Agreeing to Disagree*

Matthias Hild
Richard Jeffrey
Mathias Risse

April 24, 1997

In "Agreeing to Disagree" Robert Aumann proves that a group of agents who once agreed about the probability of some proposition for which their current probabilities are common knowledge must still agree, even if those probabilities reflect disparate observations. Perhaps one saw that a card was red and another saw that it was a heart, so that as far as that goes, their common prior probability of 1/52 for its being the Queen of hearts would change in the one case to 1/26, and in the other to 1/13. But if those are indeed their current probabilities, it cannot be the case that both know them, and both know that both know them, etc., etc.

Aumann's Theorem

Aumann's framework provides for a finite number of *agents*; call them $i = 1, \cdots, N$. These individuals are about to learn the answers to various multiple-choice questions—perhaps by making observations. The possible answers to agent i's question form a finite set $\mathcal{Q}_i$ of mutually exclusive, collectively exhaustive propositions. We think of propositions as subsets of a non-empty set Ω of *worlds*, representing all possibilities of interest for the problem at hand. Then $\mathcal{Q}_i$ is a *partition* of Ω: each world ω belongs to exactly one element of each $\mathcal{Q}_i$. We call that element '$\mathcal{Q}_i\omega$'.

Perhaps it is only certain propositions, certain subsets of Ω, that are of interest to the agents. Among them will be all propositions in any of the N partitions, and perhaps other propositions as well. We will suppose that they form a field, $\mathcal{A}$.[1]

The propositions agents know after learning the answers to their questions are the members of $\mathcal{A}$ that those answers imply: In world ω agent i knows A if and only if $\mathcal{Q}_i\omega \subseteq A \in \mathcal{A}$. Then for each i, Def. K, below, defines

*This is drawn from the Hild–Jeffrey–Risse papers listed in the bibliography.

[1]To call $\mathcal{A}$ a field is to say it is closed under the operations of denial and conjunction.

a knowledge operator K_i which, applied to any set $A \in \mathcal{A}$, yields the set $K_iA \in \mathcal{A}$ of worlds in which i knows A.

$$\textbf{Def. K}: \quad K_iA = \{\omega : \mathcal{Q}_i\omega \subseteq A\}$$

For each natural number n, Def. MK then defines an operator M_n, "n'th degree mutual knowledge": A is true, and everybody knows that, and everybody knows *that*, etc.—with n 'knows'. Finally, Def. CK defines *common knowledge*, κ, as mutual knowledge of all finite degrees:

$$\textbf{Def. MK}: \quad M_0A = A, \quad M_{n+1}A = \bigcap_{i=1}^{N} K_iM_nA$$

$$\textbf{Def. CK}: \quad \kappa A = \bigcap_{n=0}^{\infty} M_nA$$

The proof of Aumann's Theorem uses the following lemma, which says that where A is common knowledge, everyone knows it is.

Lemma: If $\omega \in \kappa A$, then for each i, $\mathcal{Q}_i\omega \subseteq \kappa A$.

Proof. If $\omega \in \kappa A$ then by Def. CK and Def. M, $\omega \in K_iM_nA$ for all agents i and degrees n of mutual knowledge. Therefore by Def. K, $\mathcal{Q}_i\omega \subseteq M_nA$ for all n, and thus by Def. CK, $\mathcal{Q}_i\omega \subseteq \kappa A$.

The first hypothesis of Aumann's Theorem, below, says that $\langle \Omega, \mathcal{A}, P\rangle$ is a probability space, that $\mathcal{A}$ includes each of the partitions $\mathcal{Q}_1, \ldots, \mathcal{Q}_N$ of Ω, and that each of those partitions is finite. In jargon: $\langle\langle \Omega, \mathcal{A}, P\rangle, \mathcal{Q}_1, \ldots, \mathcal{Q}_N\rangle$ is a *finite partition space*. $\mathcal{A}$ will then be closed under all of the operators K_i and M_n, and under κ.

P is the old probability measure that is common to all the agents. In Aumann's theorem we consider a hypothesis $H \in \mathcal{A}$ for which the various agents' probabilities are $q_1, \ldots, q_N$ after they condition P on the answers to their questions. The proposition $C \in \mathcal{A}$ identifies these probabilities:

$$\textbf{Def. C}: \quad C = \bigcap_{i=1}^{N} \{\omega : P(H|\mathcal{Q}_i\omega) = q_i\}$$

The second hypothesis says that the possibility of C's becoming common knowledge is not ruled out in advance: $P(\kappa C) \neq 0$.

Aumann's Agreement Theorem:
Suppose $\langle\langle \Omega, \mathcal{A}, P\rangle, \mathcal{Q}_1, \ldots, \mathcal{Q}_N\rangle$ is a finite partition space, and $P(\kappa C) > 0$. Then $P(H|\kappa C) = q_1 = \ldots = q_N$.

Proof. By the lemma, $\kappa A = D_{i1}$, or $\kappa A = D_{i1} \cup D_{i2}$, or $\kappa A = D_{i1} \cup D_{i2} \cup D_{i3}$, or in general $\kappa A = \bigcup_j D_{ij}$ where the D_{ij} are distinct members of $\mathcal{Q}_i$. Then

$$P(H|\kappa C) = \frac{P(H \cap \bigcup_j D_{ij})}{P(\bigcup_j D_{ij})} = \frac{\sum_j P(H|D_{ij})P(D_{ij})}{\sum_j P(D_{ij})} = \frac{\sum_j q_i P(D_{ij})}{\sum_j P(D_{ij})} = q_i,$$

where the first equation is justified by the quotient rule and the Lemma; the second by distributivity of $\cap$ over $\bigcup_j$ and by the additivity and product rules; the third by Def. C; and the fourth by factoring out the constant q_i.

Uncommon Knowledge

In common usage, 'common knowledge' refers to what everybody knows, i.e., technically, first degree mutual knowledge. Common knowledge in the technical sense is another matter, harder to come by. Non-linguistic animals may have modest degrees of mutual knowledge, but in practice, the higher degrees seem to require language, as common knowledge seems to do in principle.

Consider, e.g., the puzzle of the three dirty-faced children: Alice, Bob, Claire. Each sees that the other two have dirty faces, so each knows that *not all their faces are clean*, i.e. mutual knowledge of degree 1. And of degree 2, as well: Each sees that the other two have dirty faces, so each sees that each of the other two sees that the other one's face is dirty; and of course, in seeing a dirty face, each knows that he or she knows that not all their faces are clean. But third degree mutual knowledge that not all their faces are clean is lacking, as can be seen by considering what Alice thinks possible. As far as she knows, it is possible that her face is clean:

$$\langle Alice \rangle \ \textit{Alice has a clean face}$$

(In general, the notation $\langle i \rangle$ is to be read: *As far as i knows, it is possible that.*) And as far as Alice knows, it is possible that as far as Bob knows, both of their faces are clean—for as far as she knows, her face may be clean, in which case Bob would see that, and still think his face might be clean:

$$\langle Alice \rangle \ \langle Bob \rangle \ \textit{Alice and Bob have clean faces}$$

And, indeed, as far as Alice knows, Bob, thinking his face may be clean, may see that her face is clean, and think that as far as Clare knows, all faces are clean:

$$\langle Alice \rangle \ \langle Bob \rangle \ \langle Claire \rangle \ \textit{All their faces are clean}$$

(Indeed, any permutation of the names in this statement will leave it true.) Then the fact that *not* all their faces are clean is not mutual knowledge of degree 3, and, so cannot be common knowledge.

Now suppose their mother tells them: "Not all your faces are clean!" They believe her, and in fact it is common knowledge among them that they do. (The story is told about a mother and her children in order to make all this seem a matter of course.) Their mother has stated a physical fact that each already knows, but because of their common knowledge of their trust in her, their individual knowledge that she has made that statement informs them of mental facts they had not known—in particular, that *all know that all know that all know* the fact their mother enunciated. And this crumb of third degree mutual knowledge, combined with successive revelations of their ignorance, can tell all three children that their faces are dirty:

(1) Mother says: "Any who know your faces to be dirty, raise your hands!" No hands go up. (2) Mother asks a second time, and again no hands go up. (3) Mother asks a third time, and now all their hands go up.

What happened? Their mother's announcement that not all their faces were clean removed the possibility that $\langle Alice \rangle$ $\langle Bob \rangle$ $\langle Claire \rangle$ *All faces are clean*, and removed the possibilities corresponding to permutations of the three names in that statement. The proposition she enunciated was something they already knew; indeed, it was second degree mutual knowledge among them. But the fact that their mother enunciated it made it common knowledge among them. And the important point is that it made it mutual knowledge among them of degree three, which was exactly enough to let the second non–showing of hands show them that all their faces must be dirty; the additional degrees of mutual knowledge that make up common knowledge were not needed to solve their problem.

What their mother told them was a fragment of what they could all see for themselves. She told them that at least one of their faces was dirty, and each could each see that two of their faces were dirty. But each saw two different faces; they had three different bodies of evidence, of which each separately implied what their mother told them, and any two together implied that all their faces were dirty. Their mother structured the game so that they could not simply pool their evidence, and could not discover from what she told them that they all had dirty faces, but could discover it from the fact of her telling them what they could see for themselves.

This sort of reasoning was deployed by Geanokoplos and Polemarchakis to analyze the process by which successive announcements by $N = 2$ agents of their current probabilities for truth of A *as the announcements were made* eventually ends any disagreement in their probabilities for A—if it is common knowledge that each takes these successisve revelations into account by conditioning on the disjunction of the propositions that the other could have

conditioned upon to get the announced probability for A. In that process, too, it may be that no minds are changed until several announcements have been made, at which point the next announcement reveals total agreement ensuing from the previous announcement; and the rationale may be rather abstruse.

Evidently there are circumstances in which common knowledge can be hard to come by. Common knowledge need not be what every fool knows—as it may be in special cases, e.g., in the case of *self-evident* propositions, where[2]

$$A \text{ is self-evident} \Leftrightarrow \text{for all } i,\ \mathcal{P}_i(\omega) \subseteq A \text{ for all } \omega \in A$$

A self-evident event is a true self-evident proposition. For such events, it is evident to all that all have the same, conclusive evidence. But such conclusiveness is possible for for other events as well, e.g., for the event of the mother's declaring that not all faces were clean. Before that event took place, the fact that not all their faces were clean was not self-evident, for although all could see that it was true, all could also see that all had different visual evidence, and could puzzle out the fact that what they could plainly see was not mutual knowledge of degree three.[3]

We take these considerations to weigh against Aumann's view of his theorem as undermining the Harsanyi "Doctrine" that rational people will have the same prior:

> John Harsanyi (1968) has argued eloquently that differences in subjective probabilities should be traced exclusively to differences in information—that there is no rational basis for people who have always been fed precisely the same information to maintain different subjective probabilities. This, of course, is equivalent to the assumption of equal priors. The result of this paper might be considered evidence against this view, as there are in fact people who respect each other's opinions and nevertheless disagree heartily about subjective probabilities. (1237-8)

Now as Harsanyi points out, many games *are* aptly modelled by assuming a simple (1) *common prior*, shared by all players. On the other hand, as we have been pointing out, (2) *common knowledge of a shared posterior probability* may be very hard to come by; Geanokoplos and Polemarchakis's

[2]We adapt the terminology of Osborne and Rubinstein, p. 73.

[3]Lewis's (II, §1) often cited concept of common knowledge is often misrepresented as mutual knowledge of all finite degrees, our κ. In fact, his 'common knowledge' refers to states of affairs (sc., *bases* for common knowledge) in which, as he points out, modest degrees of mutual knowledge are expectable, with κ holding under idealized assumptions of a sort commonly made in the game theory literature. His discussion is of interest in the present context.

program for eliminating disagreement in such a way as to reach common knowledge of a shared posterior probability is not the sort of thing we easily understand and implement.[4] And, finally, as Aumann points out, we may well respect each other's opinions but (3) *disagree about posterior probabilities*. It is (1), (2), (3) *together* that are refuted by Aumann's theorem; one or more must go. We submit that in the game context, (2) is far the least plausible of the three, far the likeliest candidate for rejection.

References

Aumann, R.J. (1976), "Agreeing to Disagree", *Annals of Statistics* **4**, 1236–1239.

Aumann, R.J. (1995), "Interactive Epistemology", Discussion Paper No. 67, Center for Rationality and Interactive Decision Theory, The Hebrew University of Jerusalem.

Geanakoplos, J.D., Polemarchakis, H.M. (1982), "We Can't Disagree Forever", *Journal of Economic Theory* **28**, 192–200.

Harsanyi, J., (1967-1968), Games of Incomplete Information Played by Bayesian Players, Parts I-III, *Management Science***14**, 159-182, 320-334, 486-502

Hild, M., Jeffrey, R., and Risse, M. (forthcoming), Agreeing to Disagree: Harsanyi and Aumann, *Game Theory, Experience, Rationality*:Vienna Circle Institute Yearbook, Kluwer.

Hild, M., Jeffrey, R., and Risse, M. (forthcoming), Aumann's 'No Agreement' Theorem Generalized, *The Logic of Strategy*, C. Bicchieri, R. Jeffrey, and B. Skyrms, eds.: Oxford University Press.

Lewis, D. (1969), *Convention*, Cambridge, Mass.: Harvard University Press.

Jeffrey, R.C. (1992), *Probability and the Art of Judgment*, Cambridge: Cambridge University Press.

Osborne, M. and Rubinstein, A. (1994), *A Course in Game Theory*, Cambridge, Mass.: The MIT Press.

[4]Furthermore, it presupposes updating by conditioning on new certainties. For alternatives to simple conditioning, see Jeffrey, chapters 6 and 7..