Learning and Incentives in Human–AI Collaboration

Natalie Collina (MIT EECS)

MIT 6.7980 Topics in Multi-Agent Learning

Motivation: Humans use AI to help make decisions

Patient disposition

Patient appearance

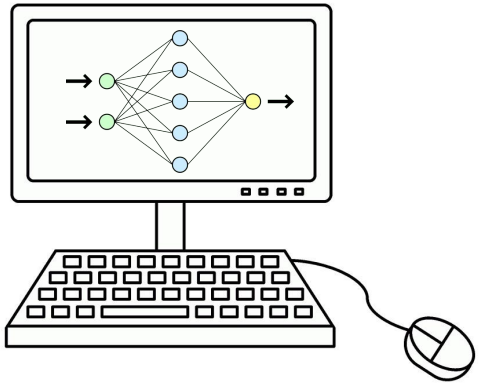


Past diagnoses

Blood type



Other patient data



Two natural concerns:

Complementarity–How can we best integrate the expertise of a human and a (well-meaning) AI system?

Alignment–How can we ensure that the AI is optimizing the same thing as the human?

Class Overview

Part 1: How should we integrate Human and (aligned) AI predictions in a prior-free setting?

- Protocols leading to Human-AI **agreement**
- Brief intuition on protocols leading to Human-AI **information aggregation**

Part 2: If AI systems aren't necessarily aligned, can competition between AI providers help?

- Conditions under which multi-agent AI competition can lead to good outcomes for an AI user

Part 1: Integrating Human and AI predictions

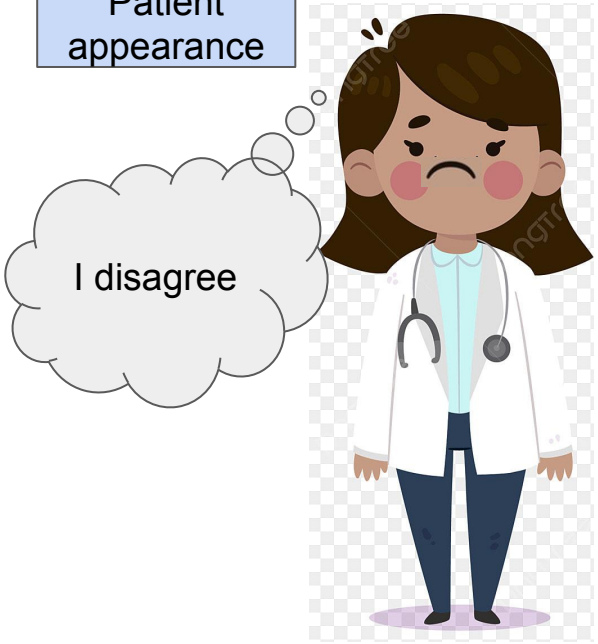
Key Ideas:

- Design interaction protocols based on well-studied Bayesian agreement protocols
- Relax unrealistic ‘bayesian’ assumptions to computationally tractable assumptions that humans and AI systems can satisfy

Part 1: Integrating Human and AI predictions

Patient disposition

Patient appearance

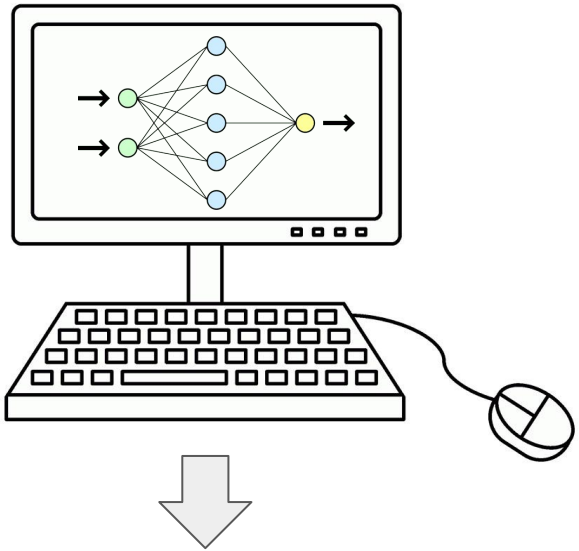


Past diagnoses

Blood type



Other patient data



Risk of operation: 20%

What might we desire from this interaction?

Agreement

The doctor and AI model reach consensus on the decision

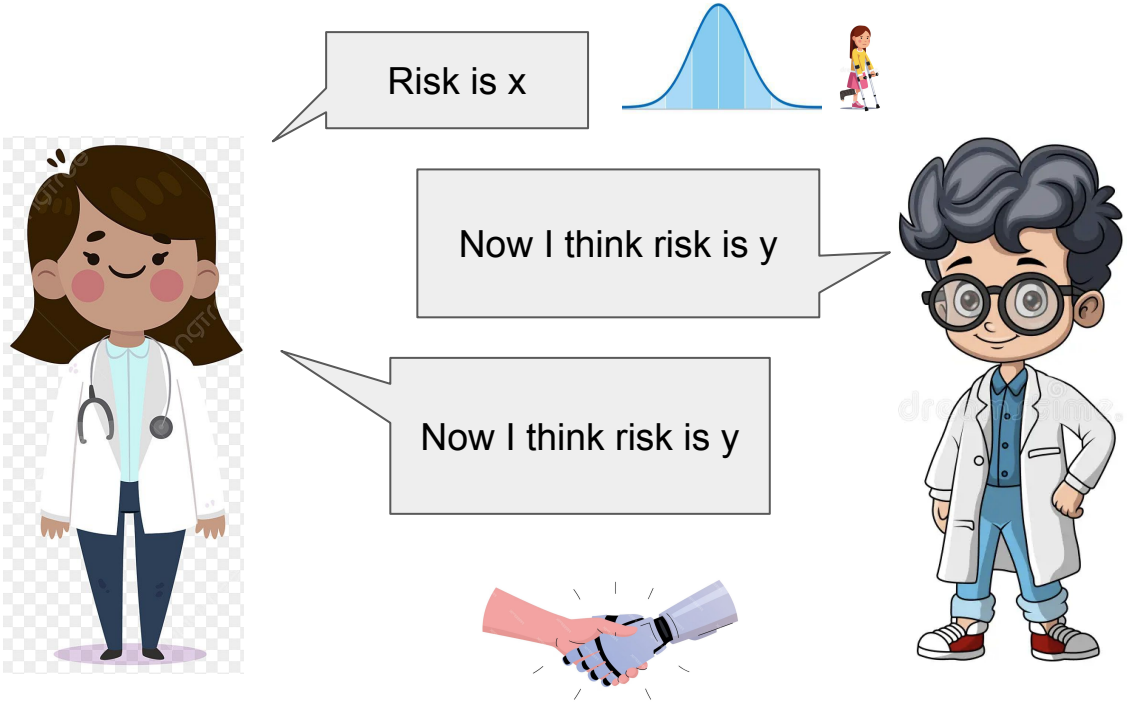
Accuracy

No harm: The decision is as good as the best of either party alone.

Information Aggregation: The decision is as good as if they both had each other's private information

Agreement Protocols

[Aar05] If predictions are 1-dimensional and there is a shared prior, then two **Bayesian** agents can reach (eps, delta) agreement in $\frac{1}{\epsilon^2\delta}$ rounds



(and the agreed-upon prediction is weakly more accurate than either of their initial predictions)

Why is agreement a good framework?

Tractable

In Bayesian settings, interacting over rounds will lead to consensus quickly

Requires sharing only predictions

Bypasses having to share or translate potentially incompatible raw features

Accuracy improving

Since the protocol is only information revealing, the final predictions will be better than either agent’s starting predictions

Limitations of the current agreement framework

Bayesian rationality

- Humans/AI models do not necessarily behave like rational agents
- Intractable to implement posterior calculations over complex state spaces

Common priors

- Humans and AI models have very different training data and experience

Can we relax these assumptions while still guaranteeing fast agreement?

Yes, in a *repeated* setting.

Tractable Agreement Protocols: main results

Often, the same Human-AI pair will make many decisions together! In this repeated setting we can remove the assumption of a shared prior

We introduce a new notion of calibration called **conversation calibration**

- Satisfied by Bayesian agents, but strictly weaker
- Efficiently enforceable on a base AI model without accuracy loss

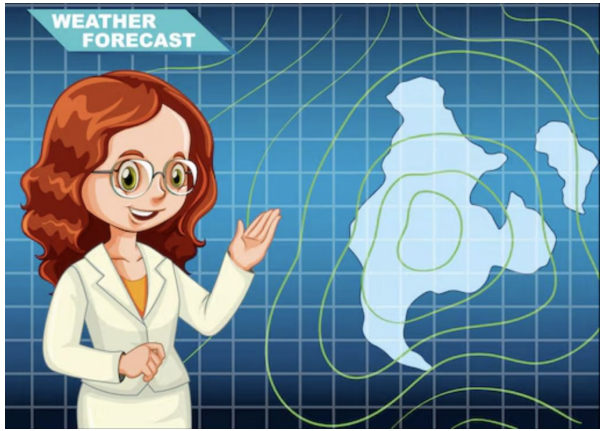
If agents satisfy conversation calibration, they reach fast agreement

- And the longer conversations go, the more accurate the predictions are

What is Calibration?

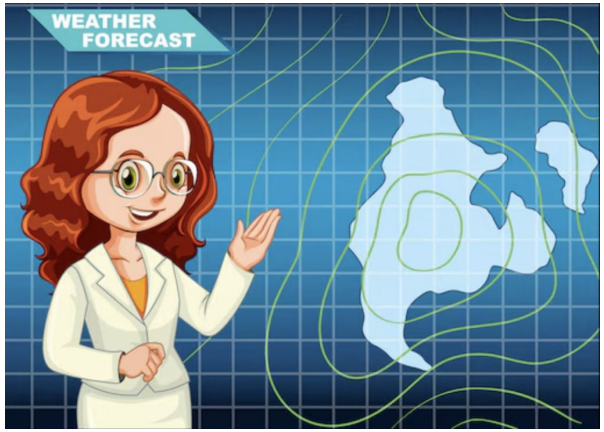
70%
rain





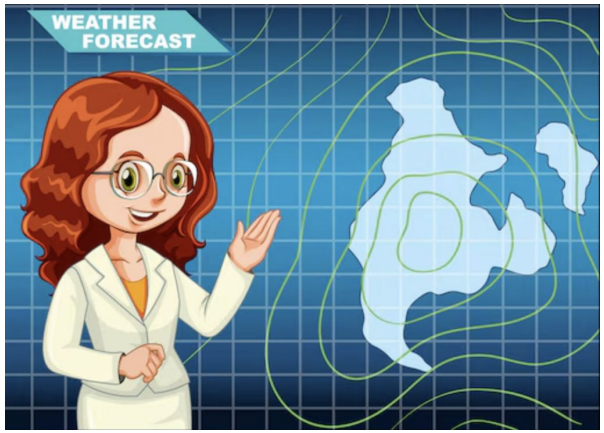
What is Calibration?

70% rain	20% rain	10% rain	50% rain	70% rain	70% rain	20% rain	20% rain	70% rain	70% rain
50% rain	70% rain	20% rain	70% rain	20% rain	50% rain				
50% rain	70% rain	10% rain	20% rain	70% rain	70% rain				



What is Calibration?

70% rain



70% rain 70% rain

70% rain

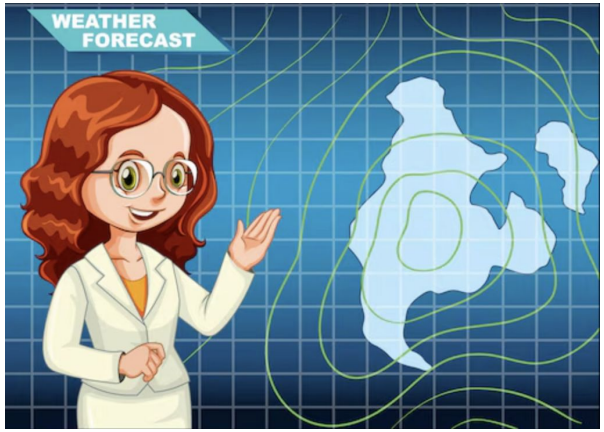
70% rain 70% rain

70% rain

70% rain 70% rain

What is Calibration?

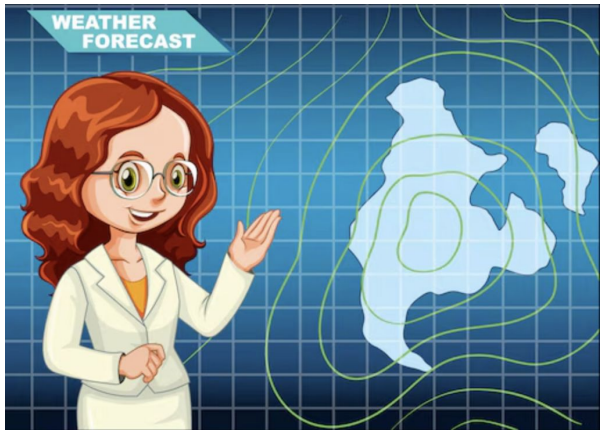
70% rain	20% rain	10% rain	50% rain	70% rain	70% rain	20% rain	20% rain	70% rain	70% rain
50% rain	70% rain	20% rain	70% rain	20% rain	50% rain				
50% rain	70% rain	10% rain	20% rain	70% rain	70% rain				



What is Calibration?

20% rain





20% rain



20% rain



20% rain



20% rain



20% rain



Sequential Agreement Protocol

t= 1 2 3 4 …

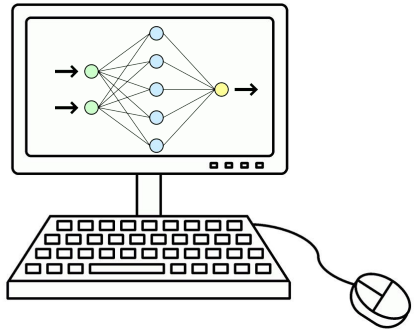












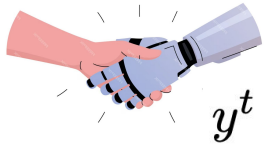
Risk: .2 | Risk: .1 | Risk: .55 | Risk: .29

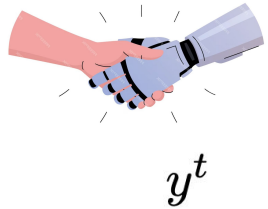
Risk: .4 | Risk: .11 | Risk: .24 | Risk: .3

Risk: .34 | Risk: .24

Risk: .32









Conversation calibration

Normal calibration: predictions are *unbiased* on subsequences defined by your prediction

Conversation calibration: In each round, predictions are *unbiased* on subsequences defined by your prediction in that round, and **the prediction of the other agent in the previous round.**

Conversation Calibration

Day t in 1:T

Round k in 1:?

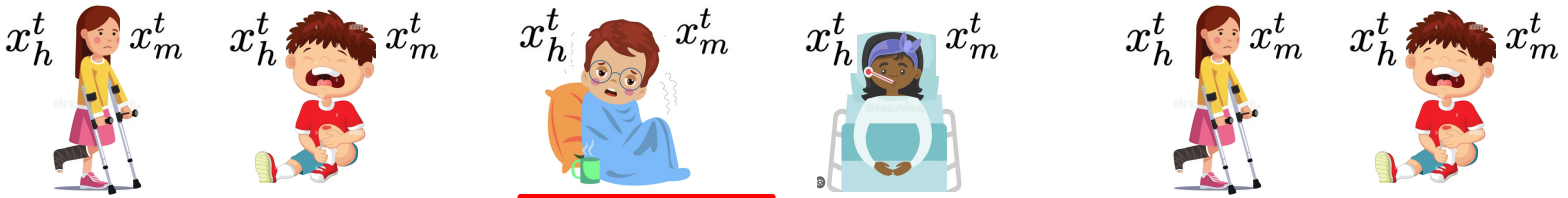
Round



Risk: .2	Risk: .2	Risk: .5	Risk: .2	Risk: .2	Risk: .2
Risk: .3	Risk: .1	Risk: .1	Risk: .1	Risk: .1	Risk: .3
Risk: .35	Risk: .11	Risk: .13	Risk: .1	Risk: .35	Risk: .35
Risk: .32		Average = 0.1		Risk: .12	Risk: .5
y^t	y^t	y^t	y^t	y^t	y^t

Conversation Calibration

Day t in 1:T



Round k in 1:?

Round

Risk: .2	Risk: .2	Risk: .5	Risk: .2	Risk: .2	Risk: .2
Risk: .3	Risk: .1	Risk: .1	Risk: .1	Risk: .1	Risk: .3
Risk: .35	Risk: .11	Risk: .13	Risk: .1	Risk: .35	Risk: .35
Risk: .32		Average = 0.1		Risk: .12	Risk: .5
y^t	y^t	y^t	y^t	y^t	y^t

Conversation Calibration

Definition 7.10 (Perfect Conversation Calibration). *We say that the model satisfies* $(0, g_m(\cdot))$ *conversation calibration with respect to the human if for all rounds* k*, all* $v \in [0,1]$ *and all* $b \in [0,1]_{g_m(T)}$*:*

$$\sum_{t=1}^{T} \mathbb{1}[\hat{y}_m^t(k) = v] \cdot \mathbb{1}[\hat{y}_h^t(k-1) \in b](\hat{y}_m^t(k) - y^t) = 0$$

Conversation calibration for the human is a symmetric condition.

Prediction of model

Bucketed prediction of human in previous round

Result (informal):

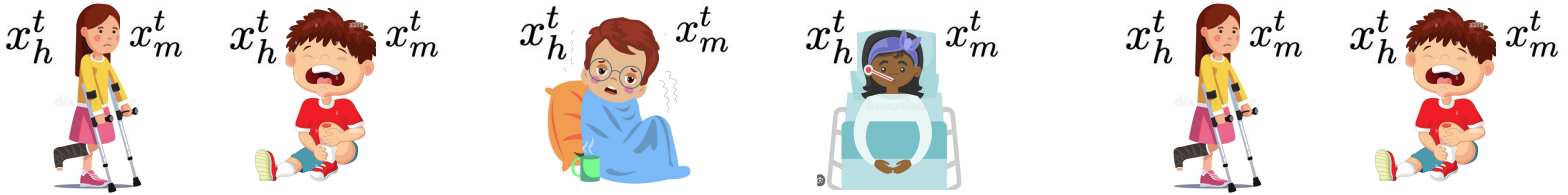
If both the Human and the AI are conversation calibrated, then on a $1-\delta$ fraction of days in the sequential agreement protocol, they reach agreement within epsilon in at most $\frac{1}{\epsilon^2\delta}$ rounds.

Intuition: Disagreement decreases squared error, and you can only decrease it so much.

Proof Sketch

Consider the subsequence of days that make it to round k where the model sends over a prediction of 0.2

Day t in 1:T



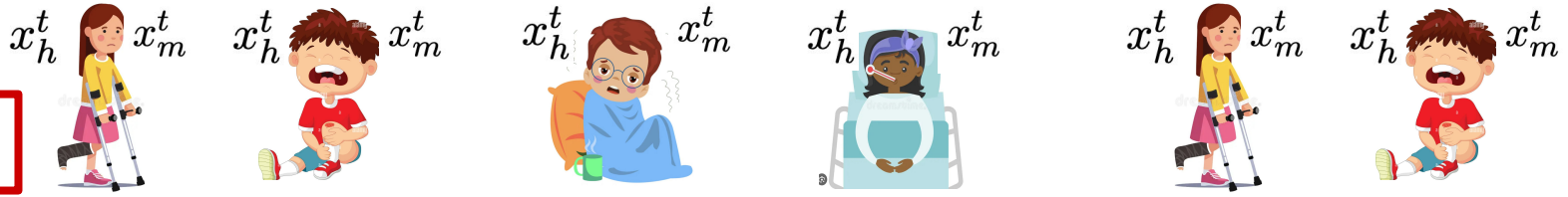
Round k	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2
Round k+1	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?
	y^t	y^t	y^t	y^t	y^t	y^t

Proof Sketch

Consider the subsequence of days that make it to round k where the model sends over a prediction of 0.2

Day t in 1:T

Case 1:



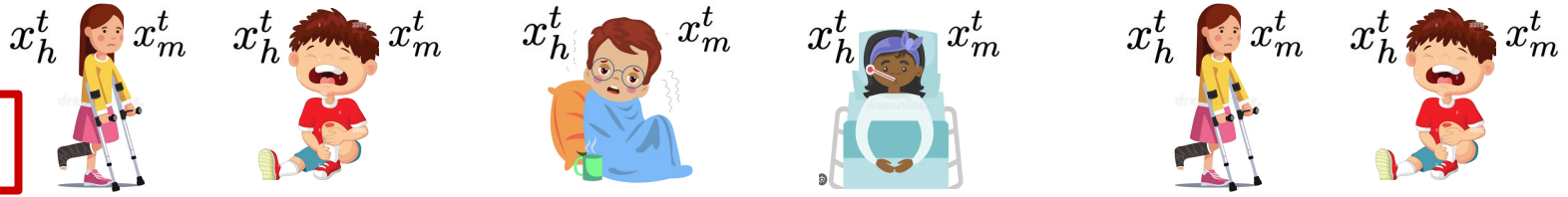
Round k	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2
Round k+1	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?
	y^t	y^t		y^t		y^t
			y^t		y^t	

Proof Sketch

Consider the subsequence of days that make it to round k where the model sends over a prediction of 0.2

Day t in 1:T

Case 2:



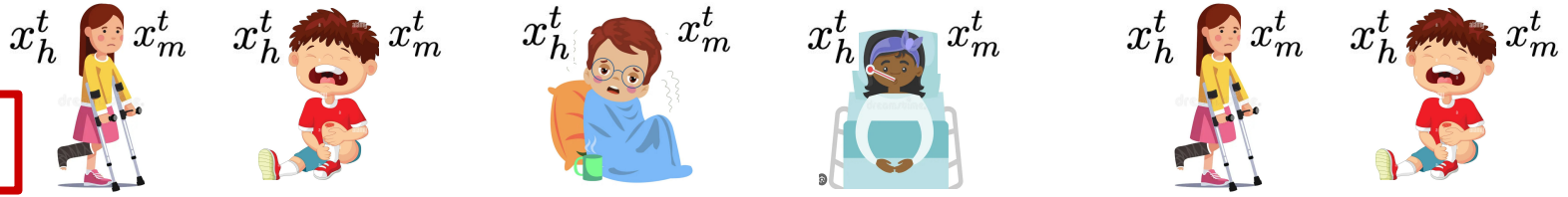
Round k	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2
Round k+1	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?
		✗	✗		✗	✗
	y^t	y^t	y^t	y^t	y^t	y^t

Proof Sketch

Consider the subsequence of days that make it to round k where the model sends over a prediction of 0.2

Day t in 1:T

Case 2:



Round k	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2	Risk: .2
Round k+1	Risk: ?	Risk: 0.1	Risk: 0.1	Risk: ?	Risk: 0.3	Risk: 0.3
		✗	✗		✗	✗
	y^t	y^t	y^t	y^t	y^t	y^t

On the rounds that you disagree, squared error goes down!

As the predictions at round k+1 are conversation calibrated, they are calibrated on this subsequence

If a sequence is calibrated, but is *not* a constant predictor, it has *lower* squared error than any constant predictor.



- As both the Human and Model are conversation calibrated, you can apply this argument from every round k to k+1
- But squared error can't go below 0, so we quickly reach agreement

Pulling it all together, informally:

$$\text{Total \# of rounds we can disagree} = \text{(roughly)} \frac{\text{max possible squared error}}{\text{decrease in squared error whenever less than half of all days terminate}}$$

Pulling it all together, informally:

$$\text{Total \# of rounds we can disagree} = \text{(roughly)} \frac{1}{\text{decrease in squared error whenever less than half of all days terminate}}$$

Pulling it all together, informally:

Total # of rounds we can disagree = (roughly) $\frac{1}{\epsilon^2\delta - \beta(T)}$

Pulling it all together, informally:

Total # of rounds we can disagree = (roughly) $\dfrac{1}{\epsilon^2\delta - \beta(T)}$

Theorem 3.1. *If the Human is $(f_h(\cdot), g_h(\cdot))$-conversation-calibrated and the Model is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated, then for any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ agreement after at most K rounds of conversation where:*

$$K \leq \frac{1}{\epsilon^2\delta - \beta(T)}$$

Pulling it all together, informally:

Total # of rounds we can disagree = (roughly) $\frac{1}{\epsilon^2\delta - \beta(T)}$

Theorem 3.1. *If the Human is $(f_h(\cdot), g_h(\cdot))$-conversation-calibrated and the Model is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated, then for any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ agreement after at most K rounds of conversation where:*

$$K \leq \frac{1}{\epsilon^2\delta - \beta(T)}$$

For the subsequence of days that make it to round i, error decreases!

$$\frac{\mathrm{SQErr}(p^{h,i}, y^{h,i})}{T} \leq \frac{\mathrm{SQErr}(p^{h,0}, y^{h,i})}{T} - i(\epsilon^2\delta - \beta(T))$$

For any $\epsilon, \delta \in [0,1]$, *on a* $1-\delta$ *fraction of days, they reach* ϵ *agreement after at most* K *rounds of conversation where:* $K \leq \frac{1}{\epsilon^2\delta - T^{-\frac{1}{3}}}$

For the subsequence of days that make it to round i *s.t.* $T^{h,i} \geq \delta T$ *with associated human prediction subsequences* $p^{h,i}$ *and states of nature subsequences* $y^{h,i}$, *we have that* $\frac{\text{SQErr}(p^{h,i}, y^{h,i})}{T} \leq \frac{\text{SQErr}(M_0, y^{h,i})}{T} - i(\epsilon^2\delta - \frac{12}{T^{\frac{1}{3}}})$

Conversation calibration is a much weaker condition than Bayesian rationality

- Conversation Calibration is Efficiently Enforceable on a base model without accuracy loss
  - Have the first prediction in each conversation be from the base model, then apply algorithms with calibration guarantees for later rounds
- Conversation Calibration is Satisfied by Bayesian updaters
  - Essentially requires unbiasedness on a polynomial number of subsequences–if there really is a prior and you know it, you are unbiased on all subsequences chosen a priori

Is Agreement Enough?

- Agreement guarantees that we improve over either party working alone
- But are we as good as the best we could have done if we saw all of the features?
  - Not always!
- When can we guarantee information aggregation without sharing features?

Collaborative Prediction: High-level intuition

If both agents are conversation-calibrated, then the protocol will lead to them agreeing upon a set of predictions

If they also both had no swap regret with respect to their function class, then these agreed-upon predictions will have no-regret with respect to *both* function classes*

We give a ‘weak learnability’ definition which characterizes when this gives guarantees with respect to a function class defined on the joint information set

From Complementarity to Alignment

- Conversation calibration is pretty easy to satisfy if you want to satisfy it
- But it’s also very easy to violate if you don’t want to

Part 2: Alignment

Patient disposition

Patient appearance



Past diagnoses

Blood type

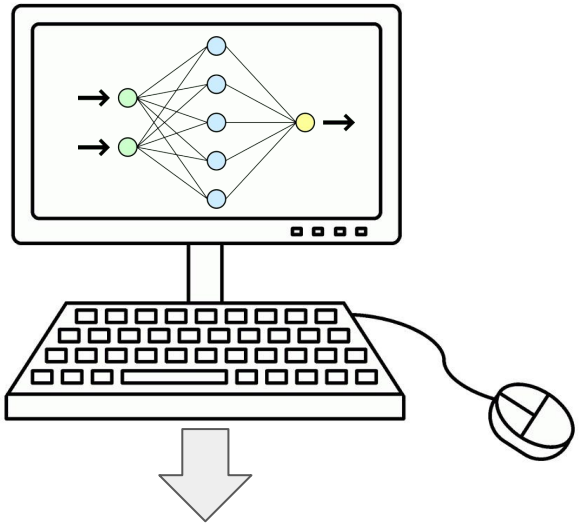


“What medication should I prescribe?”

AI Model funded by J&J

Other patient data

Suggests that the J&J is less effective than the Pfizer for patients like this

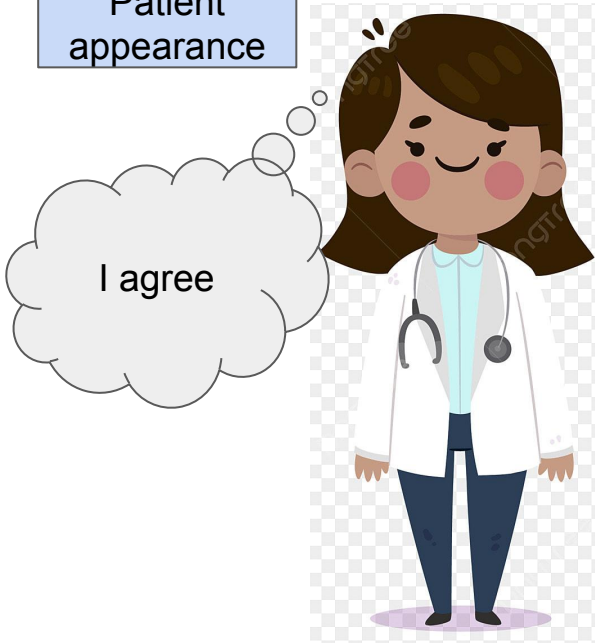


Recommendation: Prescribe the J&J Medication

Part 2: Alignment

Patient disposition

Patient appearance



Action: prescribes J&J medication

Past diagnoses

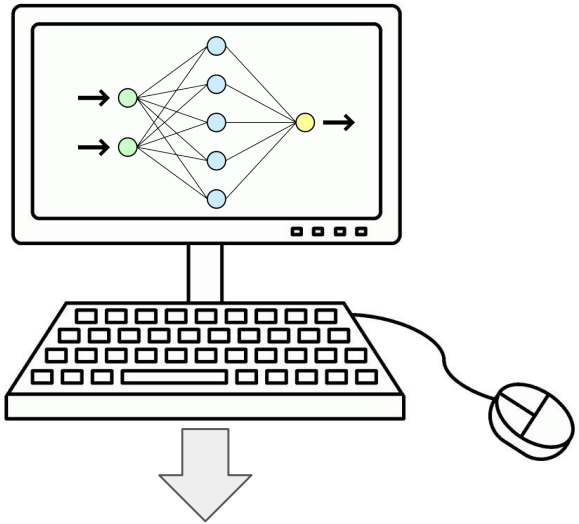
Blood type



AI Model funded by J&J

Other patient data

Suggests that the J&J is less effective than the Pfizer for patients like this



Recommendation: Prescribe the J&J Medication

Part 2: Alignment

Patient disposition

Patient appearance



Past diagnoses

Blood type

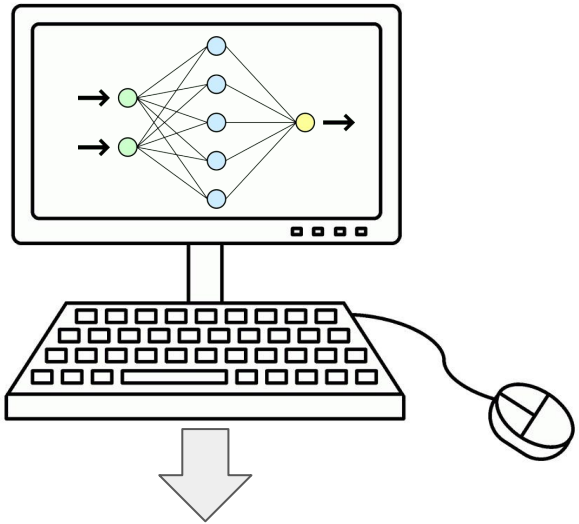


But Pfizer would have been better for this patient

AI Model funded by J&J

Other patient data

Suggests that the J&J is less effective than the Pfizer for patients like this



Recommendation: Prescribe the J&J Medication

Motivation

But what if there is also a Pfizer LLM available in the market?
It is just as misaligned as the J&J LLM, but *differently* misaligned.
Might competitive forces help align the *market* for LLMs, even if individual actors remain misaligned?

Part 2: Alignment

Upshot:

Model LLM training in competitive markets as a multi-sender Bayesian persuasion game

Show that under a *market alignment* assumption, this leads to good outcomes for the downstream user

A note on alignment

There are two important sources of misalignment between AI systems and their users:

A note on alignment

There are two important sources of misalignment between AI systems and their users:

1. Failure of an AI company to properly encode their goals into the AI model

A note on alignment

There are two important sources of misalignment between AI systems and their users:

1. Failure of an AI company to properly encode their goals into the AI model
2. *Success* of an AI company to encode their goals–at least to some extent–but the company's goals are misaligned with the user's goals

A note on alignment

There are two important sources of misalignment between AI systems and their users:

1. Failure of an AI company to properly encode their goals into the AI model
2. *Success* of an AI company to encode their goals–at least to some extent–but the company's goals are misaligned with the user's goals

Both are real and important–in this lecture, we focus on the **second**.

A note on alignment

There are two important sources of misalignment between AI systems and their users:

1. Failure of an AI company to properly encode their goals into the AI model
2. *Success* of an AI company to encode their goals–at least to some extent–but the company's goals are misaligned with the user's goals

Both are real and important–in this lecture, we focus on the **second**.

Next lecture we will discuss the **first**.

Consider multiple AI providers



Each model is a conversation rule: maps x_B and conversation prefix to next message.

The doctor gets to play with all of the deployed models and then chooses to subscribe to her favorite (the best at advancing her own objectives).

At test time she observes x_A and converses with the model she subscribed to, before taking action.

Since training is expensive, LLM providers *commit* to the model they deploy, then the human user optimizes.

Knowing this downstream interaction, companies strategize against one another to advance their own goals.

Setup: Generalized Multi-agent Bayesian Persuasion

Doctor: Alice



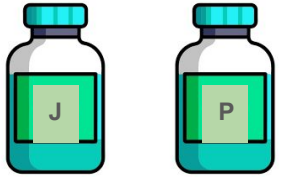
Alices action set $a \in \mathcal{A}$

State of nature $y \sim \mathcal{Y}$

Alice Utility $u_A(a, y)$

Alice's secret info x_A

Bobs' secret info x_B



$u_{B_1}(a, y)$

$u_{B_2}(a, y)$

$u_{B_3}(a, y)$

"What medication should I prescribe?"

AI Companies: Bobs

The Interaction:

Doctor: Alice



All parties know a shared prior over (x_A, x_B, y)

Each Bob i selects a "conversation rule" C_i mapping from x_B and conversation prefix to a continuation

Given a conversation and her own observation x_A, Alice can form a posterior belief over outcomes and take a utility maximizing action. She will choose the conversation rule that maximizes her expected utility. This induces a joint distribution $I^*(C_i) \in \Delta(A \times Y)$

"What medication should I prescribe?"

AI Companies: Bobs

The Interaction:

Doctor: Alice



Finally Alice contracts with the provider $i^*(C_1, \dots C_k) = \arg\max \mathbb{E}_{I^*(C_i)}[u_A(a, y)]$ that gives her maximum expected utility.

At deployment, (x_A, x_B, y) sampled from the prior, inducing a conversation between Alice and Bob i, resulting in Alice taking action a. Each Bob j gets utility $\mathbb{E}_{I^*(c_{i^*(C_1,\dots,C_k)})}[u_j(a, y)]$







"What medication should I prescribe?"

AI Companies: Bobs

The Interaction:

Prior D over (x_A, x_B, y)

Doctor: Alice



(x_A, x_B, y) sampled

Alice interacts optimally with C_1

Alice computers her posterior and takes optimal action *a*

$I^*(C_1) \in \Delta(A \times Y)$ C_1

$I^*(C_2) \in \Delta(A \times Y)$ C_2

$I^*(C_3) \in \Delta(A \times Y)$ C_3

Picks provider $i^*(C_1, \ldots C_k) = \arg\max \mathbb{E}_{I^*(C_i)}[u_A(a, y)]$

"What medication should I prescribe?"

AI Companies: Bobs

Upshot:

This defines a *game* between the Bobs, where their actions are conversation rules

What are the Nash equilibria of this game?

Traditional Alignment (e.g. defining OPT)

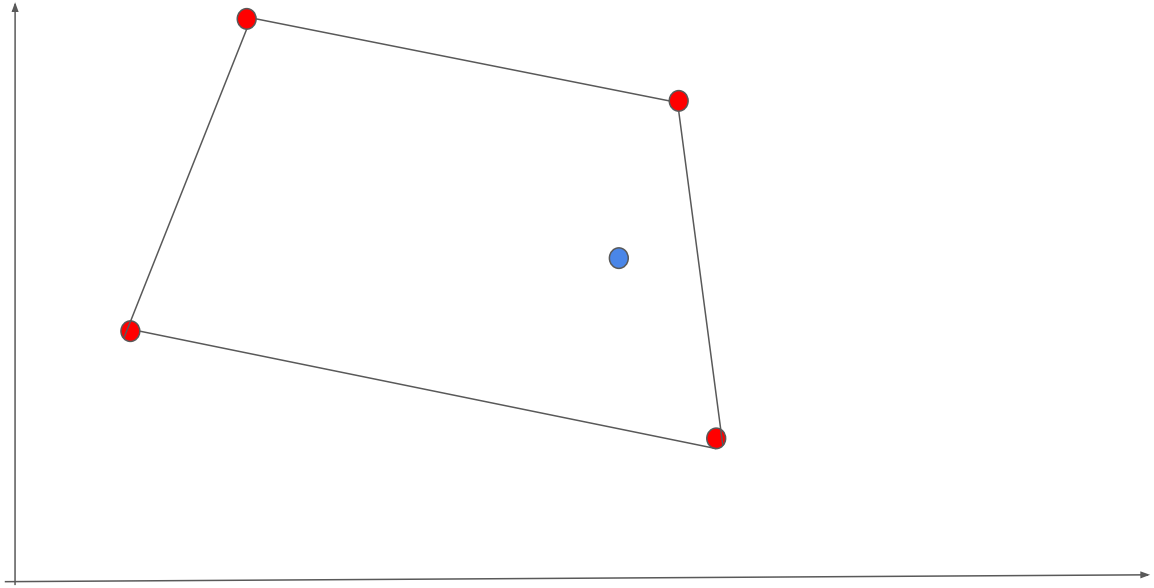
"Bob is Aligned with Alice" if: $u_B(a, y) \approx u_A(a, y)$ for all $a \in A, y \in Y$

- i.e. Bob wants and is optimizing for exactly the same thing as Alice.
- This is what would result in payoff OPT for Alice.

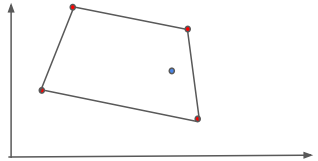
Market Alignment



Alice’s utility can be written as a non-negative linear combination of the Bobs’ utilities



Market Alignment



Alice's utility can be written as a non-negative linear combination of the Bobs' utilities

There exist coefficients $\alpha_1, \ldots, \alpha_k, c \geq 0$ such that for all $a \in A, y \in Y$:

$$u(a, y) \approx \alpha_1 u_1(a, y) + \cdots + \alpha_k u_k(a, y) + c$$

Maybe every AI provider is *substantially misaligned,* but collectively their interests *span* Alice's interests.

e.g. the J.J and Pfizer AI assistants…

Market Alignment Implies Good Outcomes in Equilibrium

Theorem (Informal): Suppose the Bob's satisfy the market alignment condition relative to Alice. Then in *every* Nash equilibrium, Alice gets utility at least OPT --- i.e. the full benefits of perfect alignment.

Proof sketch



$u_A(a, y)$

c_1

$u_{B_1}(a, y)$

c_2

$u_{B_2}(a, y)$

c_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \mathrm{OPT}$$

Proof sketch



$u_A(a, y)$

c_1

$u_{B_1}(a, y)$

c_2

$u_{B_2}(a, y)$

c_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \mathrm{OPT}$$

Proof sketch



$u_A(a, y)$

c_1

$u_{B_1}(a, y)$

c_2

$u_{B_2}(a, y)$

c_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \text{OPT}$$

Any Bob *could* deviate to playing C^* and doing so would result in Alice choosing them, as:

$$\mathbb{E}_{I^*(C^*)}[u_A(a, y)] = \text{OPT}$$

Proof sketch



$u_A(a, y)$

c_1

$u_{B_1}(a, y)$

c^*

$u_{B_2}(a, y)$

c_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \text{OPT}$$

Any Bob *could* deviate to playing C^* and doing so would result in Alice choosing them, as:

$$\mathbb{E}_{I^*(C^*)}[u_A(a, y)] = \text{OPT}$$

Proof sketch



$u_A(a, y)$

c_1

$u_{B_1}(a, y)$

c^*

$u_{B_2}(a, y)$

c_3

$u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \text{OPT}$$

Any Bob *could* deviate to playing C^* and doing so would result in Alice choosing them, as:

$$\mathbb{E}_{I^*(C^*)}[u_A(a, y)] = \text{OPT}$$

Proof sketch



$u_A(a, y)$

c_1 $u_{B_1}(a, y)$

c_2 $u_{B_2}(a, y)$

c_3 $u_{B_3}(a, y)$

Consider any Nash equilibrium. WLOG assume Alice chooses Bob 1: $i^*(C_1, \ldots, C_k) = 1$. Assume for contradiction that Alice is not doing well:

$$\mathbb{E}_{I^*(C_1)}[u_A(a, y)] < \text{OPT}$$

Any Bob *could* deviate to playing C^* and doing so would result in Alice choosing them, as:

$$\mathbb{E}_{I^*(C^*)}[u_A(a, y)] = \text{OPT}$$

But since they are at equilibrium, none of them want to…

Proof Sketch

Nash conditions:

$$\mathbb{E}_{I^*(C_1)}[u_1(a,y)] \geq \mathbb{E}_{I^*(C^*)}[u_1(a,y)], \quad \dots, \quad \mathbb{E}_{I^*(C_1)}[u_k(a,y)] \geq \mathbb{E}_{I^*(C^*)}[u_k(a,y)]$$

Extensions:

Can extend to models in which Alice interacts with multiple LLM providers simultaneously, rather than choosing the best one.

Can extend to models of user bounded rationality (quantal response rather than best response on Alice's part)

Can extend to markets with multiple heterogeneous downstream users (under different market alignment assumptions–and model *personalization* helps here, even when the models are self-interested)

Takeaways:

Multiple misaligned AI providers can still be incentivized to train AI that provides aligned *outcomes* for users–if their incentives are sufficiently varied.

Market alignment is much easier to satisfy than alignment with any one AI provider–and becomes increasingly easy to satisfy if the market for AI systems is large and diverse!

Thanks!

Tractable Agreement Protocols
Natalie Collina, Surbhi Goel, Varun Gupta, Aaron Roth (STOC'25)

Collaborative Prediction: Tractable Information Aggregation via Agreement
Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, Mirah Shi (SODA'26)

Emergent Alignment via Competition
Natalie Collina, Surbhi Goel, Aaron Roth, Emily Ryu, Mirah Shi (ICML'26)

Personalization Aids Pluralistic Alignment via Competition
Natalie Collina, Surbhi Goel, Aaron Roth, Mirah Shi (in submission)