \in \Omega)(\pi_i(\omega) = \pi_i(\omega'))$.

A frame $\mathcal{F}$ satisfies Global Uniqueness just in case $\mathcal{F}$ satisfies both Constant Prior and Global Shared Prior.

Now, finally, we consider how the assignments of posterior probabilistic beliefs p_i relate to the agents' evidence as described by the P_i and their priors as described by the π_i. First, a frame $\mathcal{F}$ satisfies

Coherence iff $(\forall i \in N)(\forall \omega, \omega' \in \Omega)(p_i(\omega)(\omega') > 0 \Leftrightarrow \omega' \in P_i(\omega))$.

From here on, it is taken to be part of the definition of an agreement frame that it satisfies Coherence, in addition to Regularity and Global Consistency.

Now given a frame $\mathcal{F}$, a given state $a \in \Omega$ satisfies

Bayesianism iff $(\forall i \in N)(p_i(a) = \pi_i(a)(\cdot|P_i(a))$.

A frame $\mathcal{F}$ satisfies Global Bayesianism just in case every state $\omega \in \Omega$ satisfies Bayesianism. In addition, we have corresponding notions for individual agents: given a frame $\mathcal{F}$, an agent $i \in N$ is a Bayesian at a state $a \in \Omega$ just in case $p_i(a) = \pi_i(a)(\cdot|P_i(a))$; i is a Global Bayesian just in case $(\forall \omega \in \Omega)(p_i(\omega) = \pi_i(\omega)(\cdot|P_i(\omega))$.

It should be clear that all agents can be rational according to Bayesian Uniqueness at a state a only if the state satisfies Shared Prior and Bayesianism. For if either of these assumptions failed at a state, then it could not be that all agents were updating appropriately on the rational prior at that state. I will also speak as if these assumptions were sufficient for the satisfaction of Bayesian Uniqueness. Since I do not know what the rational prior is (or whether there is one at all), I do not know whether the agents in my examples are updating on the rational prior. But I assume that if there is a rational prior, then whatever that prior is, it would admit examples which have the structure that mine do.

The examples in this paper will be given using *pointed* frames, where a pointed agreement frame is a structure

$$\mathcal{F}_a = \langle \Omega, a, N, (P_i)_{i \in N}, (\pi_i)_{i \in N}, (p_i)_{i \in N} \rangle$$

[17] Here, and throughout, I engage in a standard abuse of notation, writing $\pi_i(\omega)(\omega')$ instead of $\pi_i(\omega)(\{\omega'\})$, and $p_i(\omega)(\omega')$ instead of $p_i(\omega)(\{\omega'\})$. I will similarly sometimes speak informally of these distributions assigning values to worlds, instead of to the singletons containing them. I also write $\pi_i(a)(\cdot|E)$ to mean the distribution obtained by conditionalizing $\pi_i(a)$ on E, using the standard ratio definition of conditionalization.

where $a \in \Omega$, and the other elements are as in an agreement frame. The designated element a is interpreted as the "actual world".

With these notions in hand, we can define Aumann's original class of frames:

DEFINITION 2.1. *An agreement frame $\mathcal{F}$ is a* 4*-Aumann frame if and only if it satisfies*

(i) *Global Truth, Global Positive Introspection;*
(ii) *Shared Prior;*
(iib) *Constant Prior; and*
(iii) *Global Bayesianism.*

A 4*-Aumann frame is an Aumann frame if and only if it satisfies Global Negative Introspection.*

Aumann's original models satisfied both Global Negative Introspection and Global Truth (they were "$S5$" frames). But a straightforward generalization of his original theorem can be proven for 4*-Aumann frames*.[18]

Now we turn to the class of models from which the counterexamples to the key conditional will be drawn:

DEFINITION 2.2. *A pointed Aumann frame is a pointed agreement frame in which* a *satisfies*

(i) *Truth, Positive Introspection, Negative Introspection;*
(ii) *Shared Prior;*
(iii) *Bayesianism.*

Every example except the one in Figure 4 (in which Truth fails at every state) will be a pointed Aumann frame. So long as (ii) and (iii) are assumed, these pointed frames represent agents who are rational according to Bayesian Uniqueness.

In addition to the beliefs and knowledge of a single agent, we will need to describe certain relationships between the knowledge and belief of many agents. Although it is easy to define these notions relative to an arbitrary group of agents, I will further simplify matters by only considering the group which consists of all agents in a given model. I will first define what the agents mutually know^n inductively. (For the remainder of this section, I will speak only of "knowledge" in the main text, but all of the definitions I give carry over if "believe" is substituted for "know", and the reader should supply these alternatives.) The agents *mutually know*^{1} (or simply, mutually know) a proposition E at a world ω if and only if every agent in the group knows E at ω. In symbols, we define $P_N^1(\omega) = \cup_{i \in N} P_i(\omega)$. A group mutually knows^n a proposition E if and only if the group mutually knows that they mutually know^{n-1} E; or, again, $P_N^n(\omega) = \cup_{\omega' \in P_N^{n-1}(\omega)} P_N^1(\omega')$. The group *commonly knows* E if and only if, for all natural numbers n, they mutually know^n it: $P_N^{CK}(\omega) = \cup_{n \in \mathbb{N}} P_N^n$.

In the finite models we will be considering here, this infinitary definition of common knowledge is equivalent to a more elegant one in terms of what are called

[18] In fact, we can prove a more technical result, which generalizes even this one, using a property I call "local balancedness", defined in my (2014). Thus the examples in Sections 5 and 6 will describe 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.

self-evident propositions.[19] Given a frame $\mathcal{F}$ and an agent $i \in N$, a non-empty proposition E is *self-evident* to an agent i if and only if, if E obtains, i knows that it obtains $((\forall \omega \in E)(P_i(\omega) \subseteq E))$. When the P_i are interpreted as representing knowledge, such self-evident propositions are akin to what are called "luminous" propositions in Williamson (2000, p. 95).[20] A non-empty proposition is then *public* if and only if it is self-evident to all agents. Note that if two public propositions are not disjoint, their intersection will also be a public proposition.

Using these definitions, we can show that in any frame $\mathcal{F}$, a proposition $F \in \mathscr{P}(\Omega)$ is commonly known at a world ω if and only if there is some public proposition E such that, for every $i \in N$, E entails that i knows F $((\forall i \in N)(E \subseteq \{\omega' \in \Omega : P_i(\omega') \subseteq F\}))$.[21] Given this theorem, it is clear that common knowledge will be closed under conjunction: if the proposition that everyone knows F is entailed by some public proposition E such that $\omega \in E$ and if the proposition that everyone knows H is entailed by some public proposition G such that $\omega \in G$, then $E \cap G$ is non-empty (it contains ω), is public, and entails the proposition that everyone knows F and everyone knows H. But since everyone's knowledge is closed under conjunction, this last proposition entails that everyone knows F and H.

To this point, I have spoken as if making a "global" assumption is equivalent to making a "common knowledge" assumption. Thus, for example, I said that Global Uniqueness is satisfied if and only if a frame satisfies Global Shared Prior and Constant Prior, even though Global Uniqueness was originally stated as: "for some μ, the agents commonly know that the prior is μ." In fact, however, these global assumptions are slightly stronger than their common knowledge counterparts. In every agreement frame, the universe Ω is commonly known at every state; thus any "global" assumption is also assumed to be common knowledge. But the other direction does not hold, at least not without qualification. There are some models in which agents commonly know a proposition at a state even if that proposition is not equal to the universe. But this slight difference in strength won't matter in what follows: when one of my examples fails a condition imposed at all states, it will also fail the corresponding condition stated in terms of common knowledge. In fact, since all of my examples will preserve three out of the four Global assumptions, defining these at holding at all states will make my task harder, and not easier. But, in any case, to simplify the discussion, I'll continue to speak as if the Global assumptions as formalized here are equivalent to the ones stated using common knowledge in the Introduction.

In presenting the Agreement Theorem, it is convenient to have a short-hand for defining propositions by reference to the agents' posterior probabilities. In particular, the proposition that i assigns E probability k will be denoted as follows: $[p_i(E) = k] = \{\omega : p_i(\omega)(E) = k\}$.

We then state a slight generalization of Aumann's original theorem, presented in various forms by Bacharach (1985), Cave (1983), Geanakoplos (1989), Rubinstein & Wolinsky (1990), and Samet (1990):

[19] 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, p. 52–57).

[20] Williamson uses "in a position to know" in place of "know".

[21] In models of knowledge, where Global Truth 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 Global Truth 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 $E \subseteq F$.

Theorem 2.3 (Aumann 1976). *Let $\mathcal{F}$ be a 4-Aumann frame. If there is a state $\omega \in \Omega$, a proposition E, and for each i, some $k_i \in [0, 1]$ such that $[p_i(E) = k_i]$ is common knowledge at ω, then for all $i, j \in N$, $k_i = k_j$.*

The basic idea of the proof is as follows. Say that an agent i assigns constant probability to E within F if and only if there is some k_i such that $F \subseteq [p_i(E) = k_i]$. We first show that if a Global Bayesian agent who satisfies Constant Prior assigns constant probability to E within a self-evident F, then $(\forall \omega \in F)(\pi_i(\omega)(E|P_i(\omega)) = p_i(\omega)(E) = \pi_i(\omega)(E|F)$. This lemma is the "hard" part of the proof: we have to bridge the gap between the agent's posterior probabilities at states within F, where her evidence may differ at different states, and the probability of her prior conditionalized on F itself. (The details can be found in §9, where I prove the theorem as a corollary of a generalization of known qualitative agreement theorems.)

But once this is done, the remaining work is merely a question of transforming definitions. The antecedent of the key conditional says that there is some ω such that for each i, there is some k_i so that $[p_i(E) = k_i]$ is commonly known at the state ω. Thus, by hypothesis and the "elegant" formulation of common knowledge, for each agent there is a public proposition F_i such that the agent assigns constant probability to E within F_i; in fact, in our finite frames and given that common knowledge is closed under finite conjunctions, there is guaranteed to be a public proposition $F = \bigcap_{i \in N} F_i$ such that all agents assign constant probability to E within F. By definition, public propositions are self-evident to all agents, so we can use our lemma to show that, for each agent i, and every $\omega' \in F$, $\pi_i(\omega')(E|F) = k_i$. But since F is the same for all agents, and since for all states, whether in F or not, $\pi_i(\omega') = \pi_j(\omega')$, it follows immediately that for all $i, j \in N$, $k_i = k_j$.

§3. Common knowledge of common priors.

In Aumann frames, all agents have the same prior and update by conditionalizing on that prior. This assumption is often called the common prior assumption. But the name is misleading. The implementation of the common prior assumption in these models entails not just that agents have the same prior, but that they commonly know which prior they share (and commonly know that they update by conditionalizing on that prior). In other words, it entails not just that the agents satisfy Bayesian Uniqueness, but that they commonly know that they do.

This further assumption is not stated explicitly in the key conditional. Nor was it represented explicitly in Aumann's original class of frames, even though Aumann's models guarantee that it holds. In our more general agreement frames, by contrast, we can represent the assumption explicitly, and thus consider the consequences of relaxing it. In this section, I will do this in two stages. First, I will present a simple example of two agents who are rational by the lights of Bayesian Uniqueness, but who do not commonly know that they are. These agents have the same prior, but since they do not commonly know that they do, they agree to disagree. This first example illustrates how common knowledge of a strong form of rationality (Bayesian Uniqueness) is implicitly assumed in Aumann frames. Then, after some conceptual discussion, I will show that even common knowledge of Bayesian Uniqueness is not enough to restore the theorem: agents who commonly know that they satisfy the constraints of Bayesian Uniqueness may still agree to disagree, if they do not commonly know which prior they share (that is, if Constant Prior fails).

3.1. Common knowledge of rationality: First example.

Consider the frame with two agents, Evens and Odds ($N = \{e, o\}$), and six states ($\Omega = \{1, 2, 3, 4, 5, 6\}$) represented

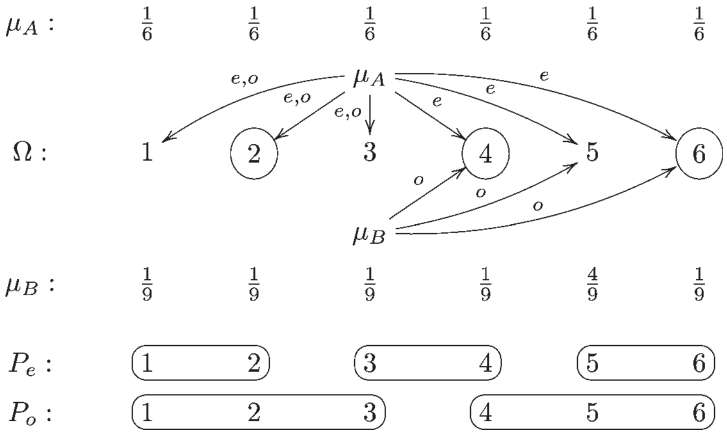

μ_A :
$\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$ $\frac{1}{6}$
μ_A
e,o
e,o
e,o
e
e
e
Ω :
1 2 3 4 5 6
o
o
o
μ_B
μ_B :
$\frac{1}{9}$ $\frac{1}{9}$ $\frac{1}{9}$ $\frac{1}{9}$ $\frac{4}{9}$ $\frac{1}{9}$
P_e :
1 2 3 4 5 6
P_o :
1 2 3 4 5 6

Fig. 1. Evens and odds: No common knowledge of shared priors.

in Figure 1. Evens has a constant prior assignment function: $(\forall\omega)(\pi_e(\omega) = \mu_A)$, where μ_A is the uniform prior which assigns every state $1/6$ $((\forall\omega \in \Omega)(\mu_A(\omega) = 1/6))$. Evens also has a simple possibility correspondence, which takes every state to a pair of worlds, as follows: for $n \in \{1, 2, 3\}$, $P_e(2n) = P_e(2n - 1) = \{2n, 2n - 1\}$. Evens is a global Bayesian: at each state, her posterior probabilities accord with her prior conditionalized on her evidence: $(\forall\omega \in \Omega)(p_e(\omega) = \pi_e(\omega)(\cdot|P_e(\omega))$.

Odds is only slightly more complex. At states 1, 2, 3, his prior agrees with Evens' prior $(\omega \in \{1, 2, 3\} \Rightarrow \pi_o(\omega) = \pi_e(\omega))$. But at states 4, 5, 6, they have different priors; at these states Odds' prior is μ_B, which assigns $1/9$ to every world, *except* 5, which Odds thinks is more likely than the other states, assigning it probability $4/9$. Odds' possibility correspondence comes in triples: for $\omega \in \{1, 2, 3\}$, $P_o(\omega) = \{1, 2, 3\}$, while for $\omega \in \{4, 5, 6\}$, $P_o(\omega) = \{4, 5, 6\}$. Like Evens, Odds is a Global Bayesian: $(\forall\omega \in \Omega)(p_o(\omega) = \pi_o(\omega)(\cdot|P_o(\omega))$.

Now it is easy to check that if we choose the designated world a to be $a \in \{1, 2, 3\}$, the resulting pointed model will be a pointed Aumann frame. (It's also easy to check that even without a designated point, the frame satisfies Global Truth, Global Negative Introspection, and Global Bayesianism; it fails to be an Aumann frame only because it fails Global Shared Prior and Constant Prior.) But note that no matter what we choose as the actual state, the agents fail to commonly know that they have the same prior. In the pointed frame where $a = 1$, for example, Odds knows that he and Evens have the prior μ_A. But he does not know that Evens knows that they share this prior. For Odds assigns positive posterior probability to $\{3\}$ and, at 3, Evens would assign positive probability to Odds having the prior μ_B, while Evens herself would still have the prior μ_A. As a result of this failure of common knowledge, the agents can agree to disagree, as I will now show.

Consider the proposition $E = \{2, 4, 6\}$, the elements of which are circled in Figure 1. It is again trivial to see that the proposition that [Evens assigns E probability $1/2$] is just the universe $([p_e(E) = 1/2] = \Omega)$. It is equally trivial to see that the proposition that [Odds assigns E probability $1/3$] is also the universe $([p_o(E) = 1/3] = \Omega)$. Since the universe is self-evident to every agent, these two facts are commonly known at every state, and thus the agents agree to disagree.

3.2. Common knowledge of shared priors. This first example already demonstrates that the key conditional does not hold in general on the class of pointed Aumann frames. In the pointed frame with $a = 1$, the agents are both rational according to Bayesian

Uniqueness: they share a prior, and they update by conditionalization. Moreover, as noted above, the frame satisfes Global Positive Introspection, Global Negative Introspection, Global Truth *and* Global Bayesianism. But still, the agents agree to disagree.

One diagnosis of what has gone wrong is that the agents in the example do not commonly know that they satisfy Shared Prior; thus they do not commonly know that they are both rational in the sense of Bayesian Uniqueness. In light of this observation, one might suspect that the Agreement Theorem corresponds to the following weakened version of the key conditional: "If agents have a common prior, *and commonly know that they are rational in the sense of Bayesian Uniqueness* (that is, that they satisfy Shared Prior and Bayesianism), then they cannot agree to disagree." A new argument from Agreement based on *this* conditional might still seem to pose a problem for (Moderate) Bayesian Uniqueness. Supposing our earlier scientists attended similar schools, read similar books and so on, it might seem plausible that they could commonly know that they have the same prior and that they update by conditionalizing this prior. To be sure, the additional requirement of common knowledge of rationality makes the argument considerably less general, but it might still seem worrisome.

But notice that pointed Aumann frames which satisfy Global Shared Prior, Global Bayesianism, Global Positive Introspection and Global Truth need not be Aumann frames, since they may fail Constant Prior. So we cannot simply apply the Agreement Theorems to this new class of frames: at this point, it remains open whether the agents represented by this class of frames can or cannot agree to disagree.

But before we turn to this this formal question, we face an antecedent conceptual issue. Frames in which Global Shared Prior holds but Constant Prior fails, are frames in which agents commonly know that they have the same prior, but don't know which prior that is. Can the theorist of Bayesian Uniqueness allow that agents who commonly know that they are rational (and thus commonly know that they have the same prior) fail to commonly know what the uniquely rational prior is?[22]

The following argument suggests that they cannot. According to this argument, the theorist of Bayesian Uniqueness should hold that the prior can be known *a priori*, and that ideally rational agents know all *a priori* truths of this sort. In the idiom of the theorist of Bayesian Uniqueness, the claim is that the unique prior μ itself must assign the proposition *that* μ *is the unique prior* probability 1. If an agent assigns a proposition probability 1, conditionalizing cannot alter her probability in that proposition, so by the lights of this version of Bayesian Uniqueness any agent who assigns the proposition *that* μ *is the unique prior* less than probability 1 will fail to be fully rational.

This argument consists of two claims, each of which can reasonably be rejected. It is certainly compatible with Bayesian Uniqueness to hold that the prior is knowable *a priori*. But this position is also not mandatory. One could equally well hold that facts about which prior is the uniquely rational one depend on facts which cannot be known *a priori*, for example, facts about the chances, which in turn depend on the physical laws. You and I may commonly know that we are rational, but if we are ignorant of the physical facts which determine the chances, we may fail to commonly know which prior is the uniquely

[22] Note that, in finite frames, even if we relax the assumption that $\pi_i = \pi_j$, there will still be substantial common knowledge about which priors all agents have. If we move to infinite frames, and allow $\pi_i \neq \pi_j$, 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 $\pi_i(\omega)(E) = x$. Thanks to an anonymous referee here.

rational one. Since according to this position rational agents can assign positive probability to the prior being different than it actually is, it follows immediately that proponents of this version of Bayesian Uniqueness will also reject the claim that the prior μ assigns the proposition *that μ is the unique prior* probability 1.

But even if one accepts this first premise – that the prior itself is knowable *a priori* – one may still reject the second – that if the prior is knowable *a priori*, ideally rational agents will know it. The theorist of Bayesian Uniqueness may wish to allow for it to be rational to assign probability less than 1 to some *a priori* truths, for example, complex mathematical truths such as Goldbach's conjecture or its negation. (There's a question about how to allow for this in the presence of a form of logical omniscience, but all we need to motivate the present point is the datum that this could be rational, and not a story about how to represent that fact.) If one is allowed to assign probability less than 1 to *some a priori* knowable truths, then it is surely coherent to allow that it is rational to assign probability less than 1 to the proposition *that μ is the prior*, even if this is knowable *a priori*. Once again, Uniqueness theorists of this second kind will similarly reject the claim that the prior μ assigns the proposition *that μ is the unique prior* probability 1.

Not only are these positions coherent, they may well be attractive to the Uniqueness theorist as ways of softening what would otherwise be an extremely restrictive theory of rationality. I myself am not certain of the proposition that [the uniquely rational prior assigns the proposition that spacetime is supersymmetric probability $< .5$]. I am also not certain of its negation. In fact, it seems to me that anyone who is certain of this proposition (or its negation) is irrationally overconfident. If the theorist of Bayesian Uniqueness wishes to respect this kind of judgment, then she will have to reject any principle which entails that rational agents must be certain of (or, in my idiom, know) the prior.

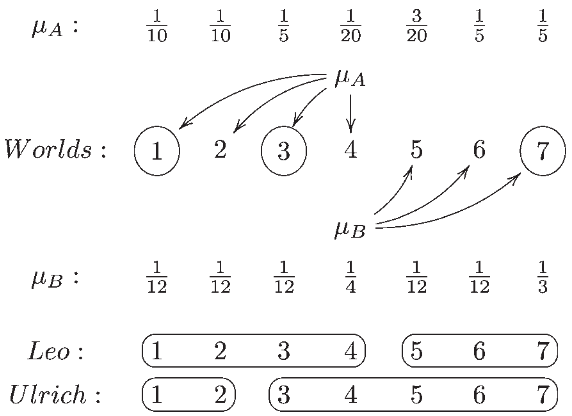
So far, we have focused on Bayesian Uniqueness, but what about subjective Bayesians who endorse Moderate Bayesian Uniqueness? I noted in the Introduction that the argument from Agreement affects these more subjective Bayesian theories, too, so long as they hold that shared priors are more common than disagreements among agents who commonly know each others' opinions. Like the theorist of Bayesian Uniqueness, this subjective Bayesian may believe that even fully rational agents who commonly know that they share a prior may fail to commonly know which prior they share (however the idea of a prior is understood in this context).[23] So this subjective Bayesian will also wonder whether agents

[23] Subjective Bayesians often interpret priors and update by conditionalization literally, holding that priors are distributions agents have at some time, and that, at a subsequent time, agents have a distribution obtained by conditionalizing the earlier, "prior" one. But subjective Bayesians may understand the combination of a common prior and update by conditionalization in other ways. One is that the prior represents agents' dispositions to form beliefs if they should come to have less information than they currently have, or if they should learn something inconsistent with what they are currently certain of. On this view, an agent's prior records what her confidences would be in various learning scenarios; there is no assumption that the prior is a distribution the agent ever has at a time. It should be clear that on this view of priors agents may fail to know their own priors by failing to know how their beliefs would change if they had different information. Thus agents might commonly know that they have the same prior, but fail to commonly know which prior it is.

In line with this alternative understanding of priors, a large literature has developed an interpretation of the *common* prior 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 agree in expectation on a class of random variables. (For extensions of this idea and variations on it, see also Bonanno and Nehring, 1997, 1999; Halpern, 2002; Heifetz, 2006;

who commonly know that they share a prior, and commonly know that they update by conditionalization can nevertheless agree to disagree.

3.3. Second example. I'll now show that such agents can in fact agree to disagree. Consider the following frame with seven worlds ($\Omega = \{1, 2, 3, 4, 5, 6, 7\}$) and two agents, Leo and Ulrich ($N = \{l, u\}$), as depicted in Figure 2. At worlds $1, 2, 3$ and 4, both Leo and Ulrich have the prior μ_A, which is specified in the figure (formally, $\omega \in \{1, 2, 3, 4\} \Rightarrow \pi_l(\omega) = \pi_u(\omega) = \mu_A$). In worlds $5, 6$ and 7, however, they both have the prior μ_B, also specified in the figure (in symbols, $\omega \in \{5, 6, 7\} \Rightarrow \pi_l(\omega) = \pi_u(\omega) = \mu_B$). At every world, the agents' priors are equal to each other's (thus the frame satisfies Global Shared Prior), but at different worlds they have different priors from the prior they would have had in other worlds (the frame fails to satisfy Constant Prior). Leo's possibility



μ_A :
$\frac{1}{10}$ $\frac{1}{10}$ $\frac{1}{5}$ $\frac{1}{20}$ $\frac{3}{20}$ $\frac{1}{5}$ $\frac{1}{5}$
μ_A
Worlds :
1 2 3 4 5 6 7
μ_B
μ_B :
$\frac{1}{12}$ $\frac{1}{12}$ $\frac{1}{12}$ $\frac{1}{4}$ $\frac{1}{12}$ $\frac{1}{12}$ $\frac{1}{3}$
Leo :
1 2 3 4
5 6 7
Ulrich :
1 2
3 4 5 6 7

Fig. 2. Agents commonly know that they share a prior, but still agree to disagree.

Sadzik, 2008; Barelli, 2009.) According to 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. But all these results tell us is that if the agents commonly know that they agree on every possible random variable (or, in Heifetz's set-up, on a particular sequence of bets), they will commonly know what their prior is. The results do not tell us whether agents who merely commonly know that they are rational must also commonly know what their prior is. In particular, they allow that agents could commonly know that there is a shared prior which renders them immune from arbitrage, but not know what exactly that prior is.

Since the models in the main text do not represent times, they are more closely related to this second interpretation of the prior. But it is worth noting that the models can also be extended to directly represent the first, "literal" subjective Bayesian interpretation of priors. Since this subjective Bayesianism takes conditionalization as form of genuine *update*, we add a set of times to the models, T, which for simplicity we may take to be countable and indexed by the natural numbers. We then define the possibility correspondences for each agent as $P_i : \Omega \times T \rightarrow \mathscr{P}(\Omega \times T)$. 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 t_1. A prior assignment function $\pi_i : \Omega \rightarrow \Delta(\Omega, t_1)$ 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, pp. 1455–1458; Heifetz, 1996, and, from a different perspective, Hanson 2006).

correspondence is described by: $\omega \in \{1, 2, 3, 4\} \Rightarrow P_l(\omega) = \{1, 2, 3, 4\}$; $\omega \in \{5, 6, 7\} \Rightarrow P_l(\omega) = \{5, 6, 7\}$, while Ulrich's is described by: $\omega \in \{1, 2\} \Rightarrow P_u(\omega) = \{1, 2\}$; $\omega \in \{3, 4, 5, 6, 7\} \Rightarrow P_u(\omega) = \{3, 4, 5, 6, 7\}$. Thus the frame satisfies Global Truth, Global Positive Introspection and Global Negative Introspection (it is an $S5$ or partitional frame). We stipulate finally that the frame satisfies Global Bayesianism: at every world, the agents update their prior at that world by conditionalization.

This (unpointed) agreement frame yields a pointed Aumann frame no matter which world we select as the actual one. Even so, at every state Leo and Ulrich agree to disagree about the proposition $A = \{1, 3, 7\}$. (In Figure 2, the elements of this proposition are circled.) Given his partition, Leo's posterior probability in this proposition (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 fact that the agents commonly know what probabilities they each assign to A, that the agents 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. This answers our earlier formal question in the affirmative: even if agents commonly know that they share a prior and are rational according to Bayesian Uniqueness, if the agents do not commonly know *which* prior they have, the key conditional may fail. Thus, even the amended version of the key conditional, which explicitly requires common knowledge of rationality, still does not hold in general on the class of pointed Aumann frames.

In concluding, it may be worth observing that if $a = 1$ or $a = 2$, then in the resulting pointed frame $\mathcal{F}_a$, each agent knows what his own prior is, and knows that he knows what it is, and so on. In fact, Ulrich knows that Leo knows Leo's prior, and that Leo knows that Leo knows Leo's prior, and so on. But although the agents mutually know what the prior is at these states, Leo does not know whether Ulrich knows Ulrich's prior, so they don't mutually know that they mutually know what the prior is. In $S5$ frames which satisfy Global Bayesianism and Global Shared Prior, but which also provide counterexamples to the key conditional, there will always be some n such that the agents fail to mutually know^n what the prior is. (Otherwise, we would effectively be in an Aumann frame, where the theorem holds.) But there is no finite bound on the n at which this fails: given any choice of $n \in \mathbb{N}$, it is easy to extend the example here to give an example in which agents mutually know^n what the prior is, but nevertheless agree to disagree.

§4. Common knowledge of update by conditionalization.

In addition to common knowledge of the common prior, Aumann originally assumed informally that the agents in the model commonly know that they update by conditionalization.[24] The proponent of Bayesian Uniqueness holds that rational agents' beliefs accord with what is obtained 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. Subjective Bayesians may similarly believe that even in cases where a group of agents do commonly know that they had the same probabilistic beliefs about a given subject matter at some earlier time, they do not commonly know that they have continued to update these beliefs appropriately.[25]

[24] In later work, Aumann himself is explicit about this assumption (e.g. Aumann, 1998; p. 929, fn. 2).

[25] See above n. 23 for discussion of alternative subjective Bayesian views of priors.

Our new models allow us to show that 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 their shared prior.

Consider the following frame, which satisfies both Global Shared Prior and Constant Prior. There are six worlds ($\Omega = \{1, 2, 3, 4, 5, 6\}$) and two agents, Arnheim and Diotima ($N = \{a, d\}$); the agents commonly know which prior they share, namely, the uniform prior over the six worlds in the model (which we call μ, for the remainder of the example). (In other words, $(\forall\omega \in \Omega)(\pi_a(\omega) = \pi_d(\omega) = \mu)$, where $(\forall\omega \in \Omega)(\mu(\omega) = 1/6)$.) Moreover, each agent updates on a partition: Arnheim's possibility correspondence is described by: $\omega \in \{1, 2, 3\} \Rightarrow P_a(\omega) = \{1, 2, 3\}$; $\omega \in \{4, 5, 6\} \Rightarrow P_a(\omega) = \{4, 5, 6\}$, while Diotima's partition is given by $\omega \in \{1, 2\} \Rightarrow P_d(\omega) = \{1, 2\}$; $\omega \in \{3, 4, 5, 6\} \Rightarrow P_d(\omega) = \{3, 4, 5, 6\}$. Thus the frame satisfies Global Truth, Global Positive Introspection and Global Negative Introspection. Arnheim is a Global Bayesian; he updates the prior at each world by conditionalization. Diotima, however, is an actual Bayesian only at worlds 1 and 2. At the other worlds her posteriors behave irregularly; at worlds 3, 4, 5 and 6, she assigns probability 1/2 to world 6, and 1/6 to each of 3, 4 and 5.

If we choose $a \in \{1, 2\}$, the resulting frame is a pointed Aumann Frame. But even though all agents are rational by the lights of Bayesian Uniqueness at these worlds, they can agree to disagree. To see this consider the proposition $\{1, 6\}$. It is commonly known at world 1 that Arnheim assigns this proposition probability 1/3, and commonly known that Diotima assigns the same proposition probability 1/2. If Diotima were a Bayesian in worlds $\{3, 4, 5, 6\}$, then Arnheim would not even *know* (never mind be a party to common knowledge of) 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. The agents agree to disagree, even though they are both rational according to Bayesian Uniqueness at world 1.

In the previous section, I argued that even in the idealized setting where agents commonly know each other to be rational in a number of substantive ways, they can still agree to disagree because they do not commonly know what their shared prior is. In the case of update by conditionalization, the analogous line of thought is implausible, at least as applied to the theories under consideration. If agents commonly know each other to be rational, then according to Bayesian Uniqueness (or the relevant form of subjective Bayesianism), there must be common knowledge that agents update by conditionalization on their prior. So examples such as the one above should simply be taken to illustrate a second way in which common knowledge of rationality is used in the Agreement Theorem. (The first was that it assumed that agents commonly know that they have the same prior, as in example in Figure 1.) The example does not bolster the stronger claim I made in Section 3.3, that even agents who commonly know each other to be rational can still agree to disagree.

The finite examples in this section and the previous one have an artificial feel. When we are uncertain about our own or others' 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 x where $.9 < x < 1$ to the hypothesis that the big bang initiated the universe as we know it, but she may be entirely uncertain what precise value x takes in this open interval. 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.[26]

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 such as the impossibility of agreeing to disagree. (What may be surprising is that common knowledge of which prior is the prior does so much of the work.) The assumption that all agents update by conditionalization, by contrast, does not appear to be an "interpersonal" assumption of this kind. But the assumption 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 (Global Bayesianism). This stronger assumption does explicitly make reference to the knowledge of many agents. And, as the example shows, it helps to lead to a striking social consequence: the impossibility of agreeing to disagree.

§5. 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 update by conditionalization, can lead to counterexamples to the key conditional. The next two sections will show that even a theorist of Bayesian Uniqueness who is attracted to these assumptions has still other ways to save the phenomenon of disagreement among rational agents who commonly know one another's attitudes to a proposition. This section discusses how failures of common knowledge that agents satisfy Positive Introspection for knowledge (or "KK") can induce failures of the key conditional, even given common knowledge that agents satisfy Bayesian Uniqueness, and common knowledge of what the prior is.

A subject in an experiment, Clarisse, is to be presented with a rectangular object at a distance.[27] 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. According to many, Clarisse's limited ability to discriminate between such objects imposes 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).[28]

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 at a distance from Clarisse. Clarisse knows that the rectangles which might be shown have the following dimensions: $10cm \times 10cm$; $10cm \times 11cm$; $10.5cm \times 10.5cm$; $11cm \times 10.5cm$; $11.5cm \times 10cm$; and $11.5cm \times 11cm$. The 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 present the relevant rectangle, but also display a large card which states in large, clear letters that the chosen object was from the set

[26] I should note also that the infinite examples gestured at here have a very different character from Samet's (1992) example, discussed in fn. 11.

[27] Thanks to Tiankai Liu and Yan Zhang for showing me the model on which this example is based.

[28] This is modeled directly on (I_i) from Williamson (2000, p. 115).

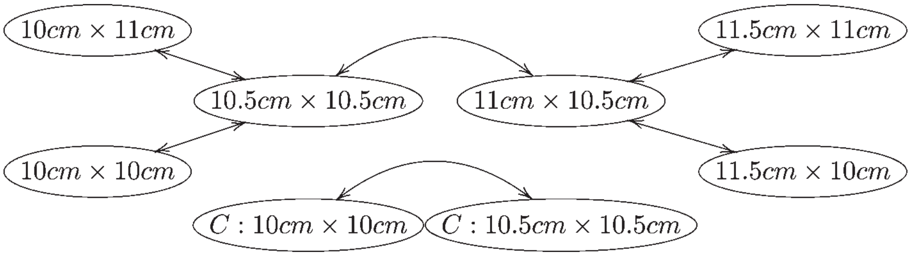


Fig. 3. Clarisse's possibility correspondence in the rectangle experiment.

$\{10cm \times 10cm, 10.5cm \times 10.5cm\}$. In this case, Clarisse knows that one of the two objects has been chosen, but does not know which. Let us assume moreover that she satisfies both Positive Introspection and Negative Introspection nonetheless. Figure 3 represents Clarisse's possibility correspondence.[29]

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; though the experimenters will publicly announce to Clarisse and Walter that some outcome has occurred, Walter will know nothing about which one it is. We may suppose furthermore that Clarisse and Walter commonly know the set-up of the situation, that they commonly know for every outcome, what Clarisse will learn in that outcome, and that they commonly know that Walter will learn nothing when the object appears.[30] Finally, we may also suppose that, as the good intellectuals they are, they have performed all relevant deductions, so that the logical assumptions implicit in the model are appropriate.

But the mere possibility that Clarisse's knowledge is inexact leads Walter and Clarisse to agree to disagree, no matter which world is actual. Consider the proposition *Extremities*, that the experimenters chose one of the extreme objects: $\{10cm \times 10cm, 10cm \times 11cm, 11.5cm \times 10cm, 11.5cm \times 11cm, C : 10cm \times 10cm\}$. Regardless of which object

[29] Note this frame satisfies not only Global Truth but also the symmetry condition which corresponds to the validity of the modal axiom **(B)**: $(\forall i \in N)(\forall \omega, \omega' \in \Omega)(\omega \in P(\omega') \Leftrightarrow \omega' \in P_i(\omega))$. The example is thus a counterexample to the validity of the key conditional on the class of pointed Aumann frames which also validate the modal logic B for knowledge.

[30] If KK fails, how can there be common knowledge at all? It is true that, for any proposition that i and j commonly know, if i's knowledge distributes over conjunctions, then i also knows that i knows that...i knows that proposition, for any finite iteration of "i knows that". But as in the example above, the agents may commonly know one proposition but nevertheless fail to be positively introspective about another proposition: 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, 2014, 2, cf. 7). 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 series of subtle variations involving knowledge itself. If, as a result, one denies the possibility of infinitely iterated knowledge of any "non-trivial" propositions, then one will be accordingly committed to the impossibility of common knowledge of non-trivial propositions. But this is a particular feature of a particular view about KK failures; there are a variety of consistent positions according to which KK failures are common (for example, because of mundane reasons concerning failure to consider the question of whether one knows), but it is nevertheless possible to have common knowledge (for example, in the cases where one does in fact consider questions about higher-order interpersonal knowledge).

is presented in which way, Clarisse's posterior probability in *Extremities* is 1/2. Similarly, regardless of which object is presented, Walter's posterior probability in *Extremities* is 5/8. Thus no matter what happens, the agents commonly know that Clarisse assigns *Extremities* 1/2 and that Walter assigns the same proposition 5/8.[31]

The example of Walter and Clarisse extends Williamson's example of Mr. Magoo (2000, pp. 116–123) to two dimensions. The frame satisfies Global Uniqueness, Global Bayesianism and Global Truth, but it fails Global Positive Introspection, since Clarisse (though not Walter) fails Positive Introspection at some worlds.

Williamson's Margins for Error premise is controversial, as are the failures of Positive Introspection which it generates. But in fact the example above does not require *actual* failures of Positive Introspection. Suppose that the actual outcome of the above experiment is $C\ :\ 10cm \times 11cm$. If we denote this state a, a satisfies Truth, Positive Introspection and Negative Introspection; the resulting frame is a pointed Aumann frame. Once we see this, it is clear that we could give an alternative description of the model, in which there were no "genuine" possibilities in which Clarisse failed Positive Introspection. Instead of taking the other six possibilities to be possible outcomes of the roll of a die, they might represent figments of one of the agents' imagination. Clarisse, for her part, knows (let us suppose she is right) that in this instance she satisfies both Positive and Negative Introspection. But poor Walter might simply be unable to make up his mind as to whether knowledge satisfies Positive Introspection. For example, he may believe that the question is an empirical one for psychology to decide. As is appropriate in such matters, he remains uncertain whether humans do or do not satisfy Positive Introspection. Walter may thus be rational in remaining uncertain: the uniquely rational prior (or the rational prior to adopt in the above circumstance) may assign some positive probability to the proposition *that human knowledge fails to be positively introspective*. In fact, this prior might even assign positive probability to the claim that *rational* agents fail to be positively introspective; merely *epistemically* possible failures of positive introspection might thus be compatible with common knowledge of rationality. But however we spell out the case, the point remains: so long as Walter's evidence does not rule out the possibility that Clarisse fails Positive Introspection, Clarisse and Walter can agree to disagree.

Thus the theorist of Bayesian Uniqueness (and also subjective Bayesians who believe that Moderate Bayesian Uniqueness applies to the scenario above) can avoid the argument from Agreement by rejecting Global Positive Introspection for knowledge. In adopting this response, they need not allow for failures of Positive Introspection in the actual state; they need only allow that the uniquely rational prior assigns positive probability to the proposition that Positive Introspection fails for the agents under consideration.

These observations bring us back to the theme introduced at the close of the last section. Positive Introspection appears to be a primarily "single-agent" principle. But if we assume Global Positive Introspection, then the agents in the model commonly know that each of them satisfies Positive Introspection. As we have seen, even if the agents actually satisfy Positive Introspection, if they merely fail to commonly know that they do, they may agree to disagree. Once again, the implicit assumption that agents commonly know that they

[31] 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. These "card" worlds are only needed to provided worlds where Clarisse *actually* satisfies the pointed axioms, as I discuss below in the main text.

satisfy Positive Introspection plays a role in generating the surprising social consequence that agents cannot agree to disagree.

§6. False beliefs. In this final section, I will consider how the proponent of Bayesian Uniqueness might respond to the argument from Agreement by denying that rational agents commonly know that they update only on true propositions. 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 μ is their shared prior, commonly believe that they update by conditionalizing the prior on what they believe, and commonly believe that they satisfy *both* Positive and Negative Introspection for belief.

The frame in Figure 4, with two worlds ($\Omega = \{\alpha, \beta\}$) and two agents, Yes and No ($N = \{y, n\}$), satisfies both Global Positive and Global Negative Introspection. Let μ be the distribution such that $\mu(\alpha) = \mu(\beta) = 1/2$, and let μ be the constant, common prior of both agents in every world ($(\forall i \in N)(\forall \omega \in \Omega)(\pi_i(\omega) = \mu)$). Whichever prior is chosen, we assume that agents conditionalize on it at every world; the resulting frame satisfies Global Shared Prior, Constant Prior and Global Bayesianism. In this frame, Yes considers only β possible, regardless of which world is actual ($(\forall \omega \in \Omega)(P_y(\omega) = \{\beta\})$), while No considers only α possible regardless of which world is actual ($(\forall \omega \in \Omega)(P_n(\omega) = \{\alpha\})$). Thus μ (and indeed any prior which gives positive probability to each world) will yield a situation in which these agents agree to strongly disagree, in the sense that they disagree not just in their posterior probabilities, but in fact each believes the negation of a proposition the other believes. Yes and No commonly believe that Yes believes $\{\beta\}$, and No believes $\{\alpha\} = \Omega \setminus \{\beta\}$, or, in other words, the negation of $\{\beta\}$.[32]

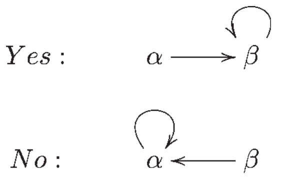


Fig. 4. False beliefs allow for common knowledge disagreements.

The frame in Figure 4 cannot provide the basis of a pointed Aumann frame; every state fails to satisfy Truth. But disagreements among agents who commonly know each others' attitudes do not require that any agents have false beliefs at the actual world: we can give such examples also in the class of pointed Aumann frames.

The frame in Figure 5, with two agents, Yes and Maybe ($N = \{y, m\}$), and two states ($\Omega = \{\alpha, \beta\}$), again satisfies both Global Positive and Global Negative Introspection.

[32] John Collins (1997) argues that van Fraassen's (1984) Reflection Principle can be used to rule out examples of common knowledge disagreement such as the one above. The Reflection Principle says, roughly: if one is certain that one will have a specific credence in a given proposition at a later time, one should now have that credence in that proposition. 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, pp. 64–69). Briggs' "qualified reflection" is restricted to cases in which one is certain one's evidence will be veridical (see (ii) on p. 69). An earlier, similar view can be found in Elga (2007). In the "prior" situation in the example above, both agents assign positive probability to their updating on a false proposition, so Reflection need not apply.

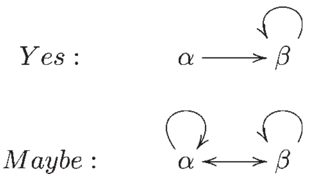


Fig. 5. Common knowledge disagreement at β with only true beliefs.

Once again, regardless of whether the actual world is α or β, Yes considers β possible and nothing else ($(\forall\omega \in \Omega)(P_y(\omega) = \{\beta\})$), while Maybe considers the elements of $\{\alpha, \beta\}$ to be possible ($(\forall\omega \in \Omega)(P_m(\omega) = \{\alpha, \beta\})$). This time, if β is specified as the actual world, and posterior beliefs are consistent with a prior which assigns nonzero probability to each world, the resulting frame *is* a pointed Aumann frame. And indeed, for any prior that gives nonzero probability to each state, it will be a matter of common belief that Yes assigns posterior probability 1 to β, and posterior probability 0 to α. It will also be common belief that Maybe assigns positive posterior probability to both α and β. The agents thus agree to disagree about their posterior probabilities in the singleton proposition $\{\alpha\}$. Moreover, 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.

This second example displays perhaps the most intuitive case for how rational agreeing to disagree is possible. We very often consider it possible that others are mistaken. Even in this simple, stylized example, the fact that one agent allows for the possibility that the other is mistaken 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.[33] This disagreement occurs, even though both agents (commonly know that they) satisfy both Positive and Negative Introspection. The model fails Global Truth, but not necessarily because anyone has false beliefs. It fails Global Truth because the agents do not *commonly know* that they each have only true beliefs. We can even think of world α as a figment of Maybe's undecided imagination; it is only Maybe's misguided view of the situation which allows for the possibility that Yes has a false belief.

These examples show that subjective Bayesians should hold that the Agreement Theorem is at best very rarely applicable, even among agents who commonly know that they are rational, and commonly know which prior they share. For subjective Bayesians, an agent's "false beliefs" are interpreted as his or her 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 are apt to claim that we rarely if ever are certain of any propositions. They 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

[33] 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.

false beliefs (certainty in a false proposition) 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. Agents may commonly know that they are (Savage) rational, but fail to commonly know that they have updated only on true propositions, and thus they may rationally agree to disagree. For subjective Bayesians of this kind, as with those who denied Positive Introspection in the previous section, even common knowledge of rationality is insufficient to restore the Agreement Theorem, since common knowledge of rationality does not entail Truth, never mind Global Truth.

The situation for the theorist of Bayesian Uniqueness is slightly subtler. Since according to this view, agents whose doxastic attitudes are determined by conditionalizing the prior on propositions stronger or weaker than their total evidence fail to be fully rational, her assessment of the agents in the above examples depends on her theory of evidence. 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 at face value, as examples of fully rational agents. But if one believes that one's evidence must consist of only true propositions, then conditionalizing 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 conception of evidence may hold that such failures of full rationality are pervasive. On this version of the view, the examples which rely on false beliefs would be examples of agents who fail to be fully rational, but may satisfy a different standard, such as reasonableness.[34] 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, and commonly known to be so.

The examples of this section also continue the elaboration of a theme we have traced now for some time. Truth 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, then it cannot be that one agent believes the negation of a proposition that another believes. (Thus the starkest examples, such as the one in Figure 4, are not possible.) But when we impose Truth in interactive frames, we typically impose the Global version of that assumption, thus making the further assumption that the agents in the frame commonly believe that they each have only true beliefs. And this stronger assumption plays a role in guaranteeing the surprising social consequence that agents cannot agree to disagree.

§7. Conclusion. Sections 3–6 have focused on four assumptions which are implicit in Aumann's original models: (1-Global Uniqueness) that there is some prior μ such that agents commonly know that μ is their shared prior; (2-Global Bayesianism) that agents commonly know that they update by conditionalization; (3-Global Positive Introspection) that agents commonly know that they satisfy positive introspection; (4-Global Truth) that agents commonly know that they update on true propositions. Our new models allowed us to make each of these assumptions explicit, and to consider relaxations of each of them. These relaxations led to countermodels to the key conditional. As noted in the Introduction

[34] For a similar move in a different context, see Aarnio (2010).

(Section 1), some have argued that observed violations of the Agreement Theorem imply that even rational people can have heterogeneous priors. The 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. Thus, even the proponent of Bayesian Uniqueness may accept the rationality of agreeing to disagree.

In the first instance, the examples show that Aumann's theorem depends on an assumption of common knowledge of rationality. But in fact even this assumption is not quite enough: in three cases, we saw that given certain standard conceptions of rationality, common knowledge of rationality is insufficient to restore the theorem. First, a theorist of Bayesian Uniqueness who believes that agents who commonly know that they are rational may fail to commonly know what prior is the uniquely rational one can allow that agents who commonly know that they are rational may still agree to disagree. Similarly, if agents who commonly know that they are rational may fail to commonly know that they satisfy Positive Introspection, then even agents who commonly know what the common prior is and commonly know that they are rational by the lights of Bayesian Uniqueness, may still agree to disagree. Finally, and perhaps most strikingly, Savageite subjective Bayesians hold that agents who commonly know that they are rational do not therefore commonly know that they have updated only on true propositions. According to this prominent version of subjective Bayesianism, agents who commonly know that they are rational and commonly know of some μ, that μ is their common prior, may still agree to disagree.

Throughout the paper, I have used pointed frames to distinguish common knowledge assumptions from conceptually related, but formally weaker assumptions. While the target of this paper has been a clearer understanding of the role of these assumptions in the Agreement Theorem, pointed frames are a tool which can be used to eliminate common knowledge assumptions from epistemic models in other contexts as well. We often wish to impose conditions on a class of models without assuming that the agents in the model know or commonly know the conditions we are imposing. But 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. Pointed frames give us a natural setting in which we can impose formal constraints on a single world, and thus make assumptions about our agents without also assuming that our agents commonly know that these assumptions hold. The usual, unwarranted common knowledge assumptions are not idle idealizations: they can cause the models to have surprising and even counterintuitive validities. But unlike the surprising deliverances of deep mathematics, these new validities simply reveal that the models contain assumptions which we did not intend to make. They are symptoms of errors in our attempts to represent agents' knowledge and beliefs, not profound discoveries about the structure of our social lives. The Agreement Theorem itself is just one symptom of a series of such errors.

§8. Appendix A. Qualitative models.

8.1. Representing alternative theories of evidence. This Appendix discusses a more general class of epistemological views than the ones addressed in the main text. I use *decision functions* to represent epistemological theories which subscribe to the Uniqueness Thesis, but which are not necessarily Bayesian, and in fact need not endorse the claim that that an agent's evidence is a set of propositions.

Let a qualitative agreement frame be a structure $\langle \Omega, N, (P_i)_{i \in N}, (d_i)_{i \in N}, A \rangle$, where A is a set of (possibly infinitely many) "actions" and $d_i : \mathscr{P}(\Omega) \to A$ is a *decision function*, which takes propositions to abstract "actions".[35] On the standard interpretation of these functions, if $P_i(\omega) = E$ 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"; they simply represent some feature of a state which is fully determined by the conjunction of all of the propositions to which a given agent i stands in the relation represented by the possibility correspondence P_i (for example, *believes* or *knows*).

We will be interested in a special class of decision functions. Let Δ be a set of doxastic attitudes: an attitude ϕ belongs to Δ if and only if there is some body of evidence E and some proposition P such that it's rational for an agent with evidence E to bear ϕ to P. We then define our $d : \mathscr{P}(\Omega) \to \Delta^{\mathscr{P}(\Omega)}$ as a function from propositions to functions from propositions to attitudes (thus we take A itself to be $\Delta^{\mathscr{P}(\Omega)}$). On the intended interpretation, the argument of this function is understood to be a proposition which stands proxy for the agent's evidence. (More on "standing proxy for" in a moment.) The value of the function at a body of evidence is itself a function, which represents the full doxastic state the agent has or should have: in particular, this new function delivers, for any proposition, the attitude the agent does or should take to that proposition, given his or her evidence. 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 there is some d which maps "proxies" for bodies of evidence to functions which describe, for any proposition, the attitude one should have to that proposition.

What does it mean for a proposition to "stand proxy for" an agent's evidence? The Uniqueness Thesis itself does not commit on the character of evidence, so if we wish to model that thesis in full generality using the formalism of decision functions we must find some way to represent agents' evidence by propositions, even if the evidence itself is not thought of as propositional. But this in itself is not a difficult task: for any agent i, and any body of evidence E^*, propositional or otherwise, we can always map E^* to the proposition *that i has evidence E^**. This brings us close to the desired result, but not quite the whole way. Decision functions produce verdicts on *all* propositions, but since there may be more propositions than there are total bodies of evidence, the map I've just described from (non-propositional) evidence may fail to be surjective, and thus give us only partially defined "decision functions". But this problem can be avoided in a variety of ways: a simple workaround is to introduce an arbitrary value a^*, and stipulate that the decision function takes this value at those propositions on which it was previously undefined.

8.2. Agreement theorems. As it stands, then, the formalism is extremely general, and can represent a vast range of views of evidence. But without any further constraints on decision functions, we can say very little of substance about them. So let's begin by imposing some simple structural constraints on our models, for example, that a qualitative agreement frame $\mathcal{F}$ satisfies

$$\textbf{Global Uniqueness} \quad \text{iff} \quad (\forall i, j \in N)(d_i = d_j).$$

The assumption that all agents have the same decision function is just a component of the Uniqueness Thesis itself. But of course the assumption that this decision function is the

[35] Decision functions were introduced to the literature on the Agreement Theorem by Cave (1983) and Bacharach (1985).

same at all worlds (one we haven't made explicit here)[36] yields the further assumption that all agents commonly know what that decision function is.

Now we can define our class of qualitative models similarly to the class of models defined with probabilities:

DEFINITION 8.1. *A qualitative agreement frame $\mathcal{F}$ is a qualitative* 4*-Aumann frame if and only if it satisfies*

(i) *Global Truth, Global Positive Introspection;*
(ii) *Global Uniqueness.*

A qualitative 4*-Aumann frame is a qualitative Aumann frame if and only if it satisfies Global Negative Introspection.*

At this point, however, we still do not have any constraints on the globally unique decision function. Let a decision function d satisfy the *formal sure-thing principle* if and only if

$$\big((\exists a \in A)(d(E) = d(F) = a) \text{ and } E \cap F = \emptyset\big) \Rightarrow d(E \cup F) = a.$$

This principle can be glossed as saying that, if E and F are incompatible, and if, if one knows E one should take action a, and if, if one knows F, one should take action a, then if one knows only that (E or F), one should also take action a. I will discuss this principle in more detail in 8.3.[37]

[36] Here we do not have an analog of "Constant Prior", and hence Global Uniqueness is not decomposed into a "Shared" condition and a "Constant" condition. We *could* have relativized decision functions to worlds, simply by adding another world argument in the "global" decision function d_i; this would have allowed us to make similar distinctions among these assumptions about qualitative frames as we did in the Bayesian case. But since I won't be concerned to offer countermodels of this form in what follows, I've omitted the extra formalism.

[37] It is perhaps worth noting that a 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; Aumann, Hart and Perry 2005; Samet 2010). The problem runs as follows. In frames which satisfy Global Truth and Global Negative Introspection, Qualitative Agreement theorems require an assumption that the shared decision function is defined on knowledge-sets the agents could not have. Consider a frame which satisfies Global Negative Introspection and Global Truth: each agent i's possibility correspondence partitions the state space (see n. 14). So for an agent i, for some E, the proposition that E is identical to the proposition that i knows E. Similarly, if i's correspondence is not the trivial one which equals the universe at every state, 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 formal 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. The difficulty is supposed to be that we should restrict a given agent's decision function so that it is defined only on bodies of knowledge that agent could have. But in our interpretation of the formalism, this problem does not apply, since the Uniqueness Thesis is an interpersonal principle. The decision function, which we might call an "evidential response function" need not be sensitive to whether i (as opposed to j or k) could know the proposition which it takes as its argument. This decision function merely gives a verdict about the doxastic attitude *someone* who had a given body of evidence should take,

For now, the interest of the principle lies in the theorem it will allow us to prove. By analogy to the abbreviation I used for posterior probabilities, I will use square brackets to represent propositions about agents' decisions, writing $[d_i = k_i] = \{\omega : d_i(P_i(\omega)) = k_i\}$. Using this notation we can state the result due to Cave (1983) and Bacharach (1985):

THEOREM 8.2 (Qualitative Agreement Theorem). *Let $\mathcal{F}$ be a qualitative Aumann frame in which for every i, d_i satisfies the formal sure-thing principle. If there is a state ω such that for each i there is some $k_i \in A$ for which the agents commonly know $[d_i = k_i]$ at ω, then $(\forall i, j \in N)(k_i = k_j)$.*

(A generalization of this theorem is proven in Appendix B (Section 9).)

In the final subsection of this Appendix, I'll consider the sure-thing principle in some detail. But for now, we should take stock of how general this new result is. As we saw above, if one has a non-propositional view of evidence, the formalism of decision functions may be applicable only via some "unnatural" or "gerrymandered" mapping from bodies of evidence to propositions. But in this case principles stated in terms of the propositions which stand proxy for bodies of evidence will lose some of their intuitive appeal. The sure-thing principle is one such principle: the conditions on Boolean relations between propositions which appear in the antecedent of the sure-thing principle may not correspond to any interesting relations between the bodies of evidence these propositions represent. So without further elaboration on the "standing proxy for" relation, we can't say whether the theorem applies.

While I think this observation raises important questions, the target of this appendix is ultimately a different problem with versions of the sure-thing principle. Accordingly, from here on I will assume that one's evidence just *is* a set of propositions.[38] Moreover, I'll assume – as we have been assuming throughout – that the set of propositions which constitutes one's total evidence satisfies the conditions of logical closure mentioned near the close of Section 2: that is, the set of propositions forms a filter on the powerset algebra of the universe Ω (the algebra of propositions) – it is closed under superset (if a proposition belongs to one's evidence, anything entailed by it does) and is closed under finite conjunction (if a finite set of propositions belongs to one's evidence, their conjunction does). This means that I will again be able to take the possibility correspondences as representing directly what propositions do and don't belong to one's evidence.

Throughout the paper I've "unofficially" adopted the view of Williamson (2000 , chap. 9), that one's evidence is what one knows. This unofficial view has helped to give us some intuitive purchase on the various results discussed in the paper, and I will continue to use it in what follows. But on this interpretation of the possibility correspondences, the assumption that agents satisfy Global Negative Introspection is implausible; philosophers of all stripes reject the claim that knowledge satisfies Negative Introspection (Hintikka, 1962; Williamson, 2000, pp. 23–27; Stalnaker, 2006; Stalnaker, 2009, p. 400).

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.[39]

irrespective of which agents can or cannot have that body of evidence. Thus on this interpretation, Moses and Nachum's problem does not arise.

[38] A version of this view is explicitly endorsed by Williamson (2000, chap. 9) and Conee and Feldman (2004, chap. 9).

[39] 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 proposition E if and only if there is a proposition $F \subseteq E$ such that, for all $\omega, \omega' \in F$,

Accordingly, say that a decision function d satisfies the *strong sure-thing principle* if and only if

$$\Big((\exists a \in A)(d(E) = d(F) = a) \text{ and } \big(E \cap F = \emptyset \text{ or } d(E \cap F) = a\big)\Big) \Rightarrow d(E \cup F) = a.$$

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.[40] But we can prove an Agreement Theorem for decision functions that satisfy the strong sure-thing principle, even if the agents fail Global Negative Introspection (in the terminology of n. 14, the logic of knowledge is merely $S4$).

THEOREM 8.3. *Let $\mathcal{F}$ be a qualitative 4-Aumann frame, so that for all i, d_i satisfies the strong sure-thing principle. If there is a state ω such that for each i there is some $k_i \in A$ so that the agents commonly know $[d_i = k_i]$ at ω, then $(\forall i, j \in N)(k_i = k_j)$.*

Update by conditionalization on a globally unique prior satisfies the strong sure-thing principle. (This is proven in Appendix B (Section 9), as Fact 9.2.) So Theorem 8.3 has Theorem 2.3 as a corollary.

8.3. The sure-thing principle: A problem for the argument from qualitative Agreement. In the Introduction (Section 1), I stated an argument from Agreement directed against Bayesian Uniqueness. The Qualitative Agreement Theorem and Theorem 8.3 suggest a more general argument, directed at the Uniqueness Thesis itself, and not merely its Bayesian elaboration. The argument runs as follows. If the decision function which represents the rational response to evidence satisfies the strong sure-thing principle, then by Theorem 8.3, the Uniqueness Thesis is inconsistent with rational agreeing to disagree among agents who satisfy Positive Introspection and Truth. But, as our earlier examples of scientists and jurors showed, such disagreements seem to be possible. So the Uniqueness Thesis is false.

The reader of this paper will be aware that this argument has a number of important gaps. We could introduce a new class of models formally, the pointed qualitative agreement frames, and also allow the decision functions to vary at different worlds and to differ across different agents. We could then impose constraints on decision functions and epistemic or doxastic axioms only at the designated point in the pointed frames. If we did this, then the four different responses to the argument from Agreement which I have developed throughout this paper could be used again in the present context. That is, we could consider denying any of the following four assumptions:

(D1) that agents commonly know which decision function represents the uniquely rational response to evidence;

(D2) that agents commonly know that they each use that uniquely rational decision function;

$P_i(\omega) \cap P_i(\omega') = \emptyset$ 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 propositions. We can prove that all locally decomposable possibility correspondences cannot agree to disagree if the globally unique decision function satisfies the formal sure-thing principle.

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

(D3) that agents commonly know that if any of them knows a proposition, she knows that she knows it;

(D4) that agents commonly know that they use only true propositions as input to the uniquely rational decision function.

Since update by conditionalization on a common prior satisfies the (strong) sure-thing principle, the counterexamples to the key conditional in the main text are also counterexamples to the new, qualitative version of the key conditional.

But in closing I want to note one further problem with the argument from Qualitative agreement, which does not immediately affect the Bayesian version discussed in the main text. The problem is that the new argument relies on a new premise: that the decision function which represents the rational response to evidence satisfies the (strong) sure-thing principle. In the case of Bayesian Uniqueness, since update by conditionalization on a globally unique prior satisfies the strong sure-thing principle, we did not need to consider this premise; it followed from other commitments of the view under consideration. But in the qualitative case, we must ask whether the Uniqueness theorist should accept the principle.

The sure-thing principle is usually motivated by stories such as the following, taken from Savage:[41]

> 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."[42]

Savage's principle, in spite of its putative "ready acceptance", is subject to counterexamples. Here is a striking one.[43] 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

[41] Bacharach (1985, p. 181), Aumann *et al.* (2005, p. 8), and Samet (2010, p. 171).

[42] Savage (1954, p. 21).

[43] Thanks to Tim Williamson and an anonymous referee for suggesting versions of this example. The example bears some structural similarities to Parfit's miners example (1983, recently revived for a different purpose by Kolodny & 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 & Perry (2005, p. 10).

victory, and since his bet would win, he would win considerably more than the bookie's fee. So he would place a bet. The businessman then considers what 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 show that Savage's argument for his principle is insufficient. If the argument from Qualitative Agreement were to be effective, the opponent of the Uniqueness Thesis would have to offer a new argument, more powerful than Savage's, to show that the decision function which represents the rational response to evidence satisfies (for example) the strong sure-thing principle. And of course, even if he succeeds in that task, he would then have to somehow rule out the counterexamples to the key conditional developed in the main text.

§9. Appendix B. Proofs.

Recall the claim to be proven.

THEOREM 9.1. *Let $\mathcal{F}$ be a qualitative* 4*-Aumann frame, so that for all i, d_i satisfies the strong sure-thing principle. If there is a state ω such that for each i there is some $k_i \in A$ so that the agents commonly know $[d_i = k_i]$ at ω, then $(\forall i, j \in N)(k_i = k_j)$.*

Proof. Say that a decision function d is constant on a proposition E if and only if for all $(\omega, \omega' \in E)(d(P_i(\omega)) = d(P_i(\omega')))$. We first show that in a frame which satisfies Global Truth and Global Positive Introspection, for any self-evident proposition E, if a decision function satisfies the strong sure-thing principle, and is constant on E with some arbitrary value a, then $d(E) = a$. (This is the "lemma" described in the first paragraph following Theorem 2.3.)

Let $\mathcal{P}_E = \{P_i(\omega) | \omega \in E\}$. The proof is by induction on the cardinality of $\mathcal{P}_E$. If $|\mathcal{P}_E| = 1$, the result follows trivially: by Global Truth, $P_i(\omega) = E$ 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 Global Positive Introspection and Global Truth, 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, $d(P_i(\alpha) \cup P_i(\beta)) = d(P_i(\alpha))$. 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 proposition E is self-evident for the new correspondence P_i^* (which still satisfies Global Truth and Global Positive Introspection), 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, for all $\omega \in E$, $d(E) = d(P_i^*(\omega))$. Since by the definition of P_i^*, for all such ω, $d(P_i^*(\omega)) = d(P_i(\omega))$, it follows that for all $\omega \in E$, $d(E) = d(P_i(\omega))$, as required. **(B2)** If $P_i(\alpha) \cap P_i(\beta) \neq \emptyset$, then by Global Positive Introspection every $\omega \in P_i(\alpha) \cap P_i(\beta)$ has $P_i(\omega) \subseteq P_i(\alpha) \cap P_i(\beta)$. It follows that $P_i(\alpha) \cap P_i(\beta)$ is itself a self-evident set, call it F, and, by choice of α, β, that $|\mathcal{P}_F|$ is at most $n - 1$. By the induction hypothesis it follows that for all $\omega \in F$, $d(F) = d(P_i(\omega))$. So $d(P_i(\alpha))= d(P_i(\beta)) = d(P_i(\alpha) \cap P_i(\beta))$, and thus by the strong sure-thing principle $d(P_i(\alpha) \cup P_i(\beta)) = a$. We then 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 proposition E, with d constant on E at value a has $d(E) = a$. The hypothesis of the theorem is that for all $i \in N$ there is a k_i and a public E_i with $\omega \in E_i$ so that $(\forall \omega' \in E_i)\ (d(P_i(\omega')) = k_i)$. But $E = \cap_{i \in N} E_i$ is also non-empty and public, with $\omega \in E$. Since public propositions are self-evident to each $i \in N$, it follows by the lemma that $(\forall i \in N)(d(E) = k_i)$, and thus $(\forall i, j \in N)(k_i = k_j)$. □

FACT 9.2. *Let* (Ω, Σ, μ) *be a probability space where* Ω *is finite and* μ *is regular. For* $E, A, B \in \Sigma$,

(1) *if* $\mu(E|A) = \mu(E|B) = a$ *and* $A \cap B = \emptyset$*; or*

(2) *if* $\mu(E|A) = \mu(E|B) = \mu(E|A \cap B) = a$,

then $\mu(E|A \cup B) = a$.

Proof. We first show that the conclusion follows given (1). (This establishes that update by conditionalization on a globally unique prior satisfies the formal 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 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 $\mu(A) + \mu(B)$ gives:

$$\frac{\mu(E \cap A) + \mu(E \cap B)}{\mu(A) + \mu(B)} = a. \tag{1}$$

But by the disjointness of A and B, $\mu(A \cup B) = \mu(A) + \mu(B)$, and, moreover, $\mu(X \cap (A \cup B)) = \mu(A \cap X) + \mu(B \cap X)$. So 1 is equivalent to $\mu(E|A \cup B) = a$, as required.

We now show (2). We first establish as a lemma that if $\mu(E|A) = \mu(E|C) = a$, where $A, C, E \in \Sigma$, and $C \subsetneq A$, then $\mu(E|A \setminus C) = a$. (Since μ is regular and $C \subsetneq A$, $\mu(A) - \mu(C) > 0$.) 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:

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

Now we can use the fact that $\mu(E \cap A) = \mu(E \cap C)\mu(A)/\mu(C)$ to give:

$$\frac{\mu(E \cap A) - \mu(E \cap C)}{\mu(A) - \mu(C)} = \frac{\mu(E \cap C)\mu(A)/\mu(C) - \mu(E \cap C)}{\mu(A) - \mu(C)} \tag{2}$$

We multiply $\mu(E \cap C)$, by $\mu(C)/\mu(C)$, to make this clearer. Then 2 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 required for the lemma.

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 fact that conditionalization satisfies the formal sure-thing principle (established as (1) above) gives us that $\mu(E|A \cup B) = a$, as required. □

Acknowledgments. Thanks to Frank Arntzenius, Peter Fritz, Alex Paseau, Richard Pettigrew, Daniel Rothschild, Tim Williamson, and two anonymous referees for detailed and insightful comments on drafts of this paper. Thanks also to Robin Hanson for correspondence about the models in Section 3, and Dan Waxman for comments on the final version. And thanks finally to Peter Fritz, Jeremy Goodman and Tiankai Liu for many happy conversations both about this material and not.

BIBLIOGRAPHY

Aarnio, M. L. (2010). Unreasonable knowledge. *Philosophical Perspectives*, **24**(1), 1–21.

Aaronson, S. (2005). The complexity of agreement. *Symposium on the Theory of Computing*, Extended Abstract. Full Version at www.scottaaronson.com/papers/agree–econ.pdf.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics*, **4**(6), 1236–1239.

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

Aumann, R. J. (1998). Common priors: A reply to Gul. *Econometrica*, **66**(4), 929–938.

Aumann, R. J., & Hart, S. (2006). *Agreeing on Decisions*. The Hebrew University of Jerusalem: Center for the Study of Rationality. Unpublished.

Aumann, R. J., Hart, S., & Perry, M. (2005). *Conditioning and the Sure-thing Principle*. The Hebrew University of Jerusalem: Center for the Study of Rationality.

Bacharach, M. (1985). Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, **37**(1), 167–190.

Barelli, P. (2009). Consistency of beliefs and epistemic conditions for nash and correlated equilibria. *Games and Economic Behavior*, **67**, 363–375.

Bonanno, G., & Nehring, K. (1997). *Agreeing to Disagree: A Survey.* Available from: http://www.econ.ucdavis.edu/faculty/bonanno/PDF/agree.pdf.

Bonanno, G., & Nehring, K. (1998). Assessing the truth axiom under incomplete information. *Mathematical Social Sciences*, **36**(1), 3–29.

Bonanno, G., & Nehring, K. (1999). How to make sense of the common prior assumption under incomplete information. *International Journal of Game Theory*, **28**(3), 409–434.

Briggs, R. (2009). Distorted reflection. *The Philosophical Review*, **118**(1), 59–85.

Cave, J. A. K. (1983). Learning to agree. *Economics Letters*, **12**(2), 147–152.

Collins, J. (1997). *How we can Agree to Disagree.* Available from: http://collins.philo.columbia.edu/disagree.pdf.

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. New York: Oxford University Press.

Cowen, T., & Hanson, R. (2014). Are disagreements honest? *Journal of Economic Methodology*. Forthcoming.

Easwaran, K. (2014). Regularity and infinitesimals. *Philosophical Review*, **123**(1), 1–41.

Elga, A. (2007). Reflection and disagreement. *Noûs*, **41**(3), 478–502.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. Y. (1995). *Reasoning About Knowledge*. Cambridge, MA: MIT Press.

Feinberg, Y. (2000). Characterizing common priors in the form of posteriors. *Journal of Economic Theory*, **91**(2), 127–179.

Feldman, R. (2007). Reasonable religious disagreements. In Antony, L., editor. *Philosophers without Gods: Meditations on Atheism and the Secular*, Chap. 17. New York: Oxford University Press, pp. 194–214.

Friedell, M. F. (1969). On the structure of shared awareness. *Behavioral Science*, **14**(1), 28–39.

Geanakoplos, J. (1989). *Game Theory without Partitions, and Applications to Speculation and Consensus.* Cowles Foundation Discussion Papers 914, Cowles Foundation for Research in Economics, Yale University, New Haven.

Geanakoplos, J. (1994). Common knowledge. In Aumann, R. J., and Hart, S., editors. *Handbook of Game Theory with Economic Applications*, Vol. 2, Chap. 40. The Netherlands: North Holland, pp. 1437–1496.

Greco, D. (2014). Could KK be OK? *Journal of Philosophy*, **111**(4), 169–197.

Hájek, A. (2010). Staying regular. Unpublished Manuscript.

Halpern, J. Y. (2002). Characterizing the common prior assumption. *Journal of Economic Theory*, **106**(2), 316–355.

Halpern, J. Y., & Pucella, R. (2011). Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial intelligence*, **175**(1), 220–235.

Hanson, R. (2006). Uncommon priors require origin disputes. *Theory and Decision*, **61**(4), 318–328.

Heifetz, A. (1996). Comment on consensus without common knowledge. *Journal of Economic Theory*, **70**(1), 273–277.

Heifetz, A. (2006). The positive foundation of the common prior assumption. *Games and Economic Behavior*, **56**(1), 105–120.

Hellman, Z. (2012). *Deludedly Agreeing to Agree*. Abstract in TARK 2013. Full paper unpublished. Available from: http://www.coll.mpg.de/sites/www.coll.mpg.de/files/workshop/DeludedPartitions21.pdf.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, NY: Cornell University Press.

Kelly, T. (2014). How to be an epistemic permissivist. In Greco, J., Steup, M., and Turri, J., editors. *Contemporary Debates in Epistemology* (second edition). Malden, MA: Blackwell Publishing Ltd.

Kolodny, N., & MacFarlane, J. (2010). Ifs and oughts. *The Journal of Philosophy*, **107**(3), 115–143.

Lederman, H. (2014). *Agreement and Equilibrium with Minimal Introspection*.

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Monderer, D., & Samet, D. (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior*, **1**(2), 170–190.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, **11**(2), 227–253.

Moses, Y., & Nachum, G. (1990). Agreeing to disagree after all. In *Proceedings of the 3rd conference on Theoretical Aspects of Reasoning and Knowledge, Pacific Grove, CA*, pp. 151–168.

Parfit, D. (1988). What we together do. Unpublished Manuscript.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, **34**(4), 517–549.

Rubinstein, A., & Wolinsky, A. (1990). On the logic of ‘agreeing to disagree’ type results. *Journal of Economic Theory*, **51**(1), 184–193.

Sadzik, T. (2008). *Impossibility of Characterizing the Common Prior Assumption*. Unpublished Manuscript. Available from: https://files.nyu.edu/ts73/public/.

Samet, D. (1990). Ignoring ignorance and agreeing to disagree. *Journal of Economic Theory*, **52**(1), 190–207.

Samet, D. (1992). Agreeing to disagree in infinite information structures. *International Journal of Game Theory*, **21**(3), 213–218.

Samet, D. (2010). Agreeing to disagree: The non-probabilistic case. *Games and Economic Behavior*, **69**(1), 169–174.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley and Sons.

Stalnaker, R. (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, **128**, 169–199.

Stalnaker, R. C. (2009). On Hawthorne and Magidor on assertion, context, and epistemic accessibility. *Mind*, **118**(470), 399–409.

van Fraassen, B. (1984). Belief and the will. *Journal of Philosophy*, **81**, 235–256.

White, R. (2005). Epistemic permissiveness. *Philosophical Perspectives*, **19**, 445–459.

White, R. (2014). Evidence cannot be permissive. In Greco, J., Steup, M., and Turri, J., editors. *Contemporary Debates in Epistemology* (second edition). Malden, MA: Blackwell Publishing Ltd, pp. 312–323.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

Williamson, T. (2007). How probable is an infinite sequence of heads? *Analysis*, **67**(3), 173–180.

HARVEY LEDERMAN
DEPARTMENT OF PHILOSOPHY
NEW YORK UNIVERSITY
5 WASHINGTON PLACE
NEW YORK, NY 10003 USA
E-mail: hsl306@nyu.edu