Agreement Through Calibrated Conversation

When Human and AI Predictions Can Correct Each Other

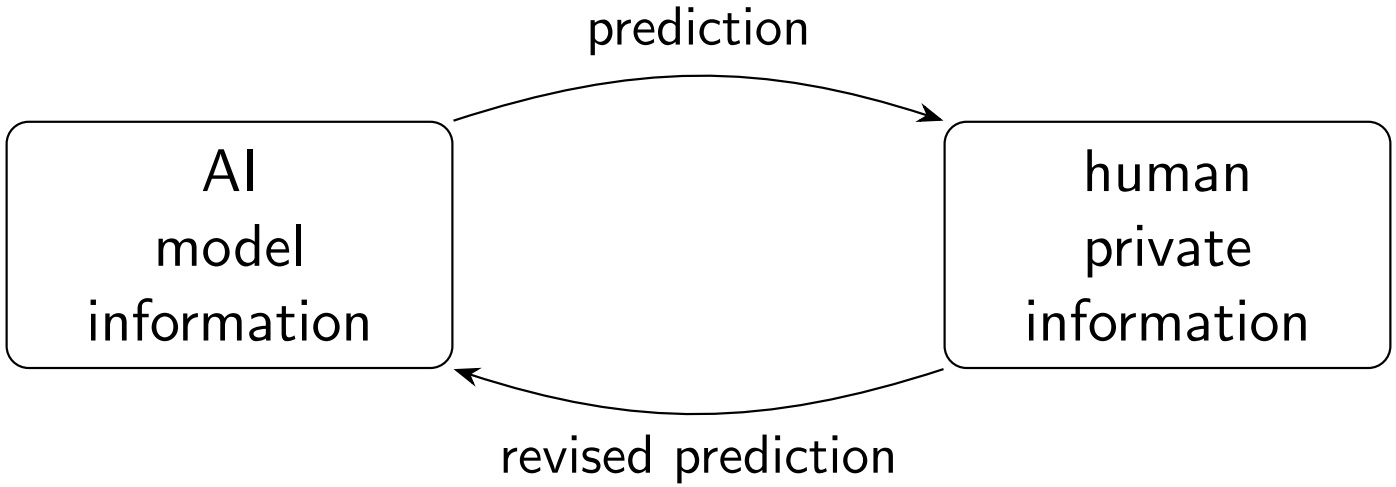
Aaron Roth

University of Pennsylvania

Fall 2026

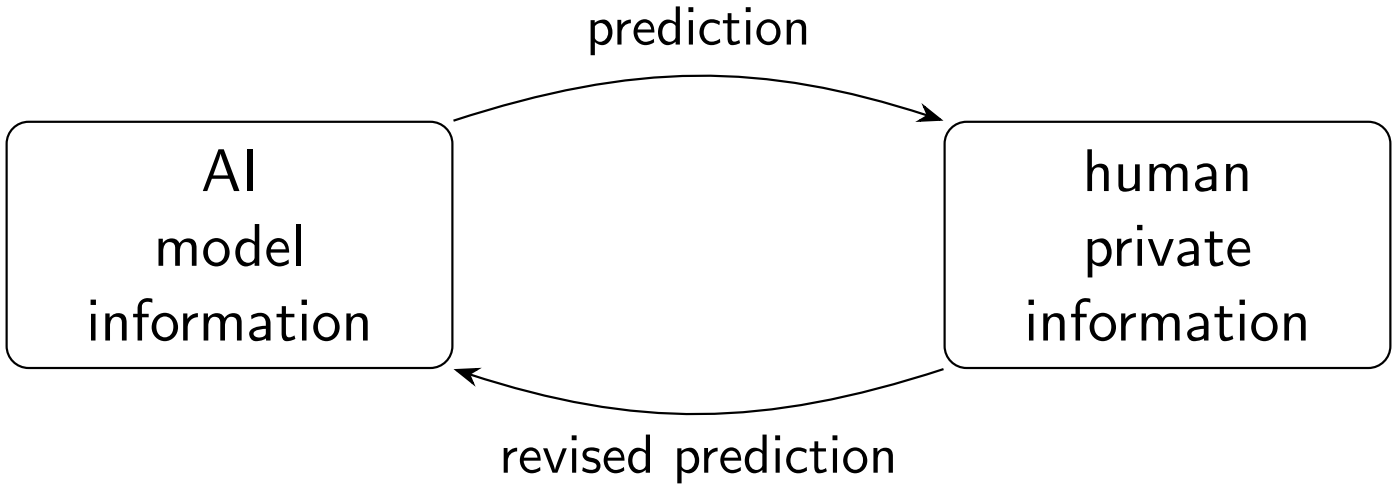
The Human Also Has Information

So far, a forecast has usually traveled from an AI system to a user. But the user may know something the AI does not.



The Human Also Has Information

So far, a forecast has usually traveled from an AI system to a user. But the user may know something the AI does not.



Both want an accurate probability for the same event, such as a safety failure. Neither party's initial estimate should automatically settle the question.

The alignment question

Lecture 7 made a calibrated forecast a one-way decision interface. Can we make the interface run in both directions such that exchanging predictions resolves disagreement and improves accuracy?

Overview

- Bayesian agents can resolve disagreement by exchanging their current predictions. Aumann's theorem describes where this process ends.

Overview

- Bayesian agents can resolve disagreement by exchanging their current predictions. Aumann's theorem describes where this process ends.
- We seek fast approximate agreement. Conversation calibration retains a learnable consequence of Bayesian rationality without requiring exact Bayesian posteriors.

Overview

- Bayesian agents can resolve disagreement by exchanging their current predictions. Aumann's theorem describes where this process ends.
- We seek fast approximate agreement. Conversation calibration retains a learnable consequence of Bayesian rationality without requiring exact Bayesian posteriors.
- Calibrated replies preserve the openings' aggregate accuracy. Continued disagreement yields further improvement, up to calibration error—so sustained disagreements are rare.

Overview

- Bayesian agents can resolve disagreement by exchanging their current predictions. Aumann's theorem describes where this process ends.
- We seek fast approximate agreement. Conversation calibration retains a learnable consequence of Bayesian rationality without requiring exact Bayesian posteriors.
- Calibrated replies preserve the openings' aggregate accuracy. Continued disagreement yields further improvement, up to calibration error—so sustained disagreements are rare.
- Approachability enforces approximate conversation calibration from repeated outcome feedback.

Private Information and Bayesian Probabilities

Let Ω be a finite set of possible states of the world. A *common prior* μ assigns each state positive probability and is shared by the human and AI.

Private Information and Bayesian Probabilities

Let Ω be a finite set of possible states of the world. A *common prior* μ assigns each state positive probability and is shared by the human and AI.
An *information partition* $\mathcal{I}_i$ divides Ω into sets that agent $i \in \{\mathrm{A}, \mathrm{H}\}$ cannot distinguish. At state ω, the agent observes the cell $\mathcal{I}_i(\omega)$.

Private Information and Bayesian Probabilities

Let Ω be a finite set of possible states of the world. A *common prior* μ assigns each state positive probability and is shared by the human and AI.
An *information partition* $\mathcal{I}_i$ divides Ω into sets that agent $i \in \{\mathrm{A}, \mathrm{H}\}$ cannot distinguish. At state ω, the agent observes the cell $\mathcal{I}_i(\omega)$.
Fix an event $E \subseteq \Omega$. Its *posterior probability* is

$$p_i(\omega) := \mu(E \mid \mathcal{I}_i(\omega)) = \frac{\mu(E \cap \mathcal{I}_i(\omega))}{\mu(\mathcal{I}_i(\omega))}.$$

Here E stays fixed: changing ω changes the information available to the agent, not the event whose probability we ask for.

Different Evidence Can Justify Different Beliefs

Two bits $X_{\mathrm{A}}, X_{\mathrm{H}}$ are independent and uniform. The AI sees X_{A}, the human sees X_{H}, and the event is $E = \{X_{\mathrm{A}} = X_{\mathrm{H}} = 1\}$.

state $(X_{\mathrm{A}}, X_{\mathrm{H}})$	prior	$\mathbf{1}_E$	AI posterior	human posterior
$(1,1)$	$1/4$	1	$1/2$	$1/2$
$(1,0)$	$1/4$	0	$1/2$	0
$(0,1)$	$1/4$	0	0	$1/2$
$(0,0)$	$1/4$	0	0	0

Different Evidence Can Justify Different Beliefs

Two bits $X_{\mathrm{A}}, X_{\mathrm{H}}$ are independent and uniform. The AI sees X_{A}, the human sees X_{H}, and the event is $E = \{X_{\mathrm{A}} = X_{\mathrm{H}} = 1\}$.

state $(X_{\mathrm{A}}, X_{\mathrm{H}})$	prior	$\mathbf{1}_E$	AI posterior	human posterior
$(1,1)$	$1/4$	1	$1/2$	$1/2$
$(1,0)$	$1/4$	0	$1/2$	0
$(0,1)$	$1/4$	0	0	$1/2$
$(0,0)$	$1/4$	0	0	0

At state $(1,0)$, both agents reason correctly but initially disagree. When the human reports probability zero, the AI learns that $X_{\mathrm{H}} = 0$ and updates its probability to zero.

Aumann's theorem concerns probabilities that are *common knowledge*, not arbitrary probabilities computed from separate evidence.

Knowledge and Common Knowledge

Agent i *knows* an event F at state ω if every state it considers possible belongs to F:

$$\mathcal{I}_i(\omega) \subseteq F.$$

Knowledge and Common Knowledge

Agent i *knows* an event F at state ω if every state it considers possible belongs to F:

$$\mathcal{I}_i(\omega) \subseteq F.$$

Common knowledge requires more than both agents knowing F: each knows it, each knows that the other knows it, and so on.

Knowledge and Common Knowledge

Agent i *knows* an event F at state ω if every state it considers possible belongs to F:

$$\mathcal{I}_i(\omega) \subseteq F.$$

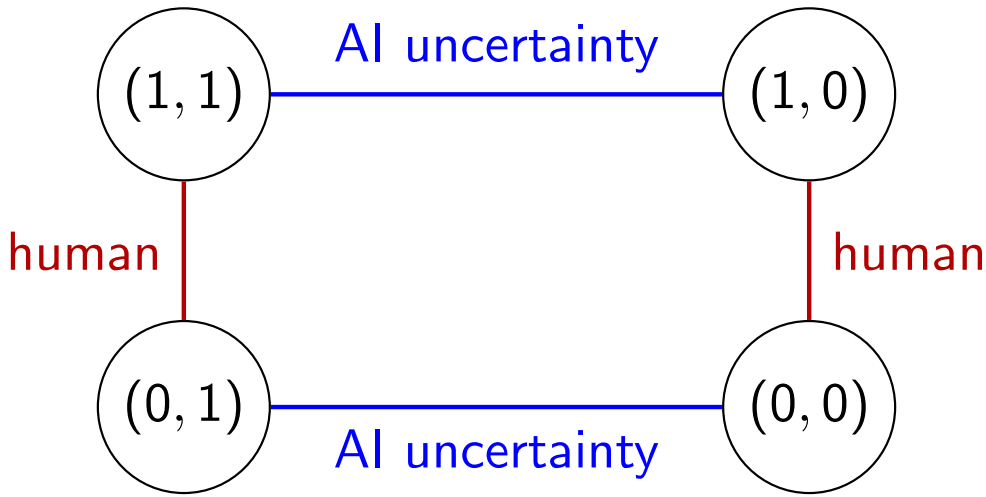
Common knowledge requires more than both agents knowing F: each knows it, each knows that the other knows it, and so on. There is a finite way to express all these requirements. Connect two states whenever either agent cannot distinguish them. Let $D(\omega)$ be the connected component containing ω. Then F is common knowledge at ω precisely when

$$\boxed{D(\omega) \subseteq F.}$$

Why: "H knows A knows F" says every $\mathcal{I}_{\mathrm{A}}$-cell touching $\mathcal{I}_{\mathrm{H}}(\omega)$ lies in F. Each further level alternates partitions once more, and the levels together sweep out exactly the connected component.

Common Knowledge Joins Whole Information Cells

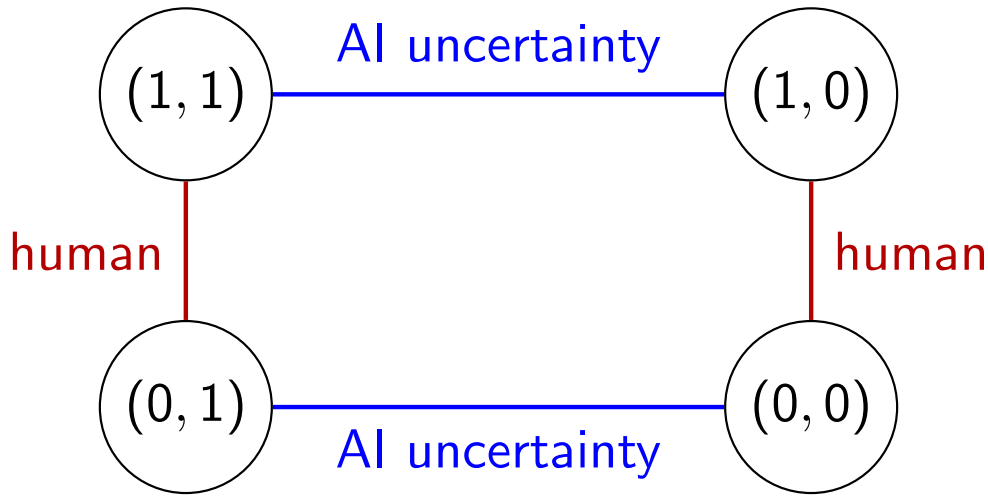
In the two-bit example, the AI cannot distinguish states in the same row, and the human cannot distinguish states in the same column.



Every state is reachable, so here $D(\omega) = \Omega$.

Common Knowledge Joins Whole Information Cells

In the two-bit example, the AI cannot distinguish states in the same row, and the human cannot distinguish states in the same column.



Every state is reachable, so here $D(\omega) = \Omega$.
In general, if one state of an information cell belongs to $D(\omega)$, every state of that cell does. Thus $D(\omega)$ is a union of whole cells of *either* agent's partition.

Aumann's Agreement Theorem

Theorem (Aumann, 1976)

Two Bayesian agents share a finite, full-support common prior μ. If at state ω it is common knowledge that their posterior probabilities of the same event E are

$$p_{\mathrm{A}} = \alpha \qquad \textit{and} \qquad p_{\mathrm{H}} = \beta,$$

then $\alpha = \beta$.

Aumann's Agreement Theorem

Theorem (Aumann, 1976)

Two Bayesian agents share a finite, full-support common prior μ. If at state ω it is common knowledge that their posterior probabilities of the same event E are

$$p_{\mathrm{A}} = \alpha \qquad \textit{and} \qquad p_{\mathrm{H}} = \beta,$$

then $\alpha = \beta$.

Recall what the state argument means:

$$p_i(\omega') = \mu(E \mid \mathcal{I}_i(\omega')).$$

This is agent i's probability of the *fixed event* E when the state is ω', not its probability of the state ω' itself.

Which Event Is Common Knowledge?

The event assumed to be common knowledge is

$$F := \{\omega' \in \Omega : p_{\mathrm{A}}(\omega') = \alpha, \quad p_{\mathrm{H}}(\omega') = \beta\}.$$

Which Event Is Common Knowledge?

The event assumed to be common knowledge is

$$F := \{\omega' \in \Omega : p_{\mathrm{A}}(\omega') = \alpha, \quad p_{\mathrm{H}}(\omega') = \beta\}.$$

Write $D = D(\omega)$. By the definition of common knowledge,

$$D \subseteq F.$$

Membership in F therefore gives

$$p_{\mathrm{A}}(\omega') = \alpha, \qquad p_{\mathrm{H}}(\omega') = \beta \quad \text{for every } \omega' \in D.$$

Which Event Is Common Knowledge?

The event assumed to be common knowledge is

$$F := \{\omega' \in \Omega : p_{\mathrm{A}}(\omega') = \alpha, \quad p_{\mathrm{H}}(\omega') = \beta\}.$$

Write $D = D(\omega)$. By the definition of common knowledge,

$$D \subseteq F.$$

Membership in F therefore gives

$$p_{\mathrm{A}}(\omega') = \alpha, \qquad p_{\mathrm{H}}(\omega') = \beta \quad \text{for every } \omega' \in D.$$

The common-knowledge claim concerns the agents' probabilities of E, not whether E occurred. The cell D may contain states both inside and outside E.

Both Probabilities Average to the Same Number

The common-knowledge cell D is a union of AI information cells.
On each such cell $I \subseteq D$, the AI's posterior is $\mu(E \mid I) = \alpha$.

Both Probabilities Average to the Same Number

The common-knowledge cell D is a union of AI information cells.
On each such cell $I \subseteq D$, the AI's posterior is $\mu(E \mid I) = \alpha$.
The law of total probability therefore gives

$$\begin{aligned}
\mu(E \mid D) &= \sum_{I \in \mathcal{I}_{\mathrm{A}} \colon I \subseteq D} \mu(I \mid D)\mu(E \mid I) \\
&= \sum_{I \in \mathcal{I}_{\mathrm{A}} \colon I \subseteq D} \mu(I \mid D)\alpha \\
&= \alpha.
\end{aligned}$$

Both Probabilities Average to the Same Number

The common-knowledge cell D is a union of AI information cells.
On each such cell $I \subseteq D$, the AI's posterior is $\mu(E \mid I) = \alpha$.
The law of total probability therefore gives

$$\begin{aligned}\mu(E \mid D) &= \sum_{I \in \mathcal{I}_{\mathrm{A}} \colon I \subseteq D} \mu(I \mid D)\mu(E \mid I) \\ &= \sum_{I \in \mathcal{I}_{\mathrm{A}} \colon I \subseteq D} \mu(I \mid D)\alpha \\ &= \alpha.\end{aligned}$$

The same calculation using the human's cells gives $\mu(E \mid D) = \beta$.
Hence $\alpha = \beta$. □

Can Conversation Create Common Knowledge?

Aumann assumes that the agents' posterior values are common knowledge. How could they get there without revealing all their private information?

Can Conversation Create Common Knowledge?

Aumann assumes that the agents' posterior values are common knowledge. How could they get there without revealing all their private information?

An alternating conversation about the same event E

1. The AI announces its current posterior probability of E.
2. The human updates using its private information and the AI's message, then announces its new posterior.
3. The AI updates and replies; the parties continue alternating.

Can Conversation Create Common Knowledge?

Aumann assumes that the agents' posterior values are common knowledge. How could they get there without revealing all their private information?

An alternating conversation about the same event E

1. The AI announces its current posterior probability of E.
2. The human updates using its private information and the AI's message, then announces its new posterior.
3. The AI updates and replies; the parties continue alternating.

Hearing a report can change your posterior, so announcing beliefs once need not settle the matter. Can continued exchange make the *current* posterior values common knowledge?

The Transcript Determines the Remaining States

Let $\Omega_j \subseteq \Omega$ be the states consistent with the first j public announcements, starting from $\Omega_0 = \Omega$. The current posterior of agent i at $\omega \in \Omega_j$ is

$$p_i^{(j)}(\omega) := \mu(E \mid \mathcal{I}_i(\omega) \cap \Omega_j).$$

The Transcript Determines the Remaining States

Let $\Omega_j \subseteq \Omega$ be the states consistent with the first j public announcements, starting from $\Omega_0 = \Omega$. The current posterior of agent i at $\omega \in \Omega_j$ is

$$p_i^{(j)}(\omega) := \mu(E \mid \mathcal{I}_i(\omega) \cap \Omega_j).$$

If the agent announces v, this set is refined:

$$\Omega_{j+1} = \{\omega' \in \Omega_j : p_i^{(j)}(\omega') = v\} \subseteq \Omega_j.$$

The same transcript narrows the possibilities for both agents, while each still has its own private information cell.

At Stabilization, Aumann Forces Agreement

The public sets form a decreasing sequence

$$\Omega_0 \supseteq \Omega_1 \supseteq \Omega_2 \supseteq \cdots .$$

These finite sets always contain the actual state, so the decreasing sequence eventually stabilizes at a nonempty set Ω_*.

At Stabilization, Aumann Forces Agreement

The public sets form a decreasing sequence

$$\Omega_0 \supseteq \Omega_1 \supseteq \Omega_2 \supseteq \cdots .$$

These finite sets always contain the actual state, so the decreasing sequence eventually stabilizes at a nonempty set Ω_*.
From then on, each agent's announcement removes no state. So for some public values α, β,

$$p_{\mathrm{A}}^{(*)}(\omega') = \alpha, \qquad p_{\mathrm{H}}^{(*)}(\omega') = \beta \quad \text{for every } \omega' \in \Omega_*.$$

At Stabilization, Aumann Forces Agreement

The public sets form a decreasing sequence

$$\Omega_0 \supseteq \Omega_1 \supseteq \Omega_2 \supseteq \cdots .$$

These finite sets always contain the actual state, so the decreasing sequence eventually stabilizes at a nonempty set Ω_*.
From then on, each agent's announcement removes no state. So for some public values α, β,

$$p_{\mathrm{A}}^{(*)}(\omega') = \alpha, \qquad p_{\mathrm{H}}^{(*)}(\omega') = \beta \quad \text{for every } \omega' \in \Omega_*.$$

After the public announcements, the state space is Ω_*, the common prior is $\mu(\,\cdot \mid \Omega_*)$, and the information cells are $\mathcal{I}_i(\omega) \cap \Omega_*$. Both posterior values hold throughout Ω_*, hence throughout every common-knowledge cell of the updated model. They are *common knowledge*, so Aumann's theorem gives

$$\boxed{\alpha = \beta.}$$

We Want Fast Approximate Agreement

Finite convergence need not mean a short conversation. There can be up to $|\Omega| - 1$ strict refinements of the public state set, and $|\Omega|$ may be enormous. Computing each exact posterior can also require expensive inference; the agreement theorem additionally assumes a common prior.

We Want Fast Approximate Agreement

Finite convergence need not mean a short conversation. There can be up to $|\Omega| - 1$ strict refinements of the public state set, and $|\Omega|$ may be enormous. Computing each exact posterior can also require expensive inference; the agreement theorem additionally assumes a common prior.

The quantitative goal

For a tolerance $\varepsilon > 0$, can consecutive reports become ε-close after few messages on most conversations, while preserving accuracy—with a round bound that does not depend on $|\Omega|$?

We Want Fast Approximate Agreement

Finite convergence need not mean a short conversation. There can be up to $|\Omega| - 1$ strict refinements of the public state set, and $|\Omega|$ may be enormous. Computing each exact posterior can also require expensive inference; the agreement theorem additionally assumes a common prior.

The quantitative goal

For a tolerance $\varepsilon > 0$, can consecutive reports become ε-close after few messages on most conversations, while preserving accuracy—with a round bound that does not depend on $|\Omega|$?

We will keep the alternating communication pattern and replace exact Bayesian updating with a calibration condition learned from outcome feedback: a *tractable relaxation of Bayesian rationality* that does not require a common prior.

Repeated Conversations with Outcome Feedback

We now repeat the interaction on days $t = 1, \ldots, T$. An arbitrary binary outcome $y_t \in \{0, 1\}$ is fixed before each conversation; each party has private information about it. Reports come from a finite grid $\mathcal{P} \subset [0, 1]$ of size $M = |\mathcal{P}|$.

Repeated Conversations with Outcome Feedback

We now repeat the interaction on days $t = 1, \ldots, T$. An arbitrary binary outcome $y_t \in \{0, 1\}$ is fixed before each conversation; each party has private information about it. Reports come from a finite grid $\mathcal{P} \subset [0, 1]$ of size $M = |\mathcal{P}|$.

Openings. In the learned protocol, allow two unrestricted opening reports: the AI states p_t^1 and the human states p_t^2, each from its own information.

Replies. From round 3 on, keep alternating: the AI sends $p_t^3, p_t^5, \ldots$ and the human sends $p_t^4, p_t^6, \ldots$.

Repeated Conversations with Outcome Feedback

We now repeat the interaction on days $t = 1, \ldots, T$. An arbitrary binary outcome $y_t \in \{0, 1\}$ is fixed before each conversation; each party has private information about it. Reports come from a finite grid $\mathcal{P} \subset [0, 1]$ of size $M = |\mathcal{P}|$.

Openings. In the learned protocol, allow two unrestricted opening reports: the AI states p_t^1 and the human states p_t^2, each from its own information.

Replies. From round 3 on, keep alternating: the AI sends $p_t^3, p_t^5, \ldots$ and the human sends $p_t^4, p_t^6, \ldots$.

Fix a tolerance $\varepsilon \in (0, 1)$ and a cap of $R \geq 3$ messages. Stop at the first reply round $k \geq 3$ with $|p_t^k - p_t^{k-1}| \leq \varepsilon$; if agreement has not occurred, stop at the cap. Then reveal y_t, which can inform future days.

Repeated Conversations with Outcome Feedback

We now repeat the interaction on days $t = 1, \ldots, T$. An arbitrary binary outcome $y_t \in \{0, 1\}$ is fixed before each conversation; each party has private information about it. Reports come from a finite grid $\mathcal{P} \subset [0, 1]$ of size $M = |\mathcal{P}|$.

Openings. In the learned protocol, allow two unrestricted opening reports: the AI states p_t^1 and the human states p_t^2, each from its own information.

Replies. From round 3 on, keep alternating: the AI sends $p_t^3, p_t^5, \ldots$ and the human sends $p_t^4, p_t^6, \ldots$.

Fix a tolerance $\varepsilon \in (0, 1)$ and a cap of $R \geq 3$ messages. Stop at the first reply round $k \geq 3$ with $|p_t^k - p_t^{k-1}| \leq \varepsilon$; if agreement has not occurred, stop at the cap. Then reveal y_t, which can inform future days.

We no longer assume a common prior or exact Bayesian updates. We will instead impose calibration conditions on the *replies* of both speakers.

Ordinary Calibration Is Not Enough

Suppose half the days have each outcome. The human sees y_t exactly and reports it. The AI always reports $1/2$, ignoring every human message.

fraction of days	outcome	AI reports	human reports
$1/2$	0	$1/2$	0
$1/2$	1	$1/2$	1

Ordinary Calibration Is Not Enough

Suppose half the days have each outcome. The human sees y_t exactly and reports it. The AI always reports $1/2$, ignoring every human message.

fraction of days	outcome	AI reports	human reports
$1/2$	0	$1/2$	0
$1/2$	1	$1/2$	1

Both speakers are calibrated conditional on their *own current report*: the AI's $1/2$ cell has outcome mean $1/2$, and the human is always correct. Both repeat their own forecasts forever.

Ordinary Calibration Is Not Enough

Suppose half the days have each outcome. The human sees y_t exactly and reports it. The AI always reports $1/2$, ignoring every human message.

fraction of days	outcome	AI reports	human reports
$1/2$	0	$1/2$	0
$1/2$	1	$1/2$	1

Both speakers are calibrated conditional on their *own current report*: the AI's $1/2$ cell has outcome mean $1/2$, and the human is always correct. Both repeat their own forecasts forever.
Every consecutive pair of reports differs by $1/2$. For $\varepsilon < 1/2$, no conversation reaches agreement, however large the cap.
Calibration alone has not required the AI to use the message it just received.

The Previous Message Exposes the Ignored Information

Look at the AI's reply in the same example:

previous human report q	current AI report p	outcome mean
0	$1/2$	0
1	$1/2$	1

The Previous Message Exposes the Ignored Information

Look at the AI's reply in the same example:

previous human report q	current AI report p	outcome mean
0	1/2	0
1	1/2	1

Pooling the two rows makes the AI look calibrated. Within either row, its report is wrong on average.
We will require calibration conditional on *both* the current report and the single most recent previous message.

The Previous Message Exposes the Ignored Information

Look at the AI's reply in the same example:

previous human report q	current AI report p	outcome mean
0	1/2	0
1	1/2	1

Pooling the two rows makes the AI look calibrated. Within either row, its report is wrong on average.

We will require calibration conditional on *both* the current report and the single most recent previous message.

This is context-dependent calibration, as in Lecture 9: the previous message supplies the context.

Which Conversations Reach k Rounds?

Let S_k be the set of days in which conversation reaches message k. Every conversation involves both openings and the first reply, so $S_1 = S_2 = S_3 = \{1, \ldots, T\}$. For $3 \leq k \leq R$,

$$S_{k+1} = \{t \in S_k : |p_t^k - p_t^{k-1}| > \varepsilon\}.$$

Thus $S_{k+1} \subseteq S_k$.

Which Conversations Reach k Rounds?

Let S_k be the set of days in which conversation reaches message k. Every conversation involves both openings and the first reply, so $S_1 = S_2 = S_3 = \{1, \ldots, T\}$. For $3 \leq k \leq R$,

$$S_{k+1} = \{t \in S_k : |p_t^k - p_t^{k-1}| > \varepsilon\}.$$

Thus $S_{k+1} \subseteq S_k$.
Recall that we allow at most R messages per conversation. The fraction that have not reached agreement at this cap is

$$\frac{|S_{R+1}|}{T}.$$

Here S_{R+1} records unresolved days, not an additional message.

Calibrate Each Reply Given the Previous Report

At reply round $k \geq 3$, group days by the previous report q and reply p:

$$C_{k,q,p} := \{t \in S_k : p_t^{k-1} = q,\ p_t^k = p\}.$$

The accumulated prediction error within this cell is

$$r_{k,q,p} := \sum_{t \in C_{k,q,p}} (y_t - p).$$

Calibrate Each Reply Given the Previous Report

At reply round $k \geq 3$, group days by the previous report q and reply p:

$$C_{k,q,p} := \{t \in S_k : p_t^{k-1} = q,\ p_t^k = p\}.$$

The accumulated prediction error within this cell is

$$r_{k,q,p} := \sum_{t \in C_{k,q,p}} (y_t - p).$$

Exact calibration requires $r_{k,q,p} = 0$ for every pair. On each nonempty cell, this says that the outcomes average to the reply p.

Calibrate Each Reply Given the Previous Report

At reply round $k \geq 3$, group days by the previous report q and reply p:

$$C_{k,q,p} := \{t \in S_k : p_t^{k-1} = q,\ p_t^k = p\}.$$

The accumulated prediction error within this cell is

$$r_{k,q,p} := \sum_{t \in C_{k,q,p}} (y_t - p).$$

Exact calibration requires $r_{k,q,p} = 0$ for every pair. On each nonempty cell, this says that the outcomes average to the reply p. These cells include all active replies, whether or not the reply ends the conversation. The previous report and current reply are both fixed within each cell.
First we will prove fast agreement under this exact condition.
Later we will allow small residual errors.

Conversation Calibration Forces Fast Agreement

Theorem

Suppose $r_{k,q,p} = 0$ *for every report pair at every reply round* $k = 3, \ldots, R$. *Then*

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2}.}$$

Conversation Calibration Forces Fast Agreement

Theorem

Suppose $r_{k,q,p} = 0$ *for every report pair at every reply round* $k = 3, \ldots, R$. *Then*

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2}.}$$

For any $\delta \in (0,1)$, at least a $1-\delta$ fraction of conversations reach ε-agreement within

$$R = 2 + \left\lceil \frac{1}{\delta\varepsilon^2} \right\rceil$$

messages: two openings followed by the indicated replies.

Conversation Calibration Forces Fast Agreement

Theorem

Suppose $r_{k,q,p} = 0$ *for every report pair at every reply round* $k = 3, \ldots, R$. *Then*

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2}.}$$

For any $\delta \in (0,1)$, at least a $1-\delta$ fraction of conversations reach ε-agreement within

$$R = 2 + \left\lceil \frac{1}{\delta\varepsilon^2} \right\rceil$$

messages: two openings followed by the indicated replies.

Proof idea: substantial disagreement forces an improvement in total squared prediction error across days. There is only a bounded amount of error to remove.

One Bounded Error Measure Tracks All the Conversations

For analysis, $\bar{p}_t^k = p_t^k$ if day t reaches round k. If its conversation ends at round h, keep using that final report:

$$\bar{p}_t^k = p_t^h \qquad \text{for every } k > h.$$

No further messages are sent; this only extends the notation beyond stopping.

One Bounded Error Measure Tracks All the Conversations

For analysis, $\bar{p}_t^k = p_t^k$ if day t reaches round k. If its conversation ends at round h, keep using that final report:

$$\bar{p}_t^k = p_t^h \qquad \text{for every } k > h.$$

No further messages are sent; this only extends the notation beyond stopping.
Our measure of prediction error is

$$\mathcal{E}_k := \sum_{t=1}^{T} (y_t - \bar{p}_t^k)^2, \qquad 0 \le \mathcal{E}_k \le T.$$

For a binary outcome, $(y - p)^2$ is also called the *Brier loss*. So $\mathcal{E}_1$ and $\mathcal{E}_2$ are the total errors of the two openings.

One Bounded Error Measure Tracks All the Conversations

For analysis, $\bar{p}_t^k = p_t^k$ if day t reaches round k. If its conversation ends at round h, keep using that final report:

$$\bar{p}_t^k = p_t^h \qquad \text{for every } k > h.$$

No further messages are sent; this only extends the notation beyond stopping.
Our measure of prediction error is

$$\mathcal{E}_k := \sum_{t=1}^{T} (y_t - \bar{p}_t^k)^2, \qquad 0 \le \mathcal{E}_k \le T.$$

For a binary outcome, $(y - p)^2$ is also called the *Brier loss*. So $\mathcal{E}_1$ and $\mathcal{E}_2$ are the total errors of the two openings.
A stopped conversation contributes the same squared error at every later round. Thus only days in S_k can change $\mathcal{E}_{k-1}$ to $\mathcal{E}_k$.

A Squared-Error Identity Measures the Gain from a Reply

On one active day, write f for a report we compare against, p for the reply, and y for the outcome. Since $y - f = (y - p) + (p - f)$,

$$\begin{aligned}(y-f)^2 - (y-p)^2 &= \big((y-p)+(p-f)\big)^2 - (y-p)^2 \\ &= (p-f)^2 + 2(y-p)(p-f).\end{aligned}$$

A Squared-Error Identity Measures the Gain from a Reply

On one active day, write f for a report we compare against, p for the reply, and y for the outcome. Since $y - f = (y - p) + (p - f)$,

$$\begin{aligned}(y-f)^2 - (y-p)^2 &= \big((y-p)+(p-f)\big)^2 - (y-p)^2 \\ &= (p-f)^2 + 2(y-p)(p-f).\end{aligned}$$

The first term is the squared change from f to the reply. The second can have either sign on an individual day.
To turn a large change into an accuracy improvement, we must control that second term after averaging over days.

Exact Calibration Cancels the Cross Term

Compare the reply with the previous report: $f_t = p_t^{k-1}$. On the cell $C_{k,q,p}$, the previous report is q and the reply is p:

$$\sum_{t \in C_{k,q,p}} 2(y_t - p)(p - q) = 2(p - q) \underbrace{\sum_{t \in C_{k,q,p}} (y_t - p)}_{r_{k,q,p}}.$$

Exact Calibration Cancels the Cross Term

Compare the reply with the previous report: $f_t = p_t^{k-1}$. On the cell $C_{k,q,p}$, the previous report is q and the reply is p:

$$\sum_{t \in C_{k,q,p}} 2(y_t - p)(p - q) = 2(p - q) \underbrace{\sum_{t \in C_{k,q,p}} (y_t - p)}_{r_{k,q,p}}.$$

Exact calibration says $r_{k,q,p} = 0$. Summing over cells therefore gives

$$\sum_{t \in S_k} 2(y_t - p_t^k)(p_t^k - p_t^{k-1}) = 0.$$

Exact Calibration Cancels the Cross Term

Compare the reply with the previous report: $f_t = p_t^{k-1}$. On the cell $C_{k,q,p}$, the previous report is q and the reply is p:

$$\sum_{t \in C_{k,q,p}} 2(y_t - p)(p - q) = 2(p - q) \underbrace{\sum_{t \in C_{k,q,p}} (y_t - p)}_{r_{k,q,p}}.$$

Exact calibration says $r_{k,q,p} = 0$. Summing over cells therefore gives

$$\sum_{t \in S_k} 2(y_t - p_t^k)(p_t^k - p_t^{k-1}) = 0.$$

This is why we conditioned on both reports: their difference is constant within a cell and can be taken outside the residual sum. We can now turn the squared-error identity into a progress guarantee.

Unresolved Conversations Force Accuracy Progress

Recall $\mathcal{E}_k = \sum_t (y_t - \bar{p}_t^k)^2$. Only active days change their report. Sum the squared-error identity over S_k; exact calibration cancels its cross term:

$$\mathcal{E}_{k-1} - \mathcal{E}_k = \sum_{t \in S_k} (p_t^k - p_t^{k-1})^2 \qquad (k \geq 3).$$

Unresolved Conversations Force Accuracy Progress

Recall $\mathcal{E}_k = \sum_t (y_t - \bar{p}_t^k)^2$. Only active days change their report. Sum the squared-error identity over S_k; exact calibration cancels its cross term:

$$\mathcal{E}_{k-1} - \mathcal{E}_k = \sum_{t \in S_k} (p_t^k - p_t^{k-1})^2 \qquad (k \geq 3).$$

Every day in S_{k+1} still has $|p_t^k - p_t^{k-1}| > \varepsilon$. All the other squared changes are nonnegative, so

$$\boxed{\mathcal{E}_{k-1} - \mathcal{E}_k \geq \varepsilon^2 |S_{k+1}| \qquad (k \geq 3).}$$

Unresolved Conversations Force Accuracy Progress

Recall $\mathcal{E}_k = \sum_t (y_t - \bar{p}_t^k)^2$. Only active days change their report. Sum the squared-error identity over S_k; exact calibration cancels its cross term:

$$\mathcal{E}_{k-1} - \mathcal{E}_k = \sum_{t \in S_k} (p_t^k - p_t^{k-1})^2 \qquad (k \geq 3).$$

Every day in S_{k+1} still has $|p_t^k - p_t^{k-1}| > \varepsilon$. All the other squared changes are nonnegative, so

$$\boxed{\mathcal{E}_{k-1} - \mathcal{E}_k \geq \varepsilon^2 |S_{k+1}| \qquad (k \geq 3).}$$

Each conversation that remains unresolved contributes at least ε^2 to the decrease in aggregate error. The same argument applies at every reply, including the first.

Bounded Error Limits the Number of Messages

Sum the progress inequalities from $k = 3$ through R:

$$\varepsilon^2 \sum_{k=3}^{R} |S_{k+1}| \leq \sum_{k=3}^{R} (\mathcal{E}_{k-1} - \mathcal{E}_k) = \mathcal{E}_2 - \mathcal{E}_R \leq T.$$

Bounded Error Limits the Number of Messages

Sum the progress inequalities from $k = 3$ through R:

$$\varepsilon^2 \sum_{k=3}^{R} |S_{k+1}| \le \sum_{k=3}^{R} (\mathcal{E}_{k-1} - \mathcal{E}_k) = \mathcal{E}_2 - \mathcal{E}_R \le T.$$

Every conversation unresolved at the cap was unresolved after all $R - 2$ reply rounds. Thus

$$(R-2)|S_{R+1}| \le \sum_{k=3}^{R} |S_{k+1}|.$$

Bounded Error Limits the Number of Messages

Sum the progress inequalities from $k = 3$ through R:

$$\varepsilon^2 \sum_{k=3}^{R} |S_{k+1}| \le \sum_{k=3}^{R} (\mathcal{E}_{k-1} - \mathcal{E}_k) = \mathcal{E}_2 - \mathcal{E}_R \le T.$$

Every conversation unresolved at the cap was unresolved after all $R - 2$ reply rounds. Thus

$$(R-2)|S_{R+1}| \le \sum_{k=3}^{R} |S_{k+1}|.$$

Combining these bounds and dividing by $(R-2)\varepsilon^2 T$ gives

$$\boxed{\frac{|S_{R+1}|}{T} \le \frac{1}{(R-2)\varepsilon^2}.}$$

This proves the theorem. □

The proof also shows that accuracy improves from the human's opening onward. What about the AI's opening?

Agreement Can Discard an Accurate Opening

The progress argument starts from $\mathcal{E}_2$, the error of the human's opening. It has not compared the replies with the AI's opening.

Agreement Can Discard an Accurate Opening

The progress argument starts from $\mathcal{E}_2$, the error of the human's opening. It has not compared the replies with the AI's opening. Suppose half the days have each outcome. The AI is initially perfect, $p_t^1 = y_t$, while the human reports $p_t^2 = 1/2$. If the AI now replies $p_t^3 = 1/2$ every day, it passes our calibration check: the single $(p^2, p^3) = (1/2, 1/2)$ cell has outcome mean $1/2$.

Agreement Can Discard an Accurate Opening

The progress argument starts from $\mathcal{E}_2$, the error of the human's opening. It has not compared the replies with the AI's opening. Suppose half the days have each outcome. The AI is initially perfect, $p_t^1 = y_t$, while the human reports $p_t^2 = 1/2$. If the AI now replies $p_t^3 = 1/2$ every day, it passes our calibration check: the single $(p^2, p^3) = (1/2, 1/2)$ cell has outcome mean $1/2$.
The conversation ends immediately, but

$$\mathcal{E}_1 = 0, \qquad \mathcal{E}_2 = \mathcal{E}_3 = T/4.$$

Agreement has discarded the AI's initial information.

The missing check

The first AI reply must also be reliable conditional on its own opening.

An Extra Check Protects Both Opening Predictions

At the first reply, also require

$$s_{q,p} := \sum_{t:\, p_t^1 = q,\ p_t^3 = p} (y_t - p) = 0 \qquad \text{for every } q, p.$$

This groups days by the AI's opening and first reply. In the example, each such cell has outcome mean 0 or 1, not $1/2$.

An Extra Check Protects Both Opening Predictions

At the first reply, also require

$$s_{q,p} := \sum_{t:\, p_t^1 = q,\ p_t^3 = p} (y_t - p) = 0 \qquad \text{for every } q, p.$$

This groups days by the AI's opening and first reply. In the example, each such cell has outcome mean 0 or 1, not $1/2$. Reuse the squared-error identity, now comparing p_t^3 with p_t^1:

$$\begin{aligned} \mathcal{E}_1 - \mathcal{E}_3 &= \sum_t (p_t^3 - p_t^1)^2 + 2 \sum_{q,p} (p - q) s_{q,p} \\ &= \sum_t (p_t^3 - p_t^1)^2 \geq 0. \end{aligned}$$

An Extra Check Protects Both Opening Predictions

At the first reply, also require

$$s_{q,p} := \sum_{t:\, p_t^1 = q,\ p_t^3 = p} (y_t - p) = 0 \qquad \text{for every } q, p.$$

This groups days by the AI's opening and first reply. In the example, each such cell has outcome mean 0 or 1, not $1/2$. Reuse the squared-error identity, now comparing p_t^3 with p_t^1:

$$\begin{aligned} \mathcal{E}_1 - \mathcal{E}_3 &= \sum_t (p_t^3 - p_t^1)^2 + 2 \sum_{q,p} (p - q) s_{q,p} \\ &= \sum_t (p_t^3 - p_t^1)^2 \geq 0. \end{aligned}$$

Previous-message calibration already gave $\mathcal{E}_3 \leq \mathcal{E}_2$. Together,

$$\boxed{\mathcal{E}_3 \leq \min\{\mathcal{E}_1, \mathcal{E}_2\}.}$$

The extra check protects the starting accuracy; the progress argument is unchanged.

Further Disagreement Improves on the Better Opening

Let $\bar{N}$ be the average number of replies after the first reply that still leave the reports more than ε apart:

$$\bar{N} := \frac{1}{T} \sum_{k=4}^{R} |S_{k+1}|.$$

This excludes the two openings, the first reply, and any reply that reaches agreement; unresolved replies at the cap count.

Further Disagreement Improves on the Better Opening

Let $\bar{N}$ be the average number of replies after the first reply that still leave the reports more than ε apart:

$$\bar{N} := \frac{1}{T} \sum_{k=4}^{R} |S_{k+1}|.$$

This excludes the two openings, the first reply, and any reply that reaches agreement; unresolved replies at the cap count.
Sum the progress inequalities for $k = 4, \ldots, R$:

$$\mathcal{E}_3 - \mathcal{E}_R \geq \varepsilon^2 T \bar{N}.$$

Since $\mathcal{E}_3 \leq \min\{\mathcal{E}_1, \mathcal{E}_2\}$,

$$\boxed{\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T} \geq \varepsilon^2 \bar{N}.}$$

Approximate Calibration Makes the Guarantee Approximate

Now measure total residual error at reply round k by

$$\Lambda_k := \sum_{q,p} |r_{k,q,p}| + \mathbf{1}\{k = 3\} \sum_{q,p} |s_{q,p}|.$$

The first reply includes both checks. Absolute values prevent errors in different cells from cancelling.

Approximate Calibration Makes the Guarantee Approximate

Now measure total residual error at reply round k by

$$\Lambda_k := \sum_{q,p} |r_{k,q,p}| + \mathbf{1}\{k = 3\} \sum_{q,p} |s_{q,p}|.$$

The first reply includes both checks. Absolute values prevent errors in different cells from cancelling.

The cross term no longer vanishes. Since $|p - q| \leq 1$,

$$2 \sum_{q,p} (p - q) r_{k,q,p} \geq -2 \sum_{q,p} |r_{k,q,p}| \geq -2\Lambda_k.$$

The same bound applies to the extra first-reply check.

Approximate Calibration Makes the Guarantee Approximate

Now measure total residual error at reply round k by

$$\Lambda_k := \sum_{q,p} |r_{k,q,p}| + \mathbf{1}\{k = 3\} \sum_{q,p} |s_{q,p}|.$$

The first reply includes both checks. Absolute values prevent errors in different cells from cancelling.
The cross term no longer vanishes. Since $|p - q| \le 1$,

$$2 \sum_{q,p} (p - q) r_{k,q,p} \ge -2 \sum_{q,p} |r_{k,q,p}| \ge -2\Lambda_k.$$

The same bound applies to the extra first-reply check.
Our two progress statements become

$$\mathcal{E}_{k-1} - \mathcal{E}_k \ge \varepsilon^2 |S_{k+1}| - 2\Lambda_k \quad (k \ge 3),$$

$$\mathcal{E}_3 \le \min\{\mathcal{E}_1, \mathcal{E}_2\} + 2\Lambda_3.$$

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities just as before:

$$\boxed{\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2} + \frac{2\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon^2}.}$$

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities just as before:

$$\boxed{\frac{|S_{R+1}|}{T} \le \frac{1}{(R-2)\varepsilon^2} + \frac{2\sum_{k=3}^{R}\Lambda_k}{T(R-2)\varepsilon^2}.}$$

Protecting both openings and summing the later accuracy gains gives

$$\boxed{\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T} \ge \varepsilon^2 \bar{N} - \frac{2}{T}\sum_{k=3}^{R}\Lambda_k.}$$

The Same Proof Gives Approximate Guarantees

Sum the corrected progress inequalities just as before:

$$\boxed{\frac{|S_{R+1}|}{T} \le \frac{1}{(R-2)\varepsilon^2} + \frac{2\sum_{k=3}^{R}\Lambda_k}{T(R-2)\varepsilon^2}.}$$

Protecting both openings and summing the later accuracy gains gives

$$\boxed{\frac{\min\{\mathcal{E}_1,\mathcal{E}_2\} - \mathcal{E}_R}{T} \ge \varepsilon^2\bar{N} - \frac{2}{T}\sum_{k=3}^{R}\Lambda_k.}$$

If $\Lambda_k/T \le \gamma$ at every reply, calibration adds $2\gamma/\varepsilon^2$ to the unresolved fraction and at most $2(R-2)\gamma$ to the accuracy allowance.

Next we learn small residuals from outcomes. We can use expected error bounds simply by taking expectations in these inequalities.

What We Assume About the Human

In the theorem, either speaker's replies must satisfy the stated calibration checks. The openings remain arbitrary, so each party can start with whatever its private information supports.

What We Assume About the Human

In the theorem, either speaker's replies must satisfy the stated calibration checks. The openings remain arbitrary, so each party can start with whatever its private information supports.
This replaces exact Bayesian updating with reliability conditional on the reports being answered. Bayesian posteriors have zero conditional mean error on these cells under their probability model; finite-sample residuals need not be exactly zero.

What We Assume About the Human

In the theorem, either speaker's replies must satisfy the stated calibration checks. The openings remain arbitrary, so each party can start with whatever its private information supports.
This replaces exact Bayesian updating with reliability conditional on the reports being answered. Bayesian posteriors have zero conditional mean error on these cells under their probability model; finite-sample residuals need not be exactly zero.
In the construction, we learn small residuals from earlier reports and past outcomes. The AI can run this procedure directly; a reporting assistant can supply such replies on the human side.
An unassisted human need not satisfy the condition. The construction shows that the reliability we ask for can be enforced efficiently.

One Forecaster Learns Each Position in the Conversation

For each reply round $k = 3, \ldots, R$, run a separate procedure across days. On a day that reaches round k:

1. Observe the previous message and, at round 3, also the AI's opening.

One Forecaster Learns Each Position in the Conversation

For each reply round $k = 3, \ldots, R$, run a separate procedure across days. On a day that reaches round k:

1. Observe the previous message and, at round 3, also the AI's opening.
2. Choose a distribution over reports, then draw and send p_t^k.

One Forecaster Learns Each Position in the Conversation

For each reply round $k = 3, \ldots, R$, run a separate procedure across days. On a day that reaches round k:

1. Observe the previous message and, at round 3, also the AI's opening.
2. Choose a distribution over reports, then draw and send p_t^k.
3. After the conversation, observe y_t and update this procedure.

One Forecaster Learns Each Position in the Conversation

For each reply round $k = 3, \ldots, R$, run a separate procedure across days. On a day that reaches round k:

1. Observe the previous message and, at round 3, also the AI's opening.
2. Choose a distribution over reports, then draw and send p_t^k.
3. After the conversation, observe y_t and update this procedure.

The outcome is fixed before the day's conversation and all its random draws. It cannot be chosen in response to a sampled report.

Private information may help choose among distributions satisfying the forecasting rule below.

Each Pairwise Check Supplies One Block of Coordinates

Fix reply round k. For an earlier report q, candidate reply p, and outcome y, define the M^2-coordinate vector

$$v(q, p, y)_{q',p'} := \mathbf{1}\{q = q',\ p = p'\}(y - p),$$

where $q', p' \in \mathcal{P}$. Only one coordinate can be nonzero.

Each Pairwise Check Supplies One Block of Coordinates

Fix reply round k. For an earlier report q, candidate reply p, and outcome y, define the M^2-coordinate vector

$$v(q,p,y)_{q',p'} := \mathbf{1}\{q = q',\ p = p'\}(y - p),$$

where $q', p' \in \mathcal{P}$. Only one coordinate can be nonzero.
On active day t, the complete vector is

$$g_t(p,y) := \begin{cases} \big(v(p_t^2,p,y), v(p_t^1,p,y)\big), & k = 3, \\ v(p_t^{k-1},p,y), & k \geq 4. \end{cases}$$

Set $g_t = 0$ on inactive days. There are at most $2M^2$ coordinates, with at most two nonzero, so $\|g_t(p,y)\|_2 \leq \sqrt{2}$.

Each Pairwise Check Supplies One Block of Coordinates

Fix reply round k. For an earlier report q, candidate reply p, and outcome y, define the M^2-coordinate vector

$$v(q,p,y)_{q',p'} := \mathbf{1}\{q = q',\ p = p'\}(y - p),$$

where $q', p' \in \mathcal{P}$. Only one coordinate can be nonzero.
On active day t, the complete vector is

$$g_t(p,y) := \begin{cases} \big(v(p_t^2,p,y), v(p_t^1,p,y)\big), & k = 3, \\ v(p_t^{k-1},p,y), & k \geq 4. \end{cases}$$

Set $g_t = 0$ on inactive days. There are at most $2M^2$ coordinates, with at most two nonzero, so $\|g_t(p,y)\|_2 \leq \sqrt{2}$.
The average $\bar{g}_T = T^{-1}\sum_{t \in S_k} g_t(p_t^k, y_t)$ consists of $r_{k,q,p}/T$, with $s_{q,p}/T$ appended at round 3. Thus

$$\|\bar{g}_T\|_1 = \Lambda_k / T.$$

The Calibration Target Is Response Satisfiable

Choose the finite grid so that every $z \in [0,1]$ is within η of a grid report. Our target is the closed convex set

$$\mathcal{K}_\eta := \{u : \|u\|_1 \le 2\eta\} \quad \text{in } \mathbb{R}^{2M^2} \text{ at round 3 and } \mathbb{R}^{M^2} \text{ afterwards.}$$

We still use Lecture 8's *Euclidean distance* to this target.

The Calibration Target Is Response Satisfiable

Choose the finite grid so that every $z \in [0,1]$ is within η of a grid report. Our target is the closed convex set

$$\mathcal{K}_\eta := \{u : \|u\|_1 \le 2\eta\} \quad \text{in } \mathbb{R}^{2M^2} \text{ at round 3 and } \mathbb{R}^{M^2} \text{ afterwards.}$$

We still use Lecture 8's *Euclidean distance* to this target.
Fix the earlier report(s) and any outcome distribution with mean z.
Choose p with $|p - z| \le \eta$. In each block, the expected vector has just one possibly nonzero entry:

$$\mathbb{E}[v(q,p,y)]_{q',p'} = \mathbf{1}\{q = q',\ p = p'\}(z - p).$$

There are at most two blocks. Hence $\|\mathbb{E}[g_t(p,y)]\|_1 \le 2\eta$, which puts the expected vector in $\mathcal{K}_\eta$. Inactive days give zero.

Blackwell Supplies the Online Forecasting Rule

The response above was allowed to know the outcome distribution. Lecture 8's minimax argument turns those hypothetical responses into one rule that works against either outcome.

Blackwell Supplies the Online Forecasting Rule

The response above was allowed to know the outcome distribution. Lecture 8's minimax argument turns those hypothetical responses into one rule that works against either outcome.

Initialize $\bar{g}_0 = 0$. After observing the earlier report(s), project $\bar{g}_{t-1}$ onto $\mathcal{K}_\eta$, obtaining c. Set $w = \bar{g}_{t-1} - c$. Choose a distribution ν_t over reports such that

$$\boxed{\mathbb{E}_{p\sim\nu_t}[\langle w, g_t(p, y) - c\rangle] \leq 0 \quad \text{for both } y \in \{0, 1\}.}$$

Response satisfiability guarantees that such a distribution exists. Finding it is a linear feasibility problem in M report probabilities, with one constraint for each of the two outcomes.

Blackwell Supplies the Online Forecasting Rule

The response above was allowed to know the outcome distribution. Lecture 8's minimax argument turns those hypothetical responses into one rule that works against either outcome.

Initialize $\bar{g}_0 = 0$. After observing the earlier report(s), project $\bar{g}_{t-1}$ onto $\mathcal{K}_\eta$, obtaining c. Set $w = \bar{g}_{t-1} - c$. Choose a distribution ν_t over reports such that

$$\boxed{\mathbb{E}_{p\sim\nu_t}[\langle w, g_t(p, y) - c\rangle] \leq 0 \quad \text{for both } y \in \{0, 1\}.}$$

Response satisfiability guarantees that such a distribution exists. Finding it is a linear feasibility problem in M report probabilities, with one constraint for each of the two outcomes.

Draw and send $p \sim \nu_t$. When y_t is revealed, add the observed vector to the running average. Each round-k procedure maintains its own average.

The Distance Bound Controls Calibration Error

Since $0 \in \mathcal{K}_\eta$ and $\|g_t\|_2 \le \sqrt{2}$, Lecture 8 gives

$$\mathbb{E}[\mathrm{dist}(\bar{g}_T, \mathcal{K}_\eta)^2] \le \frac{8}{T}.$$

Recall $\Lambda_k / T = \|\bar{g}_T\|_1$, the sum of absolute residual coordinates. Let c be the Euclidean projection of $\bar{g}_T$ onto $\mathcal{K}_\eta$.

The Distance Bound Controls Calibration Error

Since $0 \in \mathcal{K}_\eta$ and $\|g_t\|_2 \le \sqrt{2}$, Lecture 8 gives

$$\mathbb{E}[\operatorname{dist}(\bar{g}_T, \mathcal{K}_\eta)^2] \le \frac{8}{T}.$$

Recall $\Lambda_k / T = \|\bar{g}_T\|_1$, the sum of absolute residual coordinates. Let c be the Euclidean projection of $\bar{g}_T$ onto $\mathcal{K}_\eta$. The triangle inequality and Cauchy–Schwarz over at most $2M^2$ coordinates give

$$\begin{aligned} \frac{\Lambda_k}{T} &\le \|c\|_1 + \|\bar{g}_T - c\|_1 \\ &\le 2\eta + \sqrt{2}M\|\bar{g}_T - c\|_2. \end{aligned}$$

The Distance Bound Controls Calibration Error

Since $0 \in \mathcal{K}_\eta$ and $\|g_t\|_2 \le \sqrt{2}$, Lecture 8 gives

$$\mathbb{E}[\mathrm{dist}(\bar{g}_T, \mathcal{K}_\eta)^2] \le \frac{8}{T}.$$

Recall $\Lambda_k / T = \|\bar{g}_T\|_1$, the sum of absolute residual coordinates. Let c be the Euclidean projection of $\bar{g}_T$ onto $\mathcal{K}_\eta$. The triangle inequality and Cauchy–Schwarz over at most $2M^2$ coordinates give

$$\begin{aligned} \frac{\Lambda_k}{T} &\le \|c\|_1 + \|\bar{g}_T - c\|_1 \\ &\le 2\eta + \sqrt{2}M\|\bar{g}_T - c\|_2. \end{aligned}$$

Taking expectations and applying Cauchy–Schwarz once more gives

$$\begin{aligned} \mathbb{E}[\Lambda_k / T] &\le 2\eta + \sqrt{2}M\sqrt{\mathbb{E}[\mathrm{dist}(\bar{g}_T, \mathcal{K}_\eta)^2]} \\ &\le 2\eta + \frac{4M}{\sqrt{T}}. \end{aligned}$$

We use this one bound for every reply round.

Learned Calibration Gives Short Conversations

Our deterministic proof established

$$\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2} + \frac{2\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon^2}.$$

Take expectations and use the bound on each $\mathbb{E}[\Lambda_k/T]$:

$$\boxed{\mathbb{E}\left[\frac{|S_{R+1}|}{T}\right] \leq \frac{1}{(R-2)\varepsilon^2} + \frac{4\eta}{\varepsilon^2} + \frac{8M}{\varepsilon^2\sqrt{T}}.}$$

Learned Calibration Gives Short Conversations

Our deterministic proof established

$$\frac{|S_{R+1}|}{T} \leq \frac{1}{(R-2)\varepsilon^2} + \frac{2\sum_{k=3}^{R} \Lambda_k}{T(R-2)\varepsilon^2}.$$

Take expectations and use the bound on each $\mathbb{E}[\Lambda_k/T]$:

$$\boxed{\mathbb{E}\left[\frac{|S_{R+1}|}{T}\right] \leq \frac{1}{(R-2)\varepsilon^2} + \frac{4\eta}{\varepsilon^2} + \frac{8M}{\varepsilon^2\sqrt{T}}.}$$

For any desired tolerance and unresolved fraction, choose enough replies, a fine enough grid, and enough days of feedback. A uniform grid has $M = O(1/\eta)$, so all the required resources are polynomial in the inverse tolerances.
The expectation averages the protocol's random draws. The conclusion controls the fraction of these T conversations that remain unresolved, not every day or a later deployment period.

Learning Preserves the Accuracy Gain from Conversation

Recall that $\bar{N}$ averages, across days, the number of reply rounds after round 3 that still leave consecutive reports more than ε apart. We proved

$$\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T} \geq \varepsilon^2 \bar{N} - \frac{2}{T} \sum_{k=3}^{R} \Lambda_k.$$

Learning Preserves the Accuracy Gain from Conversation

Recall that $\bar{N}$ averages, across days, the number of reply rounds after round 3 that still leave consecutive reports more than ε apart. We proved

$$\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T} \geq \varepsilon^2 \bar{N} - \frac{2}{T} \sum_{k=3}^{R} \Lambda_k.$$

Using $\mathbb{E}[\Lambda_k / T] \leq 2\eta + 4M/\sqrt{T}$ for each of the $R - 2$ reply rounds gives

$$\boxed{\mathbb{E}\left[\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T}\right] \geq \varepsilon^2 \mathbb{E}[\bar{N}] - 2(R-2)\left(2\eta + \frac{4M}{\sqrt{T}}\right).}$$

Learning Preserves the Accuracy Gain from Conversation

Recall that $\bar{N}$ averages, across days, the number of reply rounds after round 3 that still leave consecutive reports more than ε apart. We proved

$$\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T} \geq \varepsilon^2 \bar{N} - \frac{2}{T} \sum_{k=3}^{R} \Lambda_k .$$

Using $\mathbb{E}[\Lambda_k / T] \leq 2\eta + 4M/\sqrt{T}$ for each of the $R - 2$ reply rounds gives

$$\boxed{\mathbb{E}\left[\frac{\min\{\mathcal{E}_1, \mathcal{E}_2\} - \mathcal{E}_R}{T}\right] \geq \varepsilon^2 \mathbb{E}[\bar{N}] - 2(R-2)\left(2\eta + \frac{4M}{\sqrt{T}}\right).}$$

The final reports match the better opening's aggregate accuracy up to the displayed allowance. They are strictly better when the progress from these reply rounds with continued disagreement exceeds that allowance.

Summary

- Public Bayesian conversation eventually makes current beliefs common knowledge, at which point Aumann forces agreement. Finiteness alone does not make the conversation short.

Summary

- Public Bayesian conversation eventually makes current beliefs common knowledge, at which point Aumann forces agreement. Finiteness alone does not make the conversation short.
- Calibrated replies turn disagreement into accuracy improvement. Bounded error makes most conversations short; approachability learns the required reliability from outcomes.

Summary

- Public Bayesian conversation eventually makes current beliefs common knowledge, at which point Aumann forces agreement. Finiteness alone does not make the conversation short.
- Calibrated replies turn disagreement into accuracy improvement. Bounded error makes most conversations short; approachability learns the required reliability from outcomes.
- Conversation can protect the better opening's aggregate accuracy, up to calibration error. Further unresolved exchanges then yield additional improvement.

Summary

- Public Bayesian conversation eventually makes current beliefs common knowledge, at which point Aumann forces agreement. Finiteness alone does not make the conversation short.
- Calibrated replies turn disagreement into accuracy improvement. Bounded error makes most conversations short; approachability learns the required reliability from outcomes.
- Conversation can protect the better opening's aggregate accuracy, up to calibration error. Further unresolved exchanges then yield additional improvement.
- Next: a shared decision may require much less agreement than matching numerical probabilities. Can the parties exchange actions instead?

Thanks!

See you next class!

References

- Robert J. Aumann. *Agreeing to Disagree*. Annals of Statistics, 1976.
- John D. Geanakoplos and Heraklis M. Polemarchakis. *We Can't Disagree Forever*. Journal of Economic Theory, 1982.
- Scott Aaronson. *The Complexity of Agreement*. STOC, 2005.
- Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. *Tractable Agreement Protocols*. STOC, 2025.
- David Blackwell. *An Analog of the Minimax Theorem for Vector Payoffs*. Pacific Journal of Mathematics, 1956.