Agreeing to Disagree

Contents

1 Introduction

The question whether two rational, Bayesian agents can agree to disagree goes back to Aumann, who presented the first formal result addressing this question. In this technical document we review preliminary theory required to understand Aumann's result, and then consider Aumann's theorem. Furthermore, we consider a few modern developments building on Aumann's theorem, and we assess the theory's applicability to reality.

2 The State-Space Model of Knowledge

The prevailing model of knowledge in information economics and game theory, originally introduced by Aumann (1976) and commonly called the state-space model of knowledge, considers a set Ω consisting of the possible *states of the world*. For each state of the world $\omega \in \Omega$, an agent's knowledge is represented by a subset $P(\omega) \subseteq \Omega$: the agent "knows" that the true state is within $P(\omega)$, but they cannot further identify whether any state in $P(\omega)$ is true. In other words, for each state of the world $\omega \in \Omega$, the agent "knows" to exclude any state in $\Omega \setminus P(\omega)$ from being the true state of the world, but they cannot further exclude any state in $P(\omega)$ from being true.

These *information functions* $P\colon \Omega \to \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ is the collection of subsets of Ω, are typically assumed to satisfy to following two conditions:

(P1) $\omega \in P(\omega)$; and
(P2) $\omega' \in P(\omega) \implies P(\omega') = P(\omega)$.

Condition (P1) states that $P(\omega)$ indeed contains the true state of the world. Condition (P2) states that the agent uses consistency of their information function to infer the true state of the world: If $P(\omega') \ni \omega'' \notin P(\omega)$, then if the true state is ω, the agent can exclude ω' from being true: $\omega'' \notin P(\omega)$, so ω'' cannot possibly be true; if, however, ω' were true, then $\omega'' \in P(\omega')$, which would not be possible since ω'' cannot possibly be true. Similarly, if $P(\omega') \not\ni \omega'' \in P(\omega)$, then if the true state is ω, the agent can exclude ω' from being true: $\omega'' \in P(\omega)$, so they cannot possibly exclude ω'' from being true;

if, however, ω' were true, then $\omega'' \notin P(\omega')$, which would not be possible since ω'' cannot be excluded from being true. Knowledge functions that satisfy (P1) and (P2) partition Ω: $I_P = \{P(\omega) : \omega \in \Omega\}$. These partitions I_P are called *information partitions*.

An *event* is represented by a set of states $E \subseteq \Omega$: the event E happens if $\omega \in \Omega$, and the event E does not happen if $\omega \notin \Omega$. For a given state of the world ω, the agent is said to "know E" if $P(\omega) \subseteq E$: for every state that the agent considers to be possible, the event E happens. An agent's *knowledge function* K gives the set of states $K(E)$ in which the agent knows E:

$$K : \Omega \to \Omega, \qquad K(E) = \{\omega \in \Omega : P(\omega) \subseteq E\}.$$

And $K(E)$ is also an event, the event in which the agent knows E. Hence, one could consider $K(K(E))$, the event in which the agent *knows that they know* E.

The knowledge function expresses a number of intuitive properties:

(K1) $K(\Omega) = \Omega$.

Property (K1), also called the principle of *necessitation*, states that in all states the agent knows that some state in Ω is true.

(K2) $K(E) \cap K(F) = K(E \cap F)$.

Property (K2), also called the principle of *conjunction*, states that the agent knows E and F if and only if the agent knows $E \cap F$.

(K3) $E \subseteq F \implies K(E) \subseteq K(F)$.

Property (K3), also called the principle of *monotonicity*, states that if F occurs whenever E occurs, then the agent knows F whenever they know E.

(K4) $K(E) \subseteq E$.

Property (K4), also called the principle of *truth*, states that if the agent knows E, then E has indeed occurred. This property depends on (P1).

(K5) $K(E) = K(K(E))$.

Property (K5), also called the principle of *transparency*, states that the agent knows E if and only if the agent knows that they know E. This property depends on (P1) and (P2).

(K6) $\Omega \setminus K(E) = K(\Omega \setminus K(E))$.

Property (K6), also called the principle of *wisdom*, states that the agent does not know E if and only if the agents knows that they do not know E. This property also depends on (P1) and (P2).

Alternatively, one may take (K1), (K2), and (K3) as axioms and let

$$P(\omega) = \bigcap \{E \subseteq \Omega : \omega \in K(E)\}.$$

Then additionally assuming (K4) and (K5) yields that P satisfies (P1) and (P2).

Consider two agents Alice A and Bob B. An event E is called *self evident* to Alice if Alice knows the event whenever it occurs; that is, E is self evident if $P_A(\omega) \subseteq E$ whenever $\omega \in E$. Note that a self-evident event E is the union of elements of the information partition: if the intersection between $P_A(\omega)$ and E is nonzero, $\omega' \in P_A(\omega) \cap E$, then $P_A(\omega) = P_A(\omega') \subseteq E$; therefore, $K_A(E) = E$ for any self-evident event E. Furthermore, at a state of the world ω, an event E is called *common knowledge* between Alice and Bob if $\omega \in F$ for some $F \subseteq E$ that is self evident to both Alice and Bob.

Alice's information partition I_A is called *finer* than Bob's information partition I_B if for every $S \in I_A$ it holds that $S \subseteq T$ for some $T \in I_B$; we also say that I_B *coarser* than I_A. Finer partitions correspond to "more" knowledge, since agents with finer partitions simply know more events.

Let the *meet* $I_A \wedge I_B$ be the finest common coarsening of I_A and I_B, which corresponds to the collection of smallest sets that are self evident to both Alice and Bob; thus, any self-evident event is made up from elements of $I_A \wedge I_B$. Denote by $(P_A \wedge P_B)(\omega)$ that element of $I_A \wedge I_B$ containing ω. It then holds that an event E is common knowledge at ω if and only if E includes $(P_A \wedge P_B)(\omega)$.

Let the *join* $I_A \vee I_B$ be the coarsest common refinement of I_A and I_B. Note that every event E self evident to Alice and Bob is the union of elements in $I_A \vee I_B$.

Finally, we define Alice's and Bob's *interactive knowledge* K^n as follows:

$$
\begin{aligned}
K^1(E) = K_A(E) \cap K_B(E)&: && \text{“Alice and Bob know } E\text{”},\\
K^2(E) = K_A(K^1(E)) \cap K_B(K^1(E))&: && \text{“Alice and Bob know that they know } E\text{”},\\
K^3(E) = K_A(K^2(E)) \cap K_B(K^2(E))&: && \text{“Alice and Bob know that they know that they know } E\text{”},\\
&&& \vdots\\
K^\infty(E) = \bigcap_{n=1}^{\infty} K^n(E)&: && E \text{ is } \textit{common knowledge} \text{ between Alice and Bob.}
\end{aligned}
$$

Theorem 2.1. The two definitions of common knowledge are equivalent.

Proof. Suppose that some event $F \subseteq E$ is self evident to both Alice and Bob. Then

$$
\begin{aligned}
& F = K_A(F) \subseteq K_A(E) && \wedge && F = K_B(F) \subseteq K_B(E) && \implies && F \subseteq K^1(E)\\
\implies\ & F = K_A(F) \subseteq K_A(K^1(E)) && \wedge && F = K_B(F) \subseteq K_B(K^1(E)) && \implies && F \subseteq K^2(E)\\
\implies\ & F = K_A(F) \subseteq K_A(K^2(E)) && \wedge && F = K_B(F) \subseteq K_B(K^2(E)) && \implies && F \subseteq K^3(E),
\end{aligned}
$$

so by induction it follows that $F \subseteq K^\infty(E)$.

Conversely, we show that $K^\infty(E)$ is self evident to both Alice and Bob. Let $\omega \in K^\infty(E)$. We show that $P(\omega) \subseteq K^\infty(E)$, which amounts to showing that $P(\omega) \subseteq K^n(E)$ for every n. Now,

$$
\omega \in K^{n+1}(E) = K_A(K^n(E)) \cap K_B(K^n(E)) \qquad \implies \qquad P_A(\omega) \subseteq K^n(E),
$$

so indeed $P_A(\omega) \subseteq K^\infty(E)$. That $K^\infty(E)$ is self evident to Bob is shown similarly. □

3 Soundness of the State-Space Model

Although the state-space model provides a set-theoretic framework that is convenient to handle, some properties of K appear dubious from an intuitive and philosophical point of view . Especially necessitation of monotonicity are often-criticised properties.

Firstly, necessitation dictates that an agent knows *every* event that is always true. In other words, if we consider the truths that hold for every state of the world, then the agent must know all of them. Secondly, let E be the axioms of a logical system, and let F be the theorems valid in that logical system. Then $E \subseteq F$, so monotonicity implies that the agent must know all theorems whenever it knows the axioms. Thirdly, the principle of knowledge states that an event occurs whenever an agent knows it. In other words, an agent cannot have false knowledge. Finally, the principle of wisdom states that an agent can only not know an event if the agents knows that they do not know it.

Two additional properties, *substitutivity* and *immediacy*, provide further insight in the assumptions ingrained in the model . The principle of substitutivity states that $K(E) = K(F)$ whenever $E = F$. Although mathematically trivial, substitutivity is philosophically troublesome. To illustrate this, let

E be the states in which triangles are equilateral, and let F be the states in which triangles are equiangular. Assuming Euclidean geometry, equilateral triangles are in fact equiangular, so $E = F$. This implies that $K(E) = K(F)$, which means that an agent cannot be unaware of the fact that equilateral triangles are equiangular.

The principle of *immediacy* states that an agent knows all events that are supersets of $P(\omega)$. Immediacy emphasises that there is no difference between potential knowledge and actual knowledge: the agent necessarily and immediately knows all such events .

3.1 Extension and Intension

To further understand the principle of substitutivity, it is helpful to consider the notions of *extension* and *intension*. The extension of a linguistic expression refers to the collection of things that the expression *designates*, whereas the intension of the expression refers to the idea or notion conveyed . For example, the extension of "computer" is the collection of physical computers that exist in our world, but the intension of "computer" is the idea of a device that can perform computation.

The principle of substitutivity now shows that the state-space model of knowledge respects extensional equality of events, but completely disregards the intensional dimension. Returning to the example of equilateral triangles, the fact that "the triangle is equilateral" is extensionally equal to "the triangle is equiangular" necessarily means that the respective events are equal illustrates the extensional nature of the model. That one may not equate the idea of an equilateral triangle to the idea of an equiangular triangle shows that the model completely disregards the intensional dimension.

3.2 Knowledge of Information Partitions

Consider the event that an agent knows that they, or some other agent, knows something. Does this require the assumption that agent has knowledge of their own and others' information partition?

A common answer is that the agents' mental states are part of the state of the world. It then follows that each agent has a unique information partition: Suppose that Alice knows that Bob's knowledge is $P_B(\omega)$ if ω is true, then for all $\omega' \in P_B(\omega)$ she must reason that Bob's knowledge is $P_B(\omega)$ as well (c.f. the justification of (P2)). Intuitively, this answer argues that for every state of the world, agents are capable of reasoning what they and other know and would have known, would the state of the world have been different.

A more satisfactory answer follows from the principles of immediacy and substitutivity. Again consider Alice and Bob, and consider an event E that happens to be equal to Alice knowing some other event F: $E = K_A(F)$. If it happens to be that $P(\omega) \subseteq E$, then by immediacy Bob necessarily knows E, which by substitutivity means that Bob knows that Alice knows F: no assumption about either Alice's of Bob's information partition is required. In other words, knowledge of knowledge is justified by the principles of immediacy and substitutivity, meaning that one should instead question the soundness of these principles.

4 Agreeing to Disagree

Suppose that Alice and Bob have a common prior μ, a probability measure on (Ω, Σ). This assumption is called the *Common Prior Assumption* (CPA). Let their posterior beliefs about some event E be respectively $q_A(E) = \mu(E \,|\, P_A(\omega))$ and $q_B(E) = \mu(E \,|\, P_B(\omega))$. Surprisingly, Alice and Bob cannot agree to disagree about E:

Theorem 4.1 (Aumann (1976))**.** Assume that Ω is countably infinite and $I_A \vee I_B$ consists of non-null events. If it is common knowledge that $q_A(E) = q_A$ and $q_B(E) = q_B$, then $q_A = q_B$.

Proof. Since $q_A(E) = q_A$ and $q_B(E) = q_B$ is common knowledge between Alice and Bob, there exists an event F that is self evident to both Alice and Bob, meaning that $\mu(E \mid F) = q_A$ and $\mu(E \mid F) = q_B$. Clearly, $q_A = q_B$.

More carefully, partition F according Alice's and Bob's information partitions:

$$\begin{aligned} I_A &= I_A^1 \cup I_A^2 \cup \cdots, & F_A^i &= F \cap I_A^i, \\ I_B &= I_B^1 \cup I_B^2 \cup \cdots, & F_B^i &= F \cap I_B^i, \end{aligned}$$

where we used the assumption that Ω is countably infinite. Then

$$q_A = \mu(E \mid F_A^i) = \frac{\mu(E \cap F_A^i)}{\mu(F_A^i)}, \qquad q_B = \mu(E \mid F_B^i) = \frac{\mu(E \cap F_B^i)}{\mu(F_B^i)},$$

where the assumption that $I_A \vee I_B$ consists of non-null events guarantees that $\mu(F_A^i) > 0$ and $\mu(F_B^i) > 0$. Therefore,

$$q_A \mu(F_A^i) = \mu(E \cap F_A^i), \qquad q_B \mu(F_B^i) = \mu(E \cap F_B^i),$$

so summing over i yields the intended result. □

Corollary 4.1. If cannot be common knowledge that $q_A(E) \neq q_B(E)$.

For simplicity, henceforth assume that Ω is finite, and that all $\omega \in \Omega$ have positive prior probability. Consider a real-valued random variable $X \colon \Omega \to \mathbb{R}$. Denote by $\mathbb{E}(X|I_A)$ the conditional expectation of X given $P_A(\omega)$; that is, $\mathbb{E}(X|I_A)$ denotes the random variable

$$\mathbb{E}(X|I_A) \colon \Omega \to \mathbb{R}, \qquad \mathbb{E}(X|I_A)(\omega) = \mathbb{E}(X \mid P_A(\omega)).$$

Lemma 4.1. $\mathbb{E}(X \mid I_A \wedge I_B) = \mathbb{E}(\mathbb{E}(X \mid I_A) \mid I_A \wedge I_B)$.

Proof. By definition of conditional expectation, we find that

$$\sum_{\omega' \in P_A(\omega)} X(\omega')\mu(\omega') = \mathbb{E}(X \mid P_A(\omega)) \sum_{\omega' \in P_A(\omega)} \mu(\omega') = \sum_{\omega' \in P_A(\omega)} \mathbb{E}(X \mid P_A(\omega'))\mu(\omega), \tag{1}$$

since $\mathbb{E}(X \mid P_A(\omega')) = \mathbb{E}(X \mid P_A(\omega))$ for all $\omega' \in P_A(\omega)$. Therefore, summing over all Equation (1) such that $P_A(\omega) \subseteq (P_A \wedge P_B)(\omega)$ yields the intended result. □

Theorem 4.2. If it is common knowledge that $\mathbb{E}(X \mid I_A) \leq \mathbb{E}(X \mid I_B)$, then $\mathbb{E}(X \mid I_A) = \mathbb{E}(X \mid I_B)$, and this is common knowledge.

Proof. Since $\mathbb{E}(X \mid I_A) \leq \mathbb{E}(X \mid I_B)$ is common knowledge between Alice and bob, there exists an event F that is self evident to both Alice and Bob, meaning that the inequality holds for all states F. Now, as a consequence of Lemma 4.1,

$$\mathbb{E}(\mathbb{E}(X \mid I_A) - \mathbb{E}(X \mid I_B) \mid I_A \wedge I_B) = 0.$$

But $\mathbb{E}(X \mid P_A(\omega')) \leq \mathbb{E}(X \mid P_B(\omega'))$ for all $\omega' \in (P_A \wedge P_B)(\omega) \subseteq F$. Therefore, $\mathbb{E}(X \mid P_A(\omega')) = \mathbb{E}(X \mid P_B(\omega'))$ for all $\omega' \in (P_A \wedge P_B)(\omega)$. □

Corollary 4.2. It cannot be common knowledge that

$\mathbb{E}(X \mid I_A) < \mathbb{E}(X \mid I_B)$: "Alice estimates lower than Bob",

$\mathbb{E}(X \mid I_A) < \mathbb{E}(\mathbb{E}(\mathbb{E}(X \mid I_B) \mid I_A) \mid I_B)$: "Alice estimates lower than Bob's estimate of Alice's estimate of Bob's estimate',

$\vdots$

$\mathbb{E}(X \mid I_A) < \mathbb{E}(\mathbb{E}(X \mid I_B) \mid I_A)$: "Alice estimates Bob's estimate to be lower",

$\mathbb{E}(X \mid I_A) < \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbb{E}(X \mid I_B) \mid I_A) \mid I_B) \mid I_A)$: "Alice estimates Bob's estimate of Alice's estimate of Bob's estimate to be lower",

$\vdots$

Proof. The first statement directly contradicts Theorem 4.2. The proofs of the remaining statements are analogous to that of Theorem 4.2, noting by repeated application of Lemma 4.1,

$$\begin{aligned}\mathbb{E}(\mathbb{E}(X \mid I_A) \mid I_A \wedge I_B) &= \mathbb{E}(\mathbb{E}(X \mid I_B) \mid I_A \wedge I_B) \\ &= \mathbb{E}(\mathbb{E}(\mathbb{E}(X \mid I_A) \mid I_B) \mid I_A \wedge I_B) \\ &= \mathbb{E}(\mathbb{E}(\mathbb{E}(\mathbb{E}(X \mid I_B) \mid I_A) \mid I_B) \mid I_A \wedge I_B) \\ &\vdots\end{aligned}$$

□

For a random variable X, let the *outcome event* $O_X(x)$ be the set of states ω such that $X(\omega) = x$, and let X's *value partition* I_X be the collection of all outcome events. We say that Alice is *informed* of X if her information partition I_A is a refinement of X's value partition I_X, the coarsest such refinement being $I_A \vee I_X$. Clearly, if Alice and Bob are informed of X, then X's outcome events are self evident to both Alice and Bob.

Theorem 4.3 (Hanson (2002))**.** If Alice informs Bob that her estimate is lower than her estimate of future Bob's estimate, then her estimate must equal her estimate of future Bob's estimate, and this is common knowledge.

Proof. Let L be the random variable that is T if Alice's estimate is lower than her estimate of Bob's estimate and F otherwise. By assumption, Bob B is informed of L, which means that future Bob B' is informed of L. Therefore, the events $O_L(\mathsf{T})$ and $O_L(\mathsf{F})$ are self evident to both Alice and future Bob.

Let ω be the true state of the world. If $L(\omega) = \mathsf{F}$, then $\mathbb{E}(X \mid I_A) > \mathbb{E}(\mathbb{E}(X \mid I_{B'}) \mid I_A)$ for all states in $O_L(\mathsf{F})$, which event is self evident to both Alice and future Bob. Hence, this inequality is common knowledge, which by Corollary 4.2 cannot be, so $L(\omega) = \mathsf{T}$. In that case, $\mathbb{E}(X \mid I_A) \leq \mathbb{E}(\mathbb{E}(X \mid I_{B'}) \mid I_A)$ for all states in $O_L(\mathsf{T})$, which event again is self evident to both Alice and future Bob. Therefore, this inequality is common knowledge, which by Corollary 4.2 means that $\mathbb{E}(X \mid I_A) = \mathbb{E}(\mathbb{E}(X \mid I_{B'}) \mid I_A)$, and this is common knowledge. □

Corollary 4.3. Alice cannot anticipate the direction of Bob's disagreement.

So far, we have not described a protocol by which Alice and Bob can come to common knowledge. If Alice and Bob repeatedly announce their current expectations, then Aaronson (2004) shows that to agree within accuracy ε with probability at least $1 - \delta$, order $1/(\delta\varepsilon^2)$ messages suffice. Even if Alice and Bob send 2 bit summaries of their expectations, order $1/(\delta\varepsilon^2)$ messages still suffice . It turns out

that the presented theory is remarkably robust to generalisations. For example, even if Alice and Bob are computationally limited, analogous results still hold . Section IV of presents a comprehensive overview of relaxations of the basic theory.

5 Conclusion

We have seen that two rational, Bayesian agents who share a common prior cannot agree to disagree, which result turns out to be remarkably robust to generalisations. But that is funny, because in practice, oftentimes arguments turn into even more heated discussions, with no hope of agreement in the foreseeable future. And if you estimate a car to be, say, 2 years old, and I estimate that same car to be 10 years old, then I can pretty reliably predict that your updated estimate will be between 2 and 10 years old, whereas no such anticipation should be possible. So we should take that as evidence to question the assumptions underlying the results in Section 4. Firstly, in Section 3 we determined that the state-space model of knowledge might not be sound, because it completely disregards the intensional dimension of events. Secondly, it is not clear whether CPA holds: it presents difficulties with transtemporal identity, and it begs the question of what determines the common prior . One might argue that prior differences due to self-favouring priors, such as self-deceit and indifference to truth , are observed in practice, suggesting that oftentimes disagreements are simply dishonest.

References

Aaronson, S. (2004). The complexity of agreement. *arXiv preprint arXiv:cs/0406061*. eprint: https://arxiv.org/abs/cs/0406061. (Cit. on p. 6)

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics*, *4*(6), 1236–1239. (Cit. on pp. 1, 5).

Hanson, R. (2002). Disagreement is unpredictable. *Economics Letters*, *77*(3), 365–369. (Cit. on p. 6).