How We Can Agree to Disagree*

John Collins
Columbia University

Knowledge entails the truth of the proposition known; that which is merely believed may be false. If I have beliefs about your beliefs, then I may believe that some of your beliefs are false. I may believe, for example, that you mistakenly believe that it is now raining outside. This is a coherent belief for me, though not for you. You cannot coherently believe that you believe falsely that it is raining, and this despite the fact that your having that false belief is clearly a logical possibility. The proposition is, for you, a kind of doxastic blindspot.[1]

As any poker player knows, there is much more to the game than simply who holds the high hand. While contemplating whether or not to call your raise I may find myself pondering your thoughts about my thoughts about your thoughts about whose hand is higher. ("Does she know that I know that she thinks that I made that flush?") Such reasoning is so prevalent in real–life strategic situations that a whole area of study—*interactive epistemology*—has arisen to provide it formal expression. "I know that you know that I know that ..." finds its limit in the notion of *common knowledge.* A proposition X is said to be common knowledge for you and me when I know that X, you know that X, I know that you know that X, you know that I know that X, I know that you know that I know that X, and so on *ad infinitum.*[2]

*Thanks are due to Isaac Levi, Horacio Arló Costa, and the other members of the Belief Revision Group at Columbia to whom this material was first presented. I am grateful to John Geanakoplos, Ariel Rubinstein, Krister Segerberg, and Roy Sorensen for discussion and helpful comments, and much indebted to Wlodek Rabinowicz, who pointed out an error in an earlier version.

[1] The term is borrowed from Roy Sorensen [9]; the observation about the logic of belief is due to G.E. Moore.

[2] Such an account of common knowledge may be found in David Lewis [5].

The richness of the concept of common knowledge is dramatically illustrated by a famous puzzle:

> Three children wearing hats are sitting in a circle. Each of them knows that the hat she is wearing is either red or white, and each can see the color of the other two hats, but none can see the color of her own hat. In fact all three hats are red. The teacher asks each child in turn if she knows which color her own hat is, and of course each child answers no. The teacher then announces that at least one of the children is wearing a red hat. Now when the children are asked again if they can determine the color of their own hats, the answers change. The first child asked answers no. So does the second. But the third child to be asked answers that she is now certain that her hat is red!

The puzzle is to explain the chain of reasoning that led the third child to her correct answer. Since the teacher merely told the children something that each of them already knew was true, how could the announcement have made a difference?

The key to the puzzle is that there is a significant difference between (a) all three children knowing that at least one hat is red, and (b) its being common knowledge among the three children that at least one hat is red. The teacher's announcement made a difference by making what each already knew a matter of common knowledge.[3]

In a paper that deserves to be better known by philosophers, Robert Aumann proved something rather surprising about common knowledge: if two agents have the same prior probabilities, and the posterior probabilities they assign to some proposition X are common knowledge, then those posterior probabilities must be equal.[4] That is so even if those posterior probabilities for X are based on completely different evidence. This surprising conclusion

[3] That clue may be sufficient for readers not already familiar with the example; I won't spoil their fun. A particularly elegant explanation of the third child's reasoning (due to Fagin, Halpern, Moses, and Vardi) can be found in Geanakoplos [3]. Variations on this example have been used so often to illustrate the importance of the concept of common knowledge that I am almost embarrassed to reproduce it here. My only excuse is that the literature on this subject is not widely (let alone commonly!) known among philosophers.

[4] See Aumann [1].

is sometimes expressed as follows: common knowledge of posterior probabilities negates asymmetric evidence. More usually, the content of the theorem is summed up like this: rational agents with the same priors cannot agree to disagree.

This is a disturbing result. The absurdity of the conclusion suggests that common knowledge assumptions, indispensible in many game–theoretic contexts, may not be as innocuous as they look. The purpose of this paper is to offer an alternative diagnosis. I shall argue that the proper subject of interactive epistemology ought to be belief rather than knowledge. The apparent problem posed by the Agreement Theorem disappears when we drop the truth requirement and focus on common belief rather than common knowledge. This is so not because the truth requirement was essential to the *proof* of the Agreement Theorem—in fact we shall prove below a version of that result that holds for common belief. The point is rather that the truth requirement is necessary for any interesting *application* of the theorem. Once we allow that rational agents may have false beliefs, we are led to admit into our model propositions that could never doxastic possibilities for agent i, despite the fact that they might well be doxastic possibilities for all agents except i. These propositions are doxastic blindspots of the kind described in the opening paragraph. This first–person/third–person asymmetry in the logic of belief makes the assumption of equal priors impossible to maintain. The Agreement Theorem, though perfectly correct, is vacuous.

The rest of the paper is organized as follows. Section 1 contains an introduction to interactive epistemology and a formal characterization of the notion of common knowledge that leads to a proof of Aumann's Agreement Theorem. In Section 2 we develop a parallel account of common belief and prove a common belief version of the Agreement Theorem. Section 3 is devoted to a discussion of the result proved in Section 2.

1 The Agreement Theorem

Let W be a non-empty set of possible worlds. For simplicity we shall assume that W is finite. Subsets of W will be called *propositions*. Let I be some set of agents. Agent i's *knowledge structure* is a function P_i which assigns to each $w \in W$ a non-empty subset of W. $P_i(w)$ is called i's *knowledge state* at w. The elements of $P_i(w)$ are those states of the world that are compatible with everything that i knows at w. $P_i(w) \subseteq A$ is interpreted as meaning that at w agent i *knows that* A. Any two elements of $P_i(w)$ are states of the world that are subjectively indistinguishable from the point of view of agent i. At world w agent i assigns non-zero subjective probability to all and only the worlds in $P_i(w)$. Let C_i be i's prior probability function, and suppose that the posterior probability function C_{iw} that i has in world w is obtained from i's priors by conditionalizing on $P_i(w)$. Then, in particular, $C_{iw}(P_i(w)) = C_i(P_i(w)/P_i(w)) = 1$.

We impose the following conditions on $P_i(w)$: (1) $w \in P_i(w)$; (2) if $v \in P_i(w)$ then $P_i(v) = P_i(w)$. The second condition ensures that for each agent i, the $P_i(w)$ form a partition of W into mutually disjoint and exhaustive possible states of knowledge.

The upshot of these two conditions can be displayed in terms more familiar to philosophers via an equivalent representation in terms of *knowledge operators*. For each agent i, the knowledge operator K_i maps each proposition A to the set of worlds in which i knows that A. In other words, $K_iA = \{w : P_i(w) \subseteq A\}$ is simply the proposition *that i knows that A*. It is straightforward to check that the two conditions just stated for knowledge structures amount to: (1) $K_iA \subseteq A$; (2a) $K_iK_iA = K_iA$; and (2b) $\neg K_iA = K_i\neg K_iA$. Requirement (1) is the *truth condition*: only propositions that are true can be said to be known. (2a) is the requirement of *positive introspection* for knowledge: if i knows that A, then i knows that she knows that A. (2b) is the assumption of *negative introspection* for knowledge: whenever i doesn't know that A, then she knows that she doesn't know that A. Given these conditions K_i is an S5 operator.

We are now in a position to give a formal account of common knowledge. A proposition A is said to be *common knowledge* among the agents in I at world w iff for any n and any sequence $i_1, i_2, \ldots, i_n \in I$ we have $K_{i_1}K_{i_2}\ldots K_{i_n}A$. It would appear from this formulation that the process of verifying that some proposition is an item of common knowledge will involve checking an infinite number of conditions. In fact this is not so. As Aumann realized, there is a very neat equivalent characterization of common knowledge for which the verification process involves only finitely many stages. Appreciating this equivalent formulation is the first step in understanding the proof of the Agreement Theorem.

Each P_i induces a partition of W into mutually disjoint cells. We may think of these cells as the equivalence classes defined by an accessibility relation R_i that obtains between worlds v and w iff $P_i(v) = P_i(w)$. Now say that $vR'_I w$ iff $vR_i w$ for some $i \in I$. The relation R'_I so defined is obviously reflexive and symmetric. Hence its transitive closure R_I is also an equivalence relation, inducing a partition of W that is the finest common coarsening of the partitions induced by each of the R_i. As Aumann realized, this finest common coarsening of all the knowledge partitions P_i is precisely the *common knowledge partition* for the agents in the set I.

Common Knowledge Characterized: A is common knowledge among the agents in I at w iff A is entailed by $R_I(w)$, the equivalence class of R_I that contains w.

To see that this alternative formulation is indeed equivalent to the one we started with, the following way of visualizing the simple two-agent case may be of assistance. Think of the space W of worlds as a large field— a paddock—criss-crossed by fences of two colors, red and blue say. Suppose that the red fences partition the paddock into Agent Red's possible knowledge states, while the blue fences play the same role for Agent Blue. The whole field is surrounded by a red fence on top of a blue fence. Let's start at some arbitrarily chosen point w in the interior of the field. Any point that can be reached from w without crossing over a red fence represents a possible state of the world that, for all Red knows, may be the way the world actually is. Similarly, any point in the field that can be reached from w without climbing over a blue fence is an epistemic possibility at w for Blue. If I cannot get to point v from point w except by crossing exactly one red fence (in other words if I am allowed to cross first any number of blue fences, then a single red fence, then any number of additional blue fences) then v is a state of the world, which, although not an epistemic possibility for Red, is nevertheless a world that Red thinks that Blue thinks that Red thinks might be the actual world. Now suppose that I am able to move around the paddock climbing freely over fences as I come to them, except when a red fence and a blue fence coincide. I am now bounded only by the doubled fences that mark the finest common coarsening of the red and blue partitions. I can now reach any point that Red thinks that Blue thinks that Red thinks that Blue thinks that . . . thinks is epistemically possible. (Similarly with 'Red' and 'Blue' reversed). The only points I cannot reach are those ruled out by everything that is common knowledge for Red and Blue. The finest common coarsening of the two partitions is precisely the common knowledge partition.

Let's now proceed with the development. Call φ_i an *epistemic function* for i if $\varphi_i(A)$ is defined whenever $C_i(A) > 0$, and φ_i satisfies the following condition:

The Sure–Thing Principle: If $\varphi_i(A) = \varphi_i(B) = a$ and $A \cap B = \emptyset$ then $\varphi_i(A \cup B) = a$.

We may now prove the following

Theorem: Suppose that for each i, φ_i is an epistemic function and that it is a matter of common knowledge at w that this function assigns the value a_i to i's knowledge state. Then there is some proposition E such that for all i, $\varphi_i(E) = a_i$.

Proof: Let $E = R_I(w)$, the cell of the common knowledge partition for I that contains w. It is common knowledge at w that φ_i assigns the value a_i to i's knowledge state. In other words, the proposition $\{v : \varphi_i(P_i(v)) = a_i\}$ is common knowledge at w. Hence $\varphi_i(P_i(v)) = a_i$ for all $v \in E$. But E is a disjoint union of $P_i(v)$, Hence by the Sure–Thing Principle $\varphi_i(E) = a_i$. This completes the proof.

Fix some proposition X. Two important interpretations of $\varphi_i(A)$ are as (i) the (prior) conditional probability $C_i(X/A)$ that agent i assigns to X given A; (ii) the expected value to agent i of X given A. It is straightforward to check that the Sure–Thing Principle is satisfied for these two interpretations. Interpreting φ_i in the Theorem as a conditional probability yields Aumann's Agreement Theorem as a corollary.

Agreement Theorem: If two agents have the same prior probabilities, and their posterior probabilities for some proposition X are common knowledge, then those posterior probabilities must be equal.

Proof: Call the two agents i and j. Let w be the actual world. Let $\varphi_i(A) = C_i(X/A)$ and define φ_j similarly. These are both epistemic functions. The posterior probabilities a_i and a_j that i and j assign to X are common knowledge, i.e. it is common knowledge that $C_{iw}(X) = C_i(X/P_i(w)) = \varphi_i(P_i(w)) = a_i$, and that $C_{jw}(X) = C_j(X/P_j(w)) = \varphi_j(P_j(w)) = a_j$. Hence, by the Theorem proved above, there must be a proposition E such that $C_i(X/E) = a_i$ and $C_j(X/E) = a_j$. But since the two agents i and j have the same priors, it must be the case that $C_i(X/E) = C_j(X/E)$, in other words, that $a_i = a_j$, and so the two agents must assign the same posterior probability to X. QED.

Applying the second interpretation to the Theorem, i.e. taking φ_i to be an expected value, yields as a second corollary a version of the Milgrom–Stokey "No–Trade Theorem" which would appear to establish the impossibility of speculative trade between rational agents.[5] More precisely, suppose that two rational agents agree upon an efficient allocation of goods prior to each of them receiving some new information. No matter how different those two pieces of new evidence are, there is still no possible exchange of goods such that it is a matter of common knowledge that each agent wants to make the trade.

2 Common Belief

Yet rational agents can, and often do, believe what is false.[6] The proper subject of interactive epistemology ought to be belief rather than knowledge. It is instructive to see to what extent the development of the preceding section can be followed once the truth assumption has been dropped.

We shall continue to use the notation P_i for agent i's belief structure. However, we now no longer require that $w \in P_i(w)$, insisting only on the weaker condition that each $P_i(w)$ be a consistent belief state. We impose, in other words, the conditions: (B1) $P_i(w) \neq \emptyset$; (B2) if $v \in P_i(w)$ then $P_i(v) = P_i(w)$. Just as before, we may define, in terms of the belief structure, a belief operator B_i for each agent $i \in I$. $B_iA = \{w : P_i(w) \subseteq A\}$ is the proposition that agent i believes that A. The belief operators satisfy: (B1) $B_i\emptyset = \emptyset$; (B2a) $B_i \subseteq B_iB_iA$; (B2b) $\neg B_iA = B_i\neg B_iA$. These are the assumptions of consistency, and of positive and negative introspection for belief.

A proposition A is said to be *common belief* among the agents in I at world w iff for any n and any sequence $i_1, i_2, \ldots, i_n \in I$ we have $B_{i_1}B_{i_2}\ldots B_{i_n}A$. If $w \notin P_i(w)$ then agent i falsely believes that $P_i(w)$. She has eliminated the actual world from her set of doxastic possibilities. Call $\{w\}$ a *doxastic blindspot* for i if $w \notin P_i(w)$. Any proposition A that is a union of doxastic

[5] See Milgrom and Stokey [6].

[6] Here and throughout this paper when I talk of "belief" I mean *full belief.* To believe a proposition in this sense is to assign it probability 1.

blindspots for i, will also be called a blindspot for i. If $\{w\}$ is a blindspot for i, then at w agent i fully, and falsely, believes that w is not the actual world. In other words i's posterior probability at w for the blindspot $\{w\}$ is zero. In general, if A is any blindspot for i then $C_{iw}(A) = 0$

As before we define an accessibility relation R_i that obtains between worlds v and w iff $P_i(v) = P_i(w)$. Each R_i is an equivalence relation. Note, however, that the equivalence classes of R_i do not coincide with the various $P_i(w)$, as was the case before. This is precisely because of the presence of blindspots which are not consistent with any of the $P_i(w)$. Instead, the conditions imposed on belief structures ensure that each equivalence class of R_i is the union of some $P_i(w)$ with the blindspot $P_i^*(w)$ which has as elements all those states at which i falsely believes that $P_i(w)$. Let $R_i(w)$ be the equivalence class of R_i that includes w.

As before, say that $vR_I'w$ iff vR_iw for some $i \in I$, and let its transitive closure be R_I. This equivalence relation induces a partition of W that is the finest common coarsening of the partitions induced by each of the R_i. This time it is the common belief partition for the agents in I.

Common Belief Characterized: A is common belief among the agents in I at w iff A is entailed by $R_I(w)$, the equivalence class of R_I that contains w.

We strike trouble however when we try to reproduce the proof of the theorems for the case of common belief rather than common knowledge. This is due to the presence of doxastic blindspots. If we define E to be $R_I(w)$ then E is a disjoint union of $R_i(v)$ for various v, but no longer a disjoint union of $P_i(v)$, since each $R_i(v) = P_i(v) \cup P_i^*(v)$, where $P_i^*(v)$ is a doxastic blindspot for i. Hence the Sure–Thing Principle alone will not suffice to ensure that $\varphi_i(E) = \varphi_i(P_i(w))$. We need to rule out any influence that the possibilities in the $P_i^*(v)$ may have on the function φ_i. This amounts to imposing the condition that i's blindspots be invisible to the function φ_i. That thought motivates the following definition:

Call φ_i a *doxastic function* for i if $\varphi_i(A)$ is defined whenever $C_i(A) > 0$ and φ_i satisfies both the Sure–Thing Principle and:

Invisibility of Blindspots: If $\varphi_i(A)$ is well-defined and B is a blindspot for i, then $\varphi_i(A \cup B) = \varphi_i(A)$.

Main Theorem: Suppose that for each $i \in I$, φ_i is a doxastic function and that it is a matter of common belief at w that this function assigns the value a_i to i's belief state. Then there is some proposition E such that for all i, $\varphi_i(E) = a_i$.

Proof: Let $E = R_I(w)$, the cell of the common belief partition for I that contains w. It is commonly believed at w that φ_i assigns the value a_i to i's belief state. In other words, the proposition $\{v : \varphi_i(P_i(v)) = a_i\}$ is a matter of common belief at w. Hence $\varphi_i(P_i(v)) = a_i$ for all $v \in E$. But E is a disjoint union of equivalence classes of R_i, each of which is of the form $P_i(v) \cup P_i^*(v)$. Hence by the Sure–Thing Principle $\varphi_i(E) = \varphi_i(P_i(w) \cup P_i^*(w))$, which is equal to $\varphi_i(P_i(w)) = a_i$ since i's blindspots are invisible to the function φ_i. This completes the proof.

Once again we may check that if $\varphi_i(A)$ is taken to be the prior conditional probability $C_i(X/A)$ that i assigns to X given A then φ_i is a doxastic function. The Sure–Thing Principle is satisfied as before, while i's blindspots will be invisible to the function φ_i if we assume that the prior probability that i assigns to those blindspots is zero.

Zero Priors for Blindspots: $C_i(A) = 0$ whenever A is a blindspot for i.

We now prove, in much the same way as before, a common belief version of the Agreement Theorem.

Agreement Theorem for Common Belief: Suppose that two agents have the same prior probabilities and each of those agents assigns zero prior probability to her own blindspots. Then, if their posterior probabilities for some proposition X are a matter of common belief, those posterior probabilities must be equal.

Proof: Call the two agents i and j. Let w be the actual world. Let $\varphi_i(A) = C_i(X/A)$ and define φ_j similarly. These are both doxastic functions since each agent assigns prior probability zero to any proposition that is a blindspot for herself. The posterior probabilities a_i and a_j that i and j assign to X are a matter of common belief, i.e. it is common belief that $C_{iw}(X) = C_i(X/P_i(w)) = \varphi_i(P_i(w)) = a_i$, and, similarly, that $C_{jw}(X) = a_j$. Hence, by the Theorem proved above, there must be a proposition E such that $C_i(X/E) = a_i$ and $C_j(X/E) = a_j$. But since the two agents i and j have the same priors, it must be the case that $C_i(X/E) = C_j(X/E)$, in other words, that $a_i = a_j$, and so the two agents must assign the same posterior probability to X.

3 Significance of the Agreement Theorem

Taking belief rather than knowledge to be the notion central to interactive epistemology allows us to develop an account of common belief parallel to the standard accounts of common knowledge in the literature, and, as we have seen, it is then possible to prove a common belief version of Aumann's Agreement Theorem. This result holds on the assumption that each agent assigns prior probability zero to any proposition that is, for her, what we termed a "doxastic blindspot". But how plausible is this assumption of Zero Priors?

If A is a blindspot for i it is obvious that i's *posterior* probability for A must be zero. But why should it also be the case that i's *prior* probability for A be zero? Agent i's posterior probabilities in w are obtained from her priors by conditionalizing on her beliefs in that world, i.e. $C_{iw}(-) = C_i(-/P_i(w))$. Hence the assumption of Zero Priors is not needed to ensure that i's blindspots are assigned zero posterior probability.

The incoherence of my belief that I believe falsely that it is raining is an essentially first–person phenomenon. It disappears on shifting to the third–person; there is nothing odd about my claiming that *he* believes falsely that it is raining. The sense of strangeness may also be dispelled in a future tense version. The thought that I *will* have the false belief that it is raining at some point in the future may just be an appropriately modest recognition of my own fallibility.

But the Agreement Theorem cannot be derived without the assumption of Zero Priors. If the agents have common priors that violate this assumption, then it is possible for their posterior probabilities to differ despite being a matter of common belief. Here is an example:

Example 1: Let the agents be i and j. Suppose that $C_i = C_j$ is the uniform distribution over W, and let $A = \{u, v, w\}$ be some proposition such that P_i assigns A to each world in A, while P_j assigns $\{u, v\}$ to each A–world. Then $\{w\}$ is a blindspot for j to which, in violation of Zero Priors, j assigns positive prior probability. Each agent has posterior probabilities that are constant over A, and hence these posteriors must be a matter of common belief at each world in A, since A is an equivalence class with respect to R_I. However those posterior probabilities are not equal. $C_{iu}(\{w\}) = C_i(\{w\}/P_i(u)) = 1/3$, while $C_{ju}(\{w\}) = 0$

The principle of Zero Priors for Blindspots is, however, likely to appeal to those Bayesians who think that there are diachronic rationality constraints on credence. Any agent who assigns positive prior probability to a proposition that is, for her, a blindspot, thereby considers it positively probable that at some particular later time she will come to assign probability 1 to a proposition that is false. Such an agent would be violating the *Reflection Principle.*

A quick way of responding to that question would be to point out that a rational agent's posterior probabilities are obtained from her priors by conditionalization. But this kind of answer is likely to appeal only to those Bayesians who think that there are diachronic rationality constraints on credence—those, for example, who are fans of the Diachronic Dutch Book Argument or the Reflection Principle.[7] Others are less likely to be impressed. That updating goes by conditionalization on one's priors has been built into the formal framework we have developed here. "So much the worse for that formal framework", reply those who are skeptical of diachronic constraints on credence. They will, accordingly, not be worried by the common belief version of the Agreement Theorem we have proved here. Something else will have to give, of course, but there are a number of options to explore.

Let's look at a sample situation in detail. Consider agent i, who, in world w has just received some information that has led her to update her belief state to $P_i(w)$. Suppose that there is a non-empty set of worlds $P_i^*(w)$ at which i falsely believes that $P_i(w)$, and suppose further that i assigns some positive prior credence to $P_i^*(w)$. How might we ensure that i's posterior credence for the blindspot $P_i^*(w)$ is zero? One way would be to hang on to conditionalization, but claim that the proper content of the information on which i updates is $P_i(w)$ rather than $P_i(w) \cup P_i^*(w)$. This would require some explanation, since on the interpretation offered above, all of the worlds in $P_i(w) \cup P_i^*(w)$ are supposed to be subjectively indistinguishable to i, in which case it is a little hard to see how any proper subset of this proposition

[7] The Diachronic Dutch Book Argument for conditionalization is due to David Lewis and described in Teller [10]. The Reflection Principle was introduced in van Fraassen [11]. For criticism of these sorts of diachronic constraint see Levi [4] and Christensen [2].

could possibly be the content of a piece of informational input for i. Another way would be simply to deny that updating is always by conditionalization, i.e. to deny that $C_{iw}(A)$ must equal $C_i(A/P_i(w) \cup P_i^*(w))$. This would certainly block the proof of the theorem given above. However, it is not clear to me that this response really defuses the Common Belief Agreement Theorem. It may well still be possible to prove the Theorem in a slightly different, and in fact, more direct way.

Let $C_{iA}(X)$ be the posterior probability that i assigns to the proposition X after updating her priors on information A. For the moment we no longer assume that $C_{iA}(X) = C_i(X/A)$. One might now reasonably argue that, for fixed X, the function $\varphi_i(A) = C_{iA}(X)$ is a doxastic function in the sense defined above. That would allow us to prove the Agreement Theorem even more directly than before, without appeal to the disputed principle of Zero Priors. Is this function φ_i a doxastic function? i's doxastic blindspots are certainly invisible to the function φ_i, since it is not disputed that the *posterior* probability i assigns to her blindspots is zero. That leaves the issue of the Sure–Thing Principle: does the function φ_i have the property of being preserved under disjoint union? I claim that the answer to that question is yes. Even if one allows that rational updating of credences sometimes proceeds by a method other than conditionalization, this weaker condition on an updating method should still apply to whatever revision rule is envisaged instead.

In summary, while one might reasonably be skeptical about the principle of Zero Priors (as about other diachronic constraints on rational credence) it is not clear that one can escape the bite of the Common Belief Agreement Theorem simply by rejecting that condition. If one's updating method satisfies the Sure–Thing Principle, then whether or not one updates by conditionalization, the undisputed principle of Zero Posteriors suffices to establish the Agreement result. It would appear that the condition of Zero Priors is not really the main point here.

It is more interesting to consider how things look if we are willing to grant the principle of Zero Priors for First–Person Blindspots. We can then certainly prove a common belief version of the Agreement Theorem. However, it seems to me that the theorem we have then proved—though perfectly valid—has no teeth, for the key assumption of its antecedent, the assumption of equal prior probabilities, has no plausibility. The point is that my blindspots and your blindspots are two quite different sets of propositions, and while I must assign zero probability to my own blindspots if my credences are to be coherent, I have no more reason to assign zero probability to a proposition that is a blindspot only for you than I have to assign zero probability to any other contingent proposition. What this means is that if we have beliefs about each other's beliefs, you and I cannot coherently have equal priors.

Once our two agents are assumed to have the same priors, then the condition of Zero Priors for First–Person Blindspots amounts to the totally implausible

Zero Priors for Blindspots: $C_i(A) = 0$ whenever A is a blindspot for any agent in I.

and we arrive at the following, more revealing, statement of the theorem.

Common Belief Agreement Theorem (Restated): Suppose that two agents have the same prior probabilities and each of those agents assigns zero prior probability to any proposition that is a doxastic blindspot *either for herself or for the other agent*. Then if their posterior probabilities for some proposition X are a matter of common belief, those posterior probabilities must be equal.[8]

Our posterior probabilities for a certain proposition may be a matter of common belief, and yet unequal, whenever my evidence suggests to me that you are mistaken, or your evidence leads you to suspect I am mistaken. This point, once stated, might seem so obvious as not to be worth making, and it wouldn't be worth making, except for the fact that it has been obscured by a large literature that has focussed almost entirely on knowledge rather than on belief.

The following simple example should suffice to demonstrate the possiblity of two rational agents agreeing to disagree.

Example: Let the agents be i and j. Let $P_i(w_1) = P_i(w_2) = P_i(w_3) =$

[8] Dov Samet considers dropping the truth requirement in in [7]. As Samet puts it (p.191): "we may allow for false propositions to be 'known' ". Samet's Theorem 8 (p.202) is similar to the result proved here. The antecedent of Samet's Theorem explicitly includes the requirement that each agent assign zero probability to any proposition which is, in our terms, a doxastic blindspot either for herself *or for any other agent*. Samet's description of this as a "consistency" requirement is rather misleading, since the crucial difference between the first–person and the third–person cases has been ignored.

$\{w_1, w_2, w_3\}$, and $P_j(w_1) = P_j(w_2) = P_j(w_3) = \{w_1, w_2\}$. Note that $\{w_3\}$ is a blindspot for j, but not for i. Suppose that w_1 is the actual world, and let X be the proposition $\{w_1\}$. We have $R_i(w_1) = R_j(w_1) = \{w_1, w_2, w_3\}$, so the proposition whose existence is guaranteed by the Main Theorem is $E = \{w_1, w_2, w_3\}$. Suppose that $C_i(w_1) = C_i(w_2) = C_i(w_3)$, and that $C_j(w_1) = C_j(w_2)$, while $C_i(w_3) = 0$. Then $C_i(X/E) = 1/3$ and $C_j(X/E) = 1/2$. These posterior probabilities are a matter of common belief at w_1 but they are not equal. This is because i thinks that j may be mistaken in having eliminated the possibility w_3.

Finally, an opponent might challenge my claim that it is belief rather than knowledge that ought to be central to interactive epistemology. My response to this is simply to point out that agents, even rational agents, can and do get things wrong. This is not a controversial claim, just the commonplace observation that rational agents sometimes have false beliefs. The reason for this is not hard to find. It is because the input on which we update is sometimes misleading and sometimes downright false. To demand that everything an agent fully believes be true is not to state a requirement of rationality but rather to demand that the agent be invariably lucky in the course of her experience. Being completely rational is one thing; always being lucky is another.

I suspect that many economists working on these matters are unconcerned about the distinction between knowledge and belief and about the strength of the truth requirement because they are convinced that there is a level of description of informational input at which the agent cannot get it wrong.[9]

Suppose that I am looking at an object. I may be wrong in my perceptual judgement that the object *is* yellow, but surely I cannot be wrong about the fact that it *looks* or *seems to me* to be yellow. The thought then is that, when I look at the object, an expression of the perceptual content on which I properly update my credences ought to be couched in the "looks" or "seems" language that carries with it the guarantee of truth. Then if I start from coherent priors and proceed to update by conditionalization on informational contents of this kind I can be certain that I will never fully believe anything that is false.

[9] John Geanakoplos made a suggestion to me along these lines in conversation following a paper he read at Columbia in 1994. My apologies to him if I have misremembered or misunderstood the point he was making.

This is a familiar strategy. The difficulties with it are also familiar. The main problem with this line of thought is that it proves far more difficult than one might at first imagine to get from the epistemically privileged language of "looks" and "seems" to the ordinary everyday language of physical objects. And this gap will have to be bridged, since we want our agents to have beliefs about the world rather simply about their own mental states. Such problems are well-known to anyone acquainted with the history of the heroic failure of the philosophical program of phenomenalism.[10]

In a context other than that of interactive epistemology, we might perhaps agree that differences in subjective probability assignments should always be traceable to differences in evidence. Once, however, we have allowed that agents may have beliefs about the (possibly false) beliefs of others, we find that a first–person/third–person asymmetry in the logic of belief makes the assumption of equal priors impossible to maintain. If B is, for me, what we have called a blindspot proposition, then I must assign a probability of zero to B. Neither you, nor any other agent, are so constrained in your probability assignment. The real culprit in the Agreement Theorem is not the assumption of common belief, but rather the truth condition required for that belief to count as knowledge. Common belief is not problematic in the way that common knowledge is. Rational agents can agree to disagree.

References

[1] Aumann, Robert. 'Agreeing to Disagree,' *The Annals of Statistics* (1976) pp. 1236–1239.

[2] Christensen, David. 'Clever Bookies and Coherent Beliefs,' *The Philosophical Review* (1991) pp. 229–247.

[3] Geanakoplos, John. 'Common Knowledge,' in *Handbook of Game Theory*, Vol.2, Aumann and Hart (eds.) Elsevier (1994).

[4] Levi, Isaac. 'The Demons of Decision,' *The Monist* (1987) pp. 193–211.

[5] Lewis, David. *Convention*, Harvard University Press (1969).

[10] This is not the place to recount that history. One famous diagnosis of the failure of the phenomenalist program may be found in Sellars [8].

[6] Milgrom, P. and Stokey, N. ‘Information, Trade, and Common Knowledge,’ *Journal of Economic Theory* (1982) pp. 17–27.

[7] Samet, Dov. ‘Ignoring Ignorance and Agreeing to Disagree,’ *Journal of Economic Theory* (1990) pp. 190–207.

[8] Sellars, Wilfrid. ‘Empiricism and the Philosophy of Mind,’ in *Science, Perception, and Reality*, Routledge and Kegan Paul (1963).

[9] Sorensen, Roy A. *Blindspots*, Oxford University Press (1988).

[10] Teller, Paul. ‘Conditionalization and Observation,’ *Synthese* (1973) pp. 218–258.

[11] van Fraassen, Bas. ‘Belief and the Will,’ *The Journal of Philosophy* (1984) pp. 235–256.