Data Science: Probabilistic and Optimization Methods II
ICTS Bengaluru

August 2025

COLLABORATIVE PREDICTION

VIA TRACTABLE AGREEMENT PROTOCOLS

Surbhi Goel
University of Pennsylvania

Based on joint works with:

Natalie Collina
UPenn

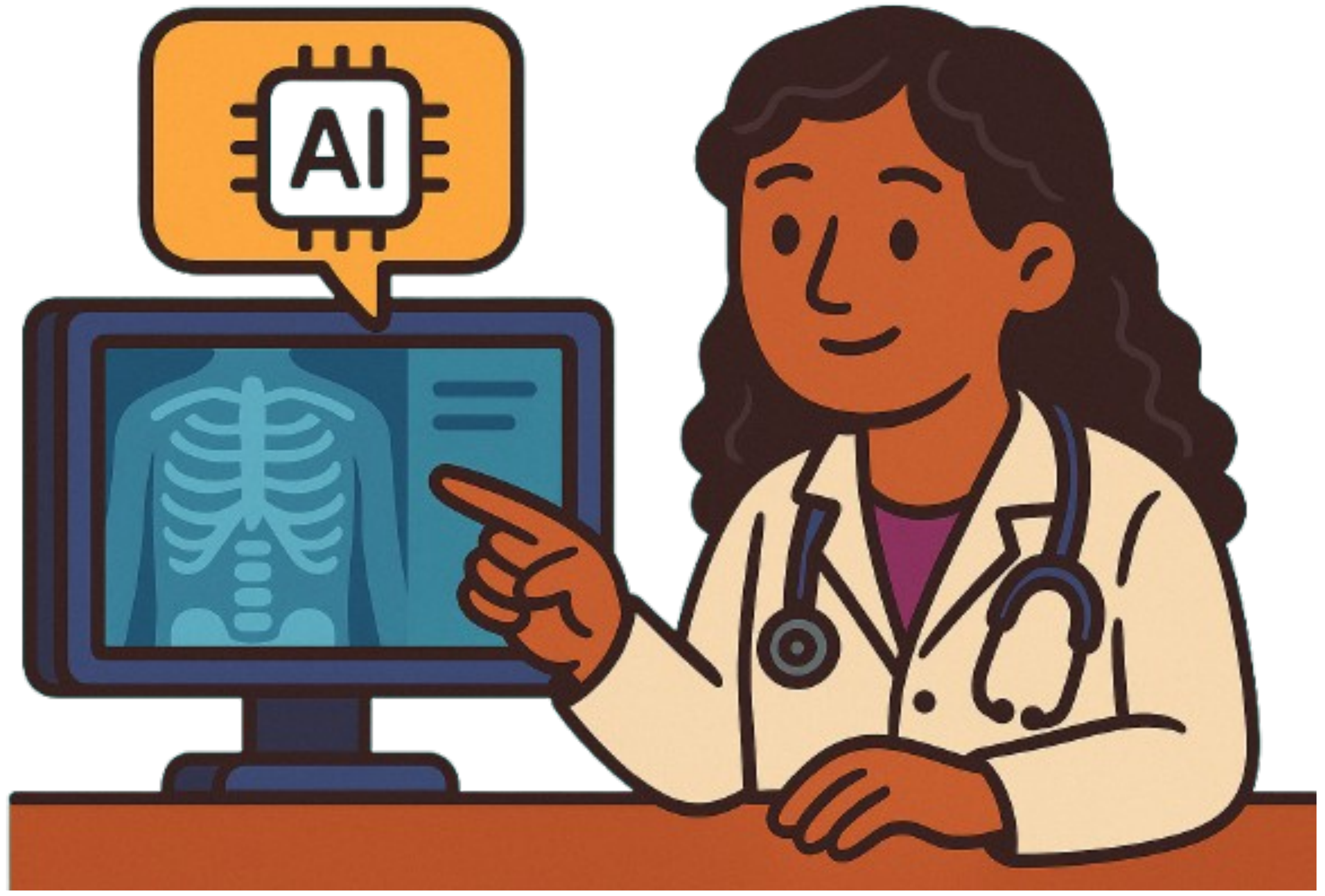
Ira Globus-Harris
UPenn → Cornell

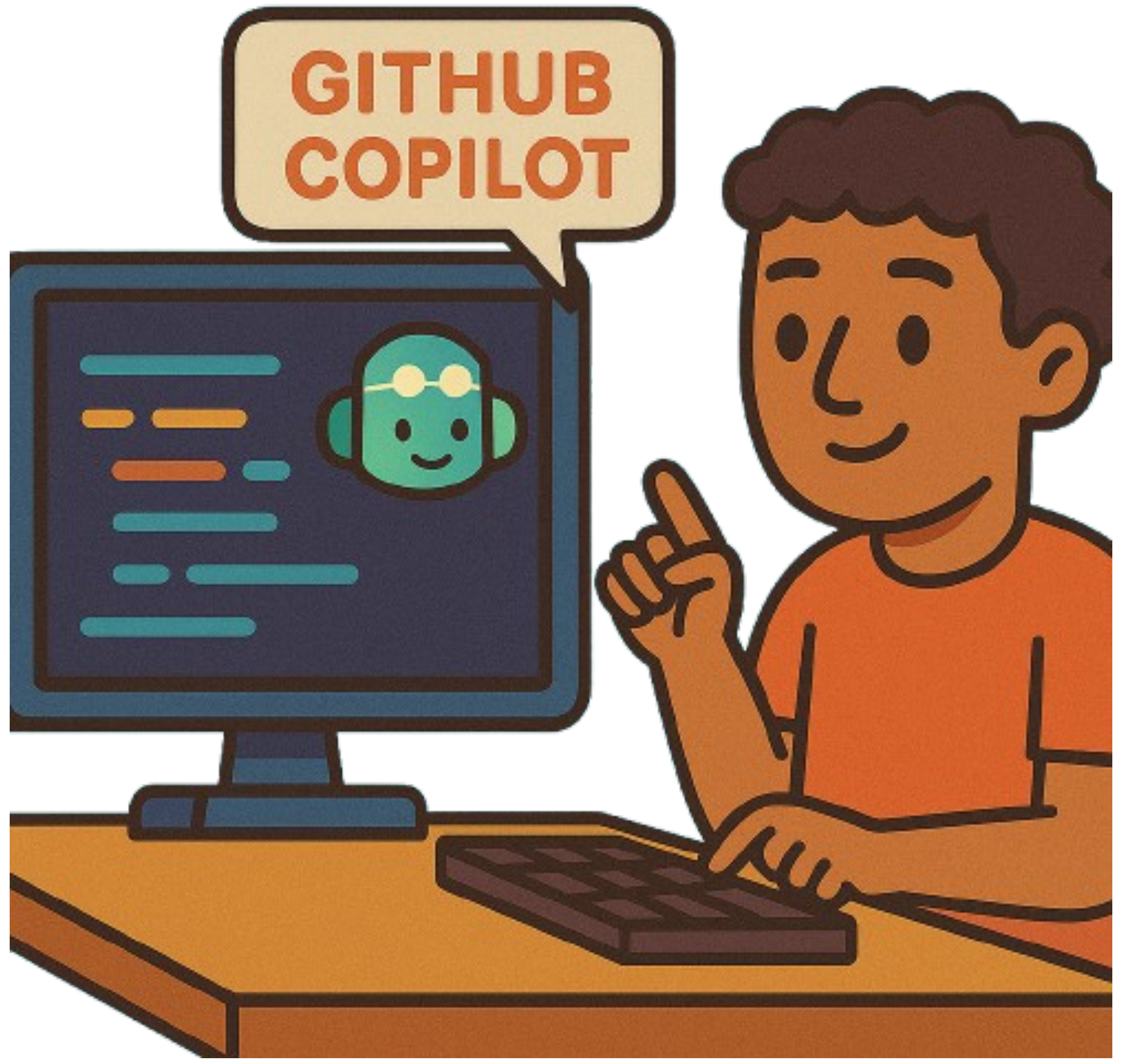
Varun Gupta
UPenn → Vector Institute

Aaron Roth
UPenn

Mirah Shi
UPenn

HOW DO WE USE AI SYSTEMS TODAY?

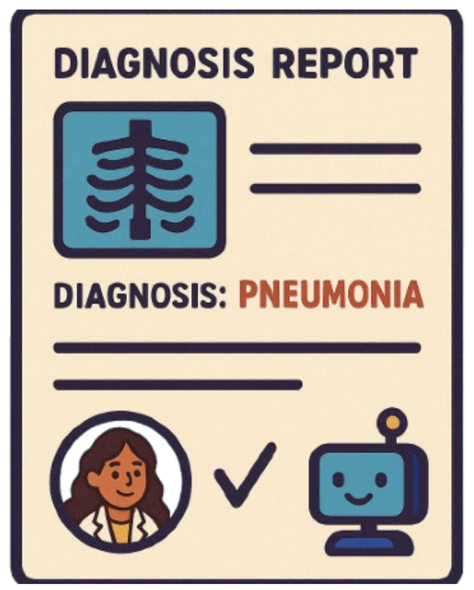




Increasingly, we are interacting with AI agents to do tasks and make decisions

Note: All visuals in this talk are created in collaboration with GPT

EXAMPLE: DOCTOR USING AI SYSTEM



Patient disposition

Patient appearance

Relevant context

Medical history

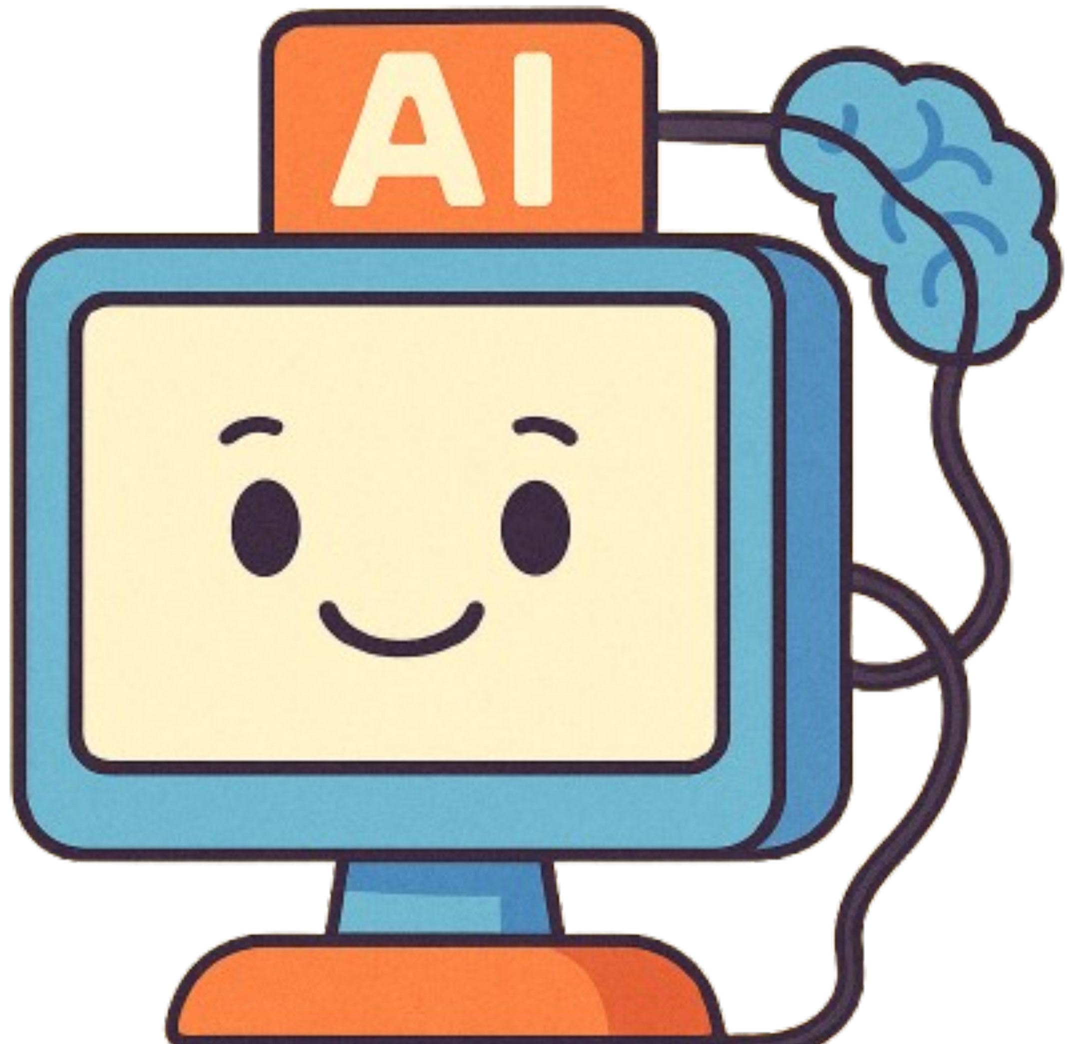
Family history

Lab reports

Patient records from other hospitals

Medical Journals

Pretrained information



GOALS: WHAT DO WE DESIRE FROM THIS INTERACTION?

- **Complementarity**
  - The interaction leverages the complementary skills of the AI and the human
- **Agreement**
  - The doctor and AI model reach consensus on the decision
- **Accuracy**
  - The end outcome for the patient is positive
- **Information Aggregation**
  - Outcome is as good as if they both had each other's complete information

We want the team to improve over either human or AI working alone

REALITY: HUMANS USING AI SYSTEMS

Agree to disagree: the symmetry of burden of proof in human–AI collaboration

Karin Rolanda Jongsma # 1, Martin Sand # 2

Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review

Clara Cestonaro [1], Arianna Delicati [1], Beatrice Marcante [1], Luciana Caenazzo [1], Pamela Tozzo [1,*]

A.I. Chatbots Defeated Doctors at Diagnosing Illness

A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.

AI slows down some experienced software developers, study finds

By **Anna Tong**

July 10, 2025 7:31 PM GMT+5:30 · Updated July 10, 2025

HEALTH · AI · FEATURES

Google's healthcare AI made up a body part – what happens when doctors don't notice?

Google dubbed an error from its Med-Gemini model a typo. Experts say it demonstrates the risks of AI in medicine.

PRESS

Humans and AI: Do they work better together or alone?

by MIT Sloan Office of Communications | Oct 28, 2024

Incorrect AI Advice Influences Diagnostic Decisions

System developers must consider how AI explanation might impact reliance on AI advice

India's Apollo Hospitals bets on AI to tackle staff workload

By **Rishika Sadam**

March 13, 2025 5:11 PM GMT+5:30 · Updated March 13, 2025

Aa

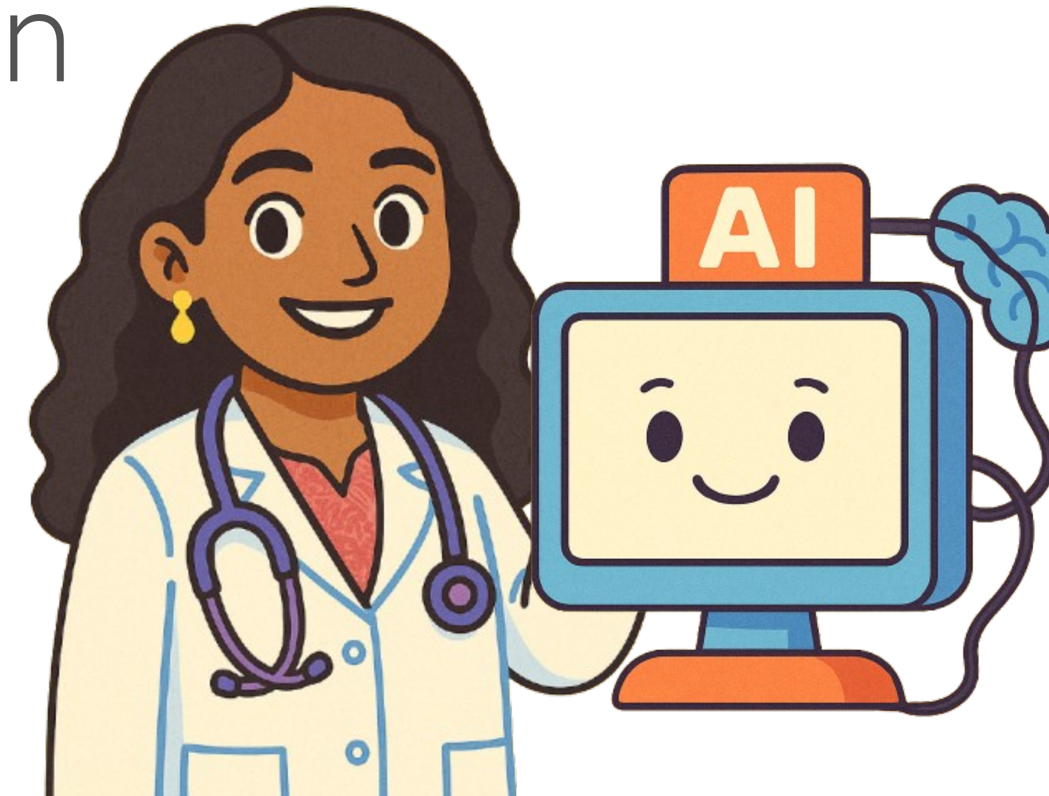
These systems are already being used,

while achieving these goals remains challenging!

Can we design systems that guarantee that humans make better decisions when using AI?

Roadmap:

- Collaboration via Bayesian Agreement Protocols
- Show how to relax 'Bayesian' assumptions to make these protocols tractable using calibration
- Show when such agreement protocols provably lead to information aggregation



Part 1:
Bayesian Agreement Protocols

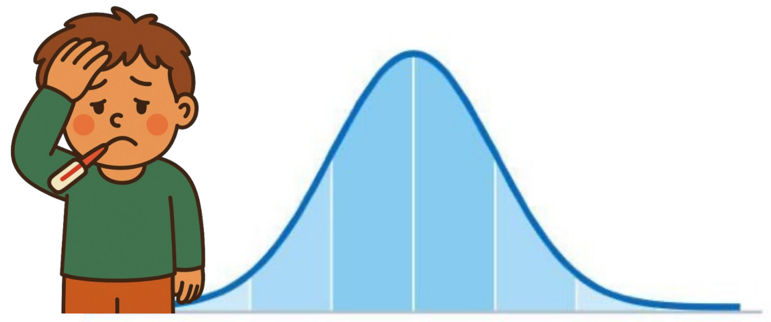
[Aumann'76, Geanakoplos-Polemarchakis'82, Aaronson'05]

AGREEMENT

Theorem [Aumann'76]

If two Bayesian agents have a shared prior and *common* knowledge of each other's posterior expectation, the posterior expectation will be the same.

Bayesian agents with common knowledge cannot **agree to disagree**



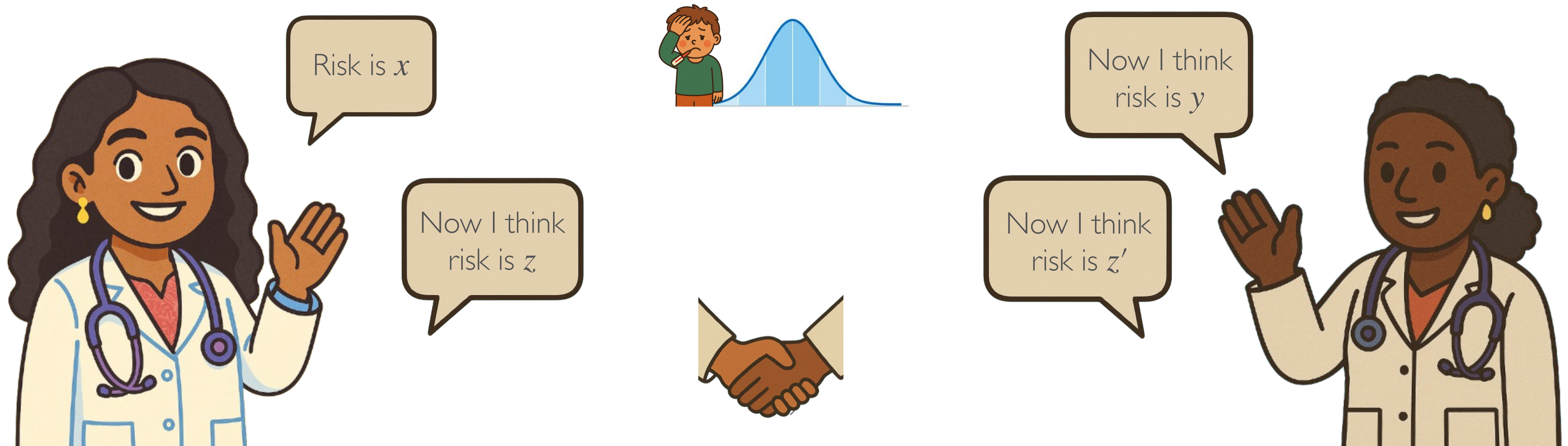
…

AGREEMENT

Theorem [Geanakoplos-Polemarchakis'82]

If the underlying state space is *finite*, agreement happens in a finite number of rounds, if each agent shares the expectation in each round.

Bayesian agents **agree** in finite time

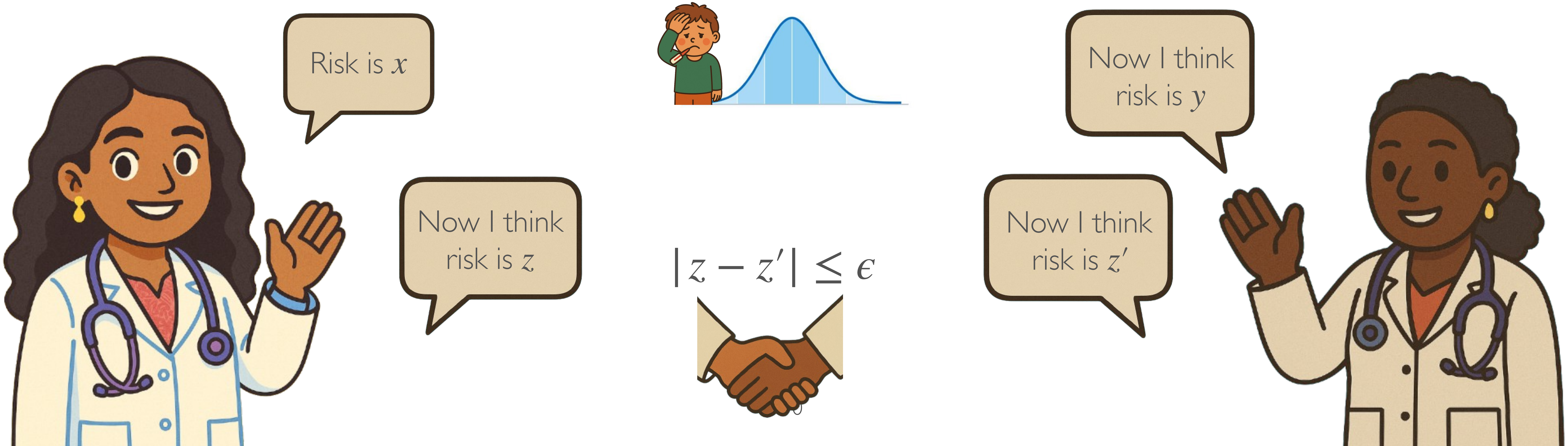


AGREEMENT

Theorem [Aaronson'05]

If each agent shares their posteriors at each round, for scalar predictions, with probability $1 - \delta$, they reach ϵ-agreement in at most $\dfrac{1}{\epsilon^2 \delta}$ rounds.

Bayesian agents agree in #rounds **independent** of state size!



AGREEMENT: WHY IS THIS A GOOD FRAMEWORK?

- **Guaranteed Agreement**
  - Shows that interacting over rounds will lead to consensus quickly, independent of size of features each agent has
- **Sharing only Predictions**
  - The protocol requires only sharing predictions bypassing the need to directly share or translate potentially incompatible raw features or explanations
- **Accuracy improving**
  - Since the protocol is only information revealing, the final predictions will be better than either agents starting predictions

OTHER APPROACHES TO COLLABORATION

- **Vertically Federated Learning**
  - Use techniques like homomorphic encryption [Hardy et al.'17] to jointly train one model on combined features without revealing the raw data
  - Requires cryptographic overhead, and compatible features
- **Explanations**
  - AI provides an "explanation" for its reasoning to help the human
  - Explanations can often be complex and even misleading [Bansal et al.'21, Goh et al.'24]
- **Multi-modal Learning** [Hardy et al.'17]
  - Combine different data types either by merging features at the start ("early fusion") or by averaging final predictions ("late fusion")
  - Requires either feature alignment or provides simple averaging which is insufficient

AGREEMENT: WHAT ARE THE LIMITATIONS?

- **Bayesian Rationality**
  - Humans/AI models do not behave like bayesian rational agents
  - It is intractable to implement posterior calculations over complex state spaces and long interaction histories
- **Common Priors**
  - Unclear where a common prior would come from for a human and AI model given their different training data and experience

Can we relax these assumptions while still guaranteeing fast agreement?

Part 2:
Tractable Agreement Protocols

Natalie Collina
UPenn

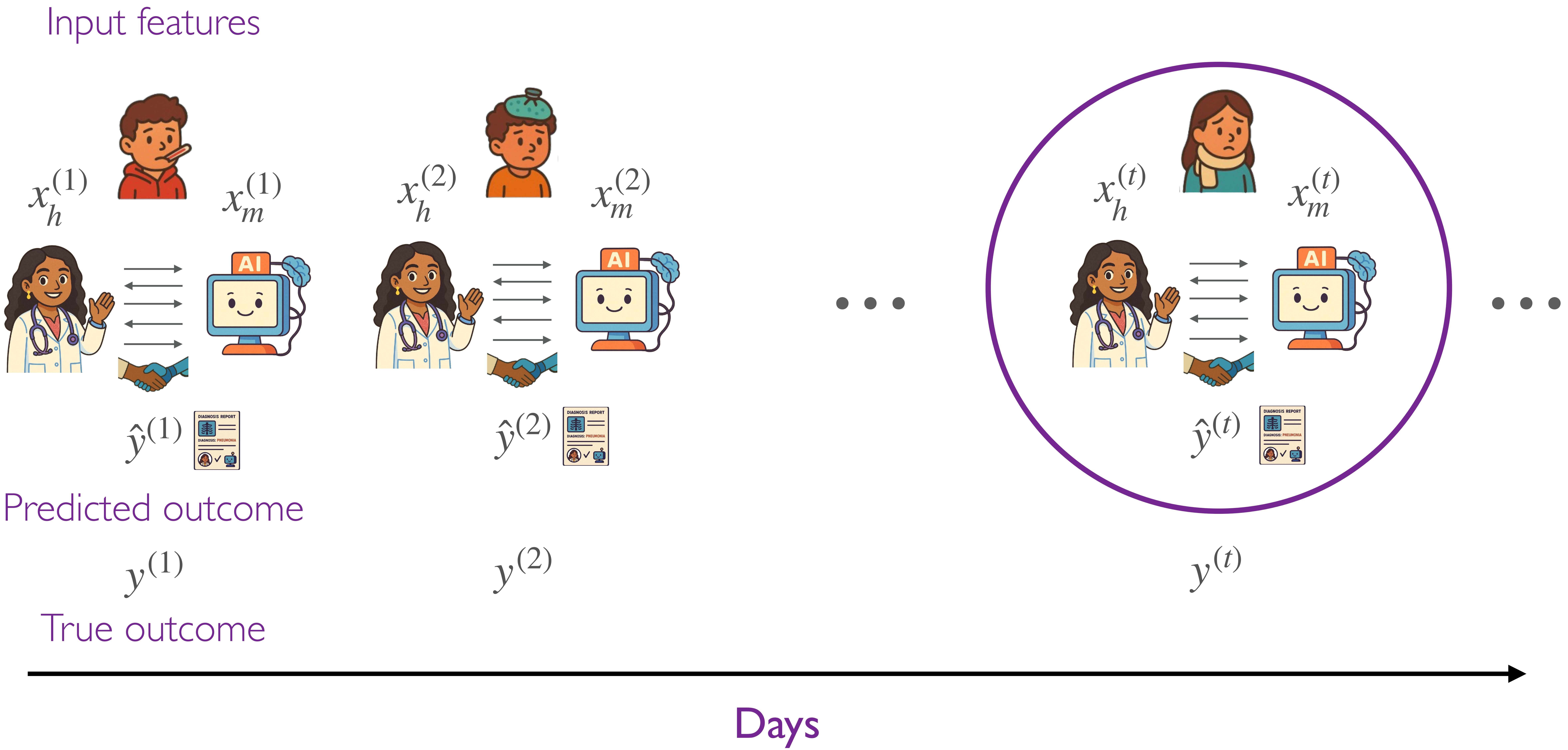
Varun Gupta
UPenn → Vector Institute

Aaron Roth
UPenn

TRACTABLE AGREEMENT PROTOCOLS: MAIN RESULTS

- We move to a **repeated setting** to remove the assumption of priors
- We introduce a new notion of calibration we call **conversation calibration**
  - Satisfied by Bayesians, but strictly weaker
  - Enforceable computationally efficiently on base model without loss of accuracy
- If agents satisfy **conversation calibration** then they reach **fast agreement**
  - The longer the conversation goes, the more accurate the prediction
- Can recover the same rates as [Aaronson'04] in one-shot Bayesian setting
- Extends beyond 1-dimensional setting to multi-dimensional and action feedback

SETUP



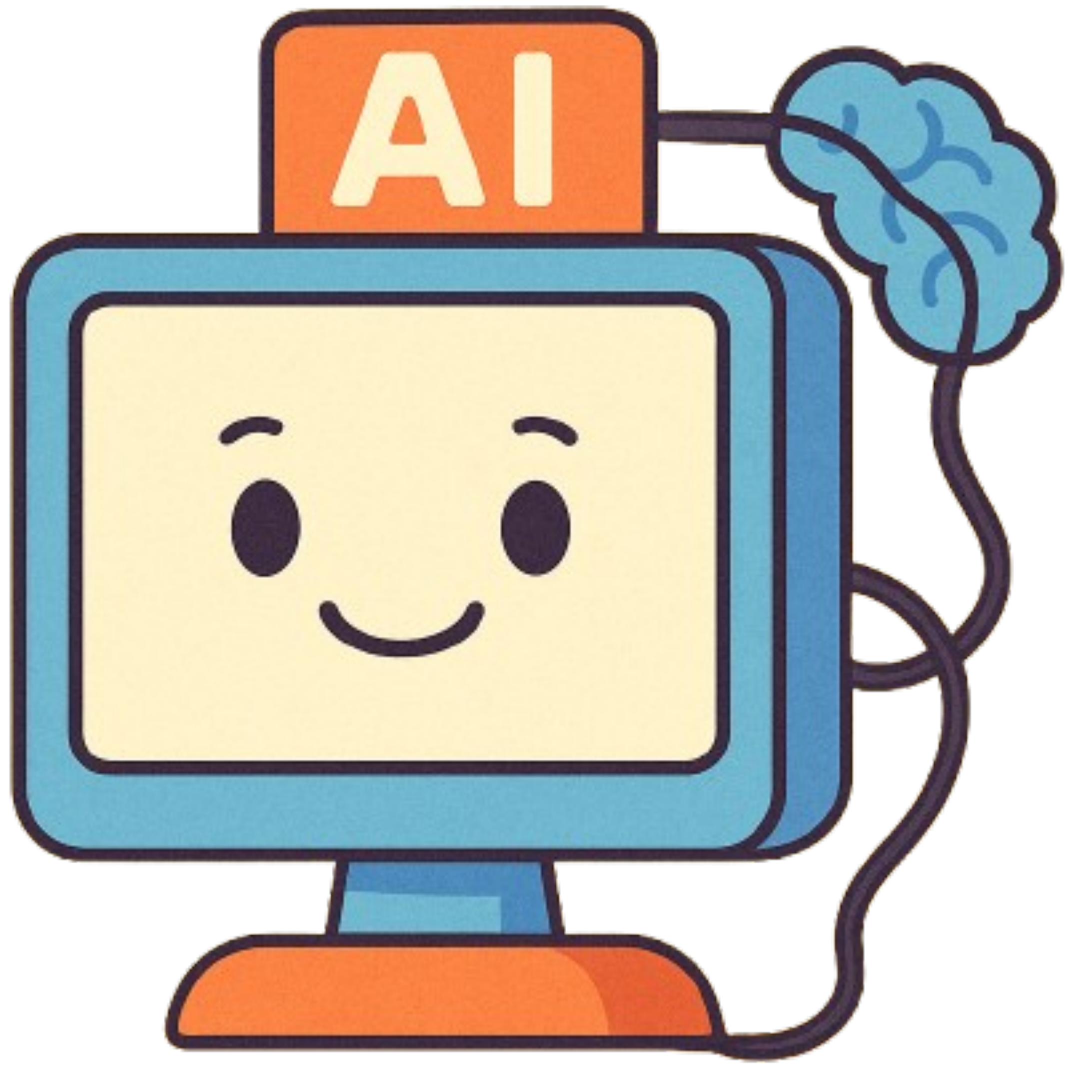
Input features
$x_h^{(1)}$ $x_m^{(1)}$
$x_h^{(2)}$ $x_m^{(2)}$
$x_h^{(t)}$ $x_m^{(t)}$
AI
$\hat{y}^{(1)}$ $\hat{y}^{(2)}$ $\hat{y}^{(t)}$
Predicted outcome
$y^{(1)}$ $y^{(2)}$ $y^{(t)}$
True outcome
Days

SETUP

Rounds

$x_h^{(t)}$

$x_m^{(t)}$



$\hat{y}_h^{(t)}(1) \longrightarrow$

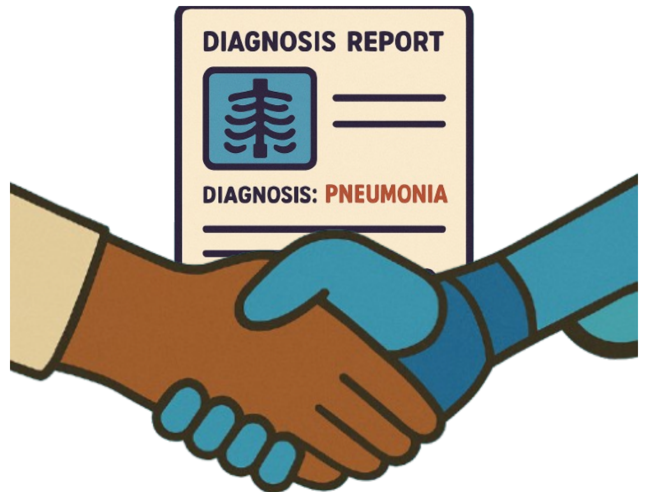
$\longleftarrow \hat{y}_m^{(t)}(2)$

$\hat{y}_h^{(t)}(3) \longrightarrow$

$\vdots$

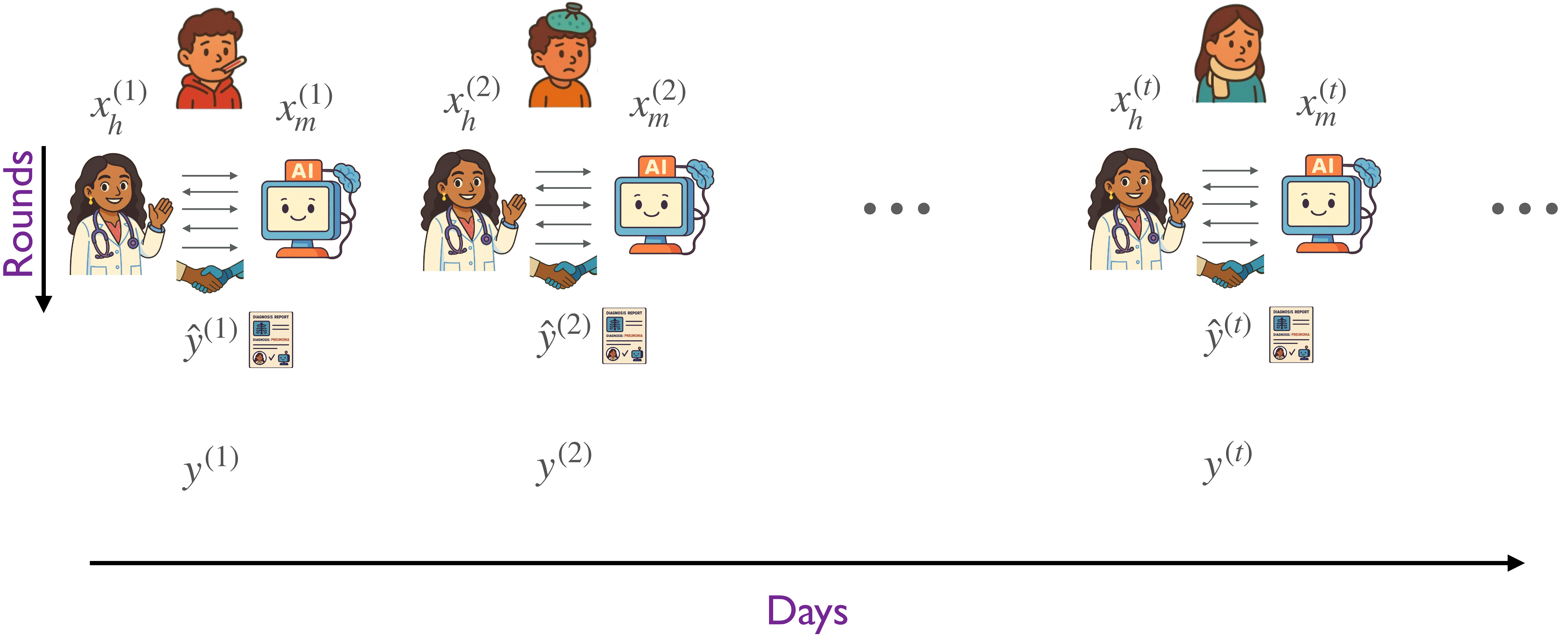
$\longleftarrow \hat{y}_m^{(t)}(k)$

$$|\,\hat{y}_m^{(t)}(k) - \hat{y}_h^{(t)}(k-1)\,| \leq \epsilon$$



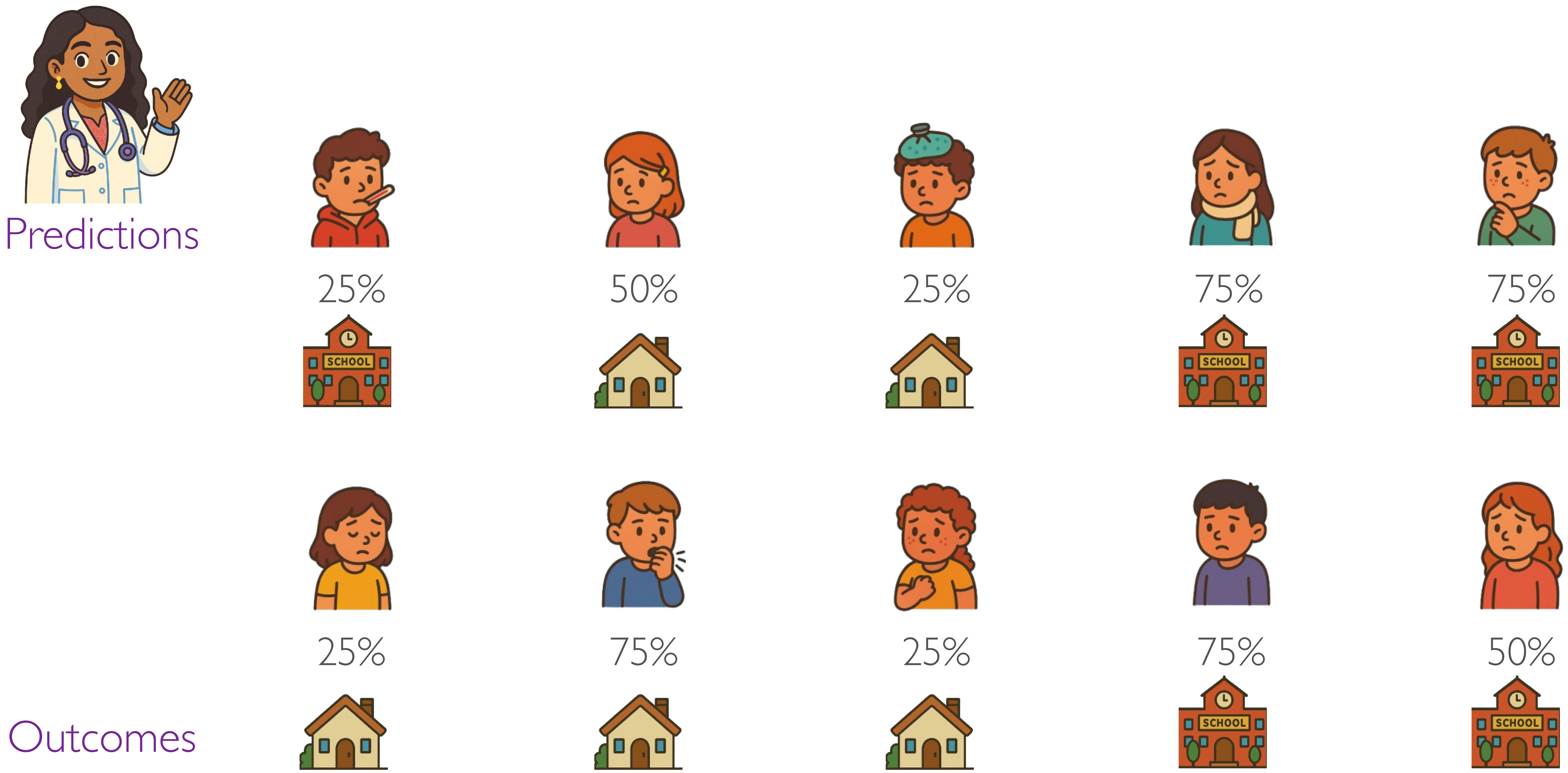
SETUP

Goal: We want $1 - \delta$ fraction of the days achieve ϵ agreement in few rounds



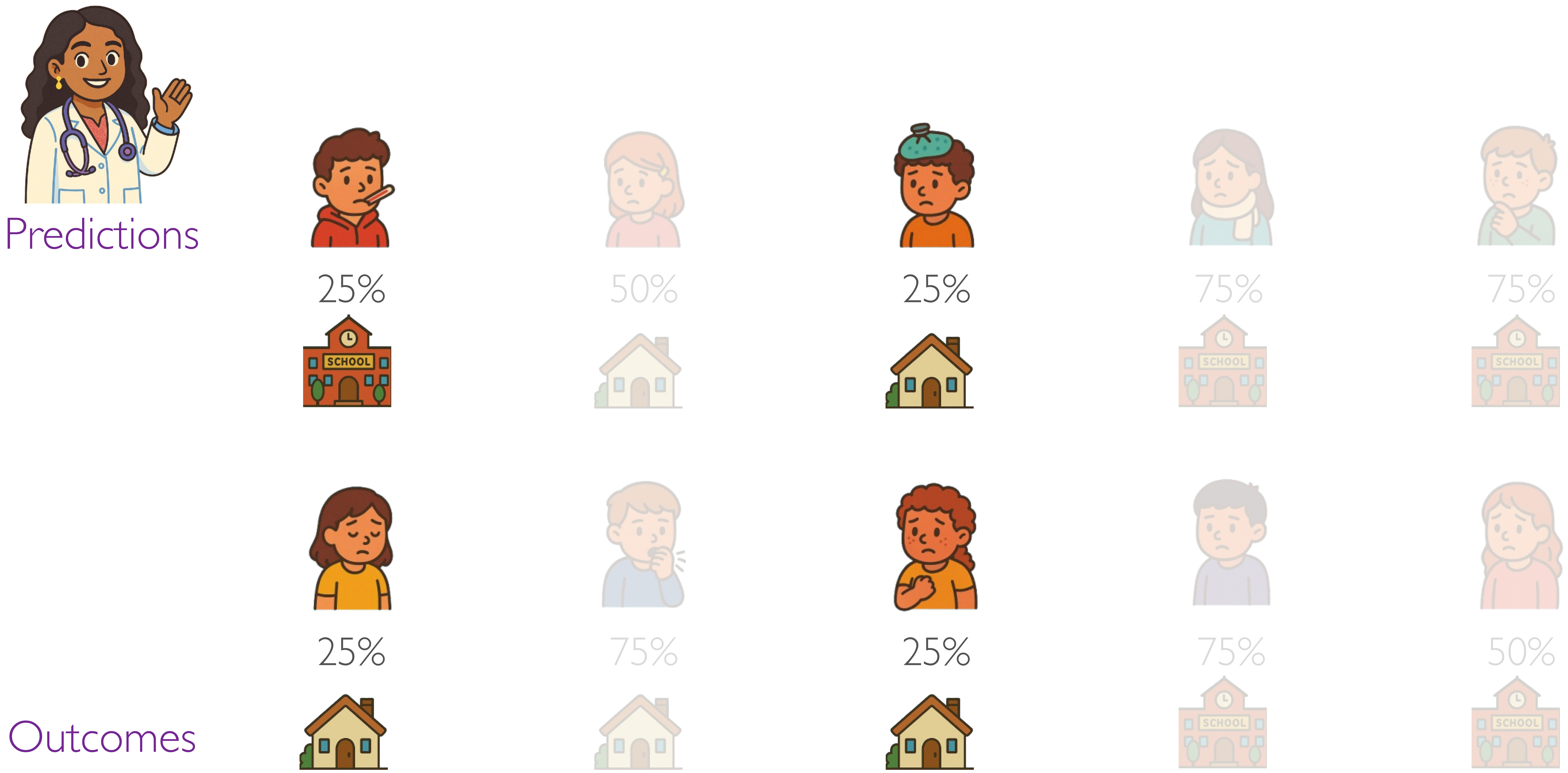
CALIBRATION [Dawid'82]

Predictions should "mean what they say"



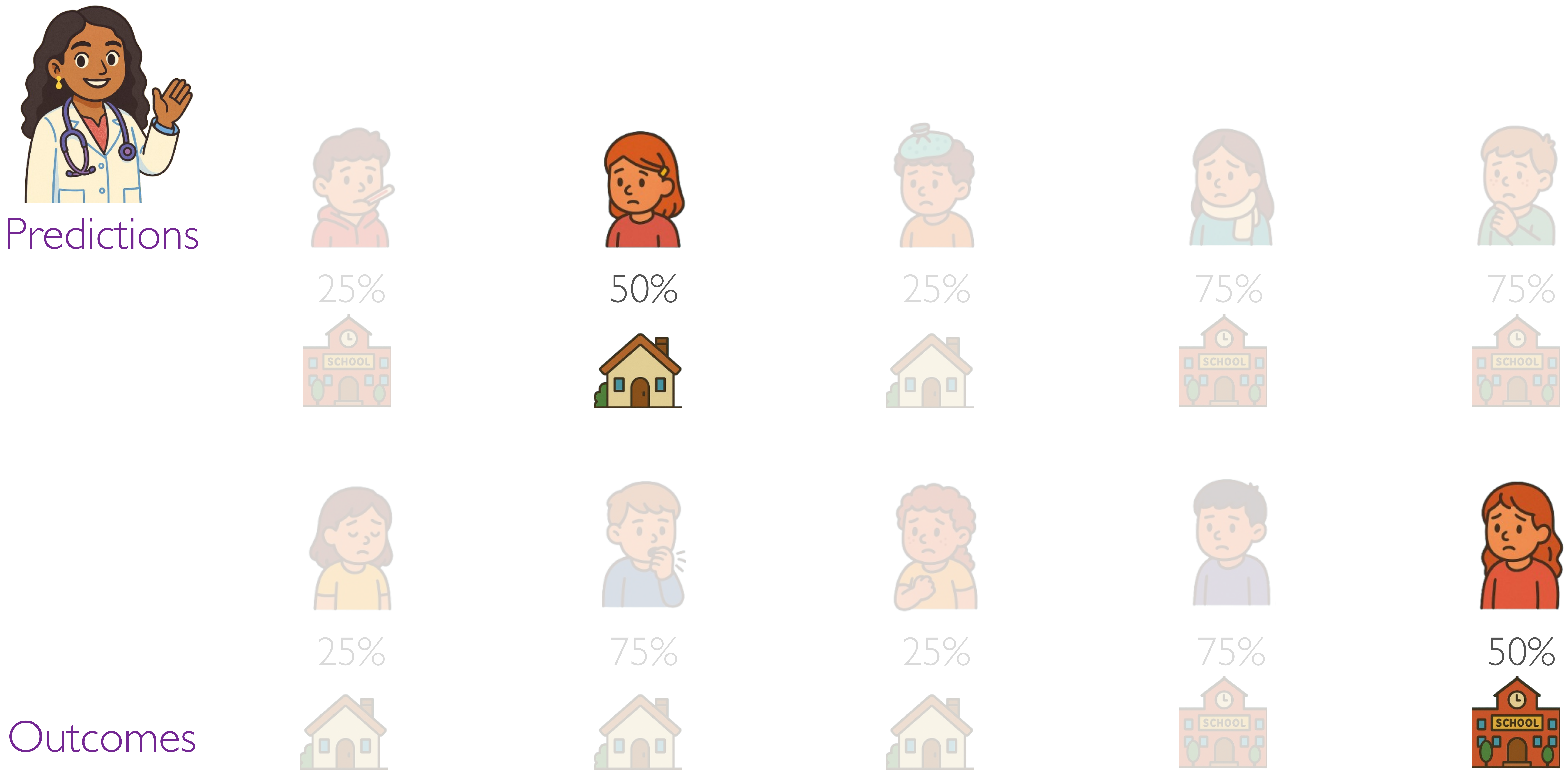
CALIBRATION [Dawid'82]

Predictions should "mean what they say"



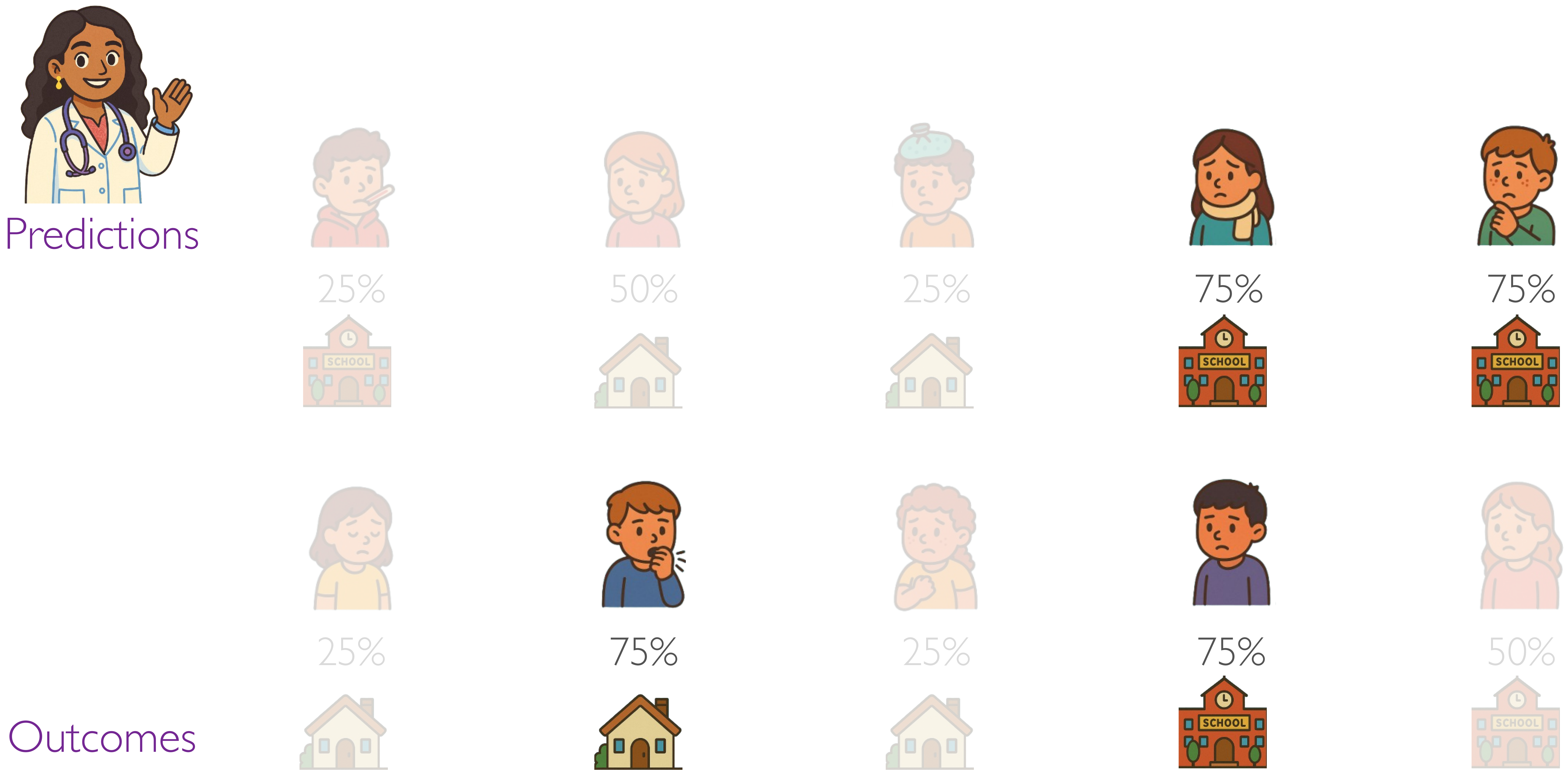
CALIBRATION [Dawid'82]

Predictions should "mean what they say"



CALIBRATION [Dawid'82]

Predictions should "mean what they say"



CONVERSATION CALIBRATION

- **Calibration:** Predictions should be unbiased conditional on the prediction itself.

  For all $p \in [0,1]$, $\sum_{t=1}^{T} \mathbb{I}[\hat{y}_m^t = p](p - y^t) = 0.$

  Over many days

  when AI predicts p

  AI's predictions are correct on average

- **Conversation Calibration:** Predictions should be unbiased conditional on the predictions of the other agent in the previous round. For AI, for all even rounds k, and $p, p' \in [0,1]$,

$$\sum_{t=1}^{T} \mathbb{I}[\hat{y}_m^{(t)}(k) = p]\mathbb{I}[\hat{y}_h^{(t)}(k-1) = p'](p - y^{(t)}) = 0.$$

Over many days

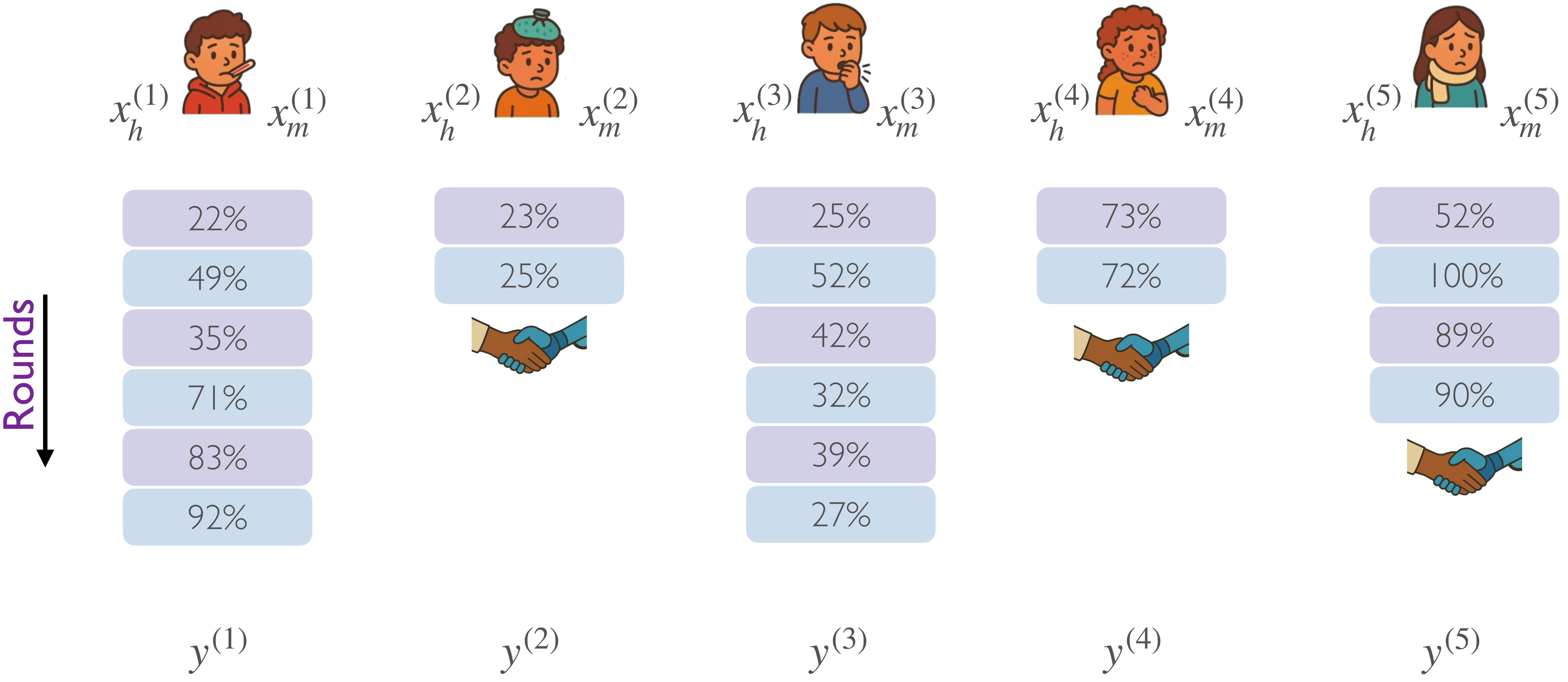
when AI predicts p

after human predicts p'

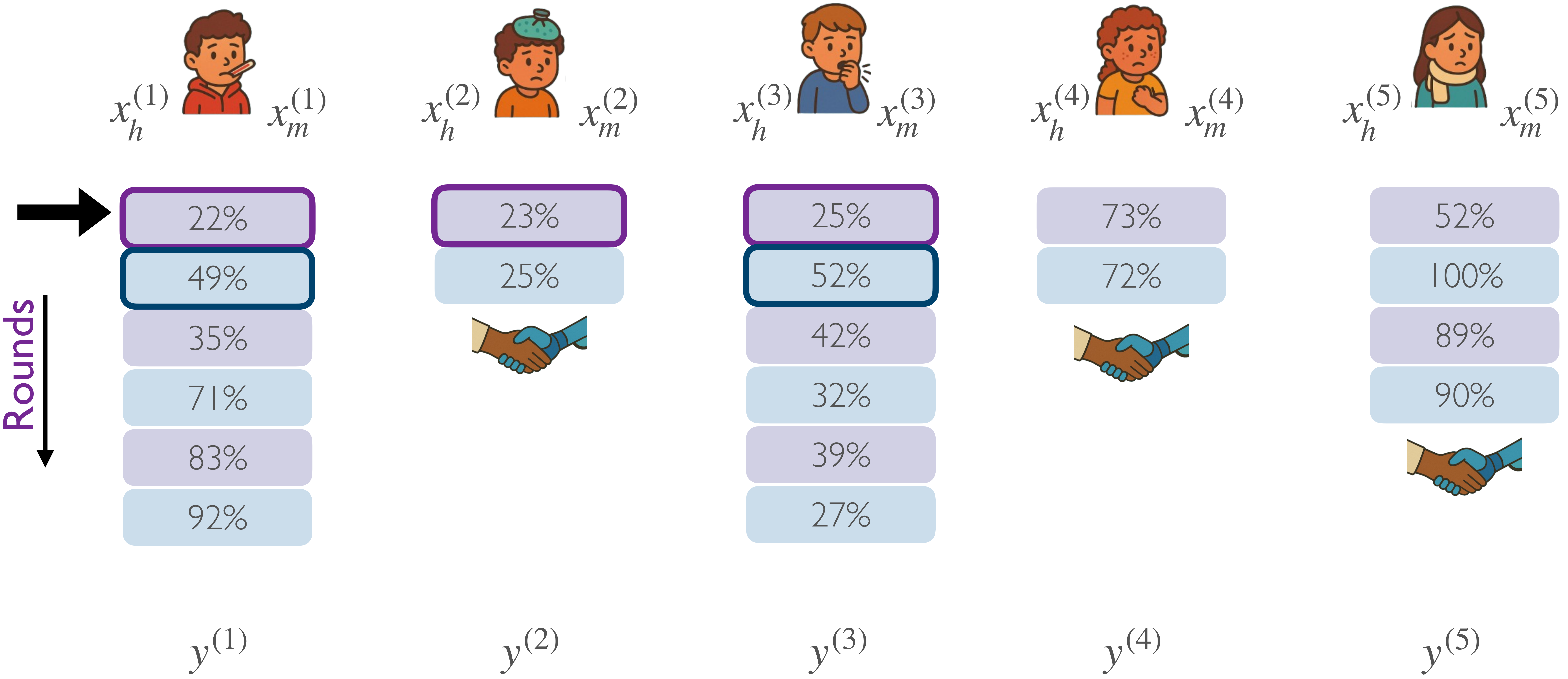
AI's predictions are correct on average

We relax this to approximate calibration and bucketing of the predictions of the other agent

CONVERSATION CALIBRATION



CONVERSATION CALIBRATION



If the AI is conversation calibrated then the expectation of outcome on days 1 and 3 should be roughly 50%

CONVERSATION CALIBRATION $\implies$ FAST AGREEMENT

Theorem [Colina-G-Gupta-Roth'24]

If both the human and AI are (approximately) conversation calibrated then on a $1 - \delta$ fraction of the days, they achieve ϵ-agreement after at most K rounds for

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

$\beta(T)$ goes to 0 as $T \to \infty$ for the appropriate choice of bucketing and distance to calibration [Blasiok et al.'23] for both predictors

Using prior work [Qiao-Zheng'24, Arunachaleswaran et al.'25], we can design efficient algorithms with $\beta(T) \approx T^{-1/3}$

CONVERSATION CALIBRATION $\implies$ FAST AGREEMENT

Proof sketch:

Consider the days on which we haven't reached agreement by round k, we know that the predictions at round k are

- at least ϵ far from predictions at round $k-1$, and
- calibrated conditional on the predictions at round $k-1$

Lemma

If a sequence 2 is calibrated conditional on sequence 1 then sequence 2 has lower (or equal) squared error than sequence 1.

Sequence 2 can make better predictions within the level sets of sequence 1

CONVERSATION CALIBRATION $\Longrightarrow$ FAST AGREEMENT

Two cases:

- Either $1 - \delta$ fraction of the rounds reach agreement, or 
- On at least δ fraction of the rounds, in round k, we improve upon the squared error by ϵ^2 *(since predictions were ϵ different from the predictions in round $k - 1$)*

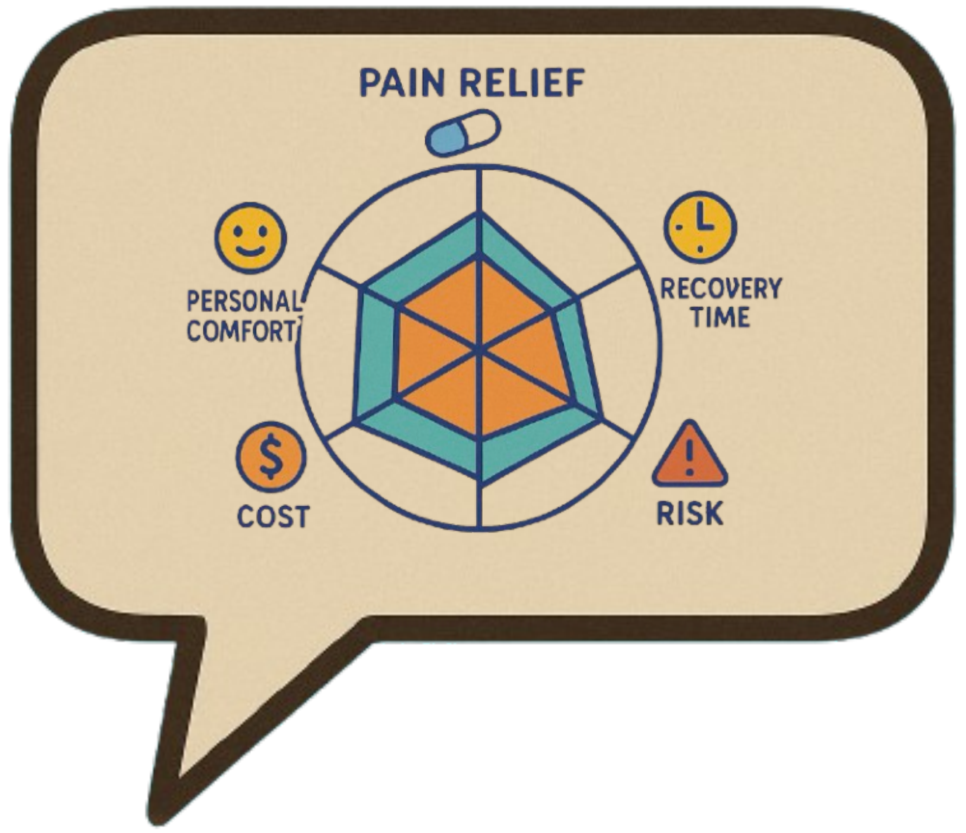
Till we reach case 1, at each round we decrease average squared error by $\epsilon^2\delta$

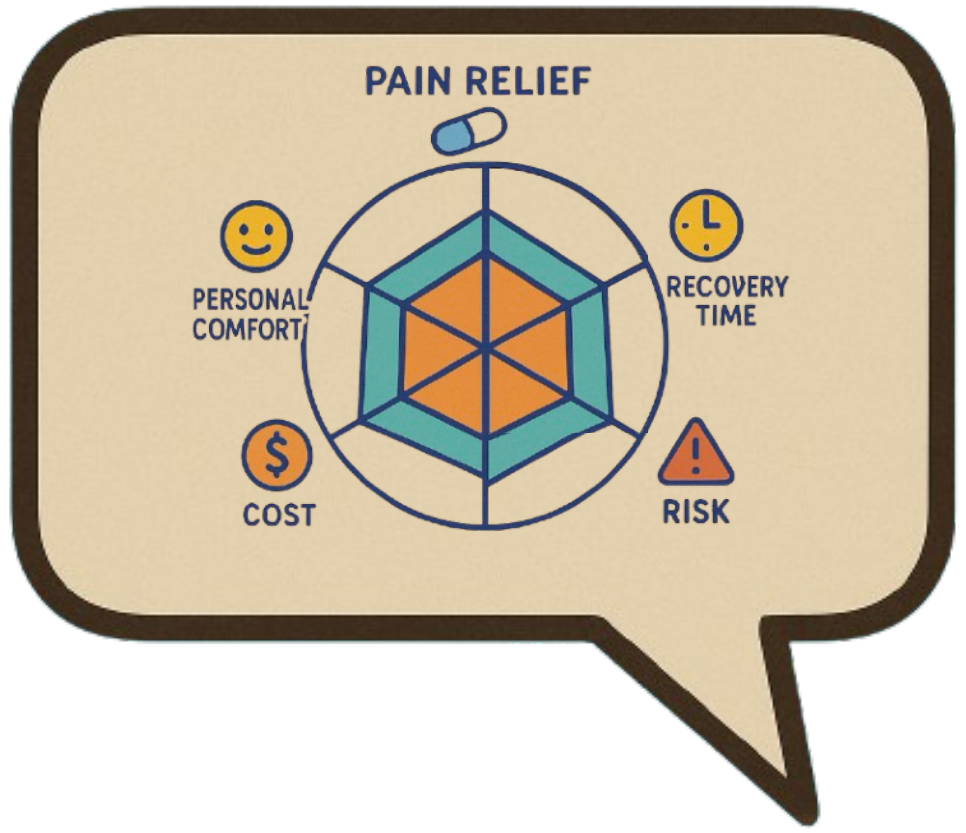
$$\text{Total \#rounds we can disagree} = \frac{\text{Max possible average squared error}}{\text{Decrease in average squared error at each round}}$$

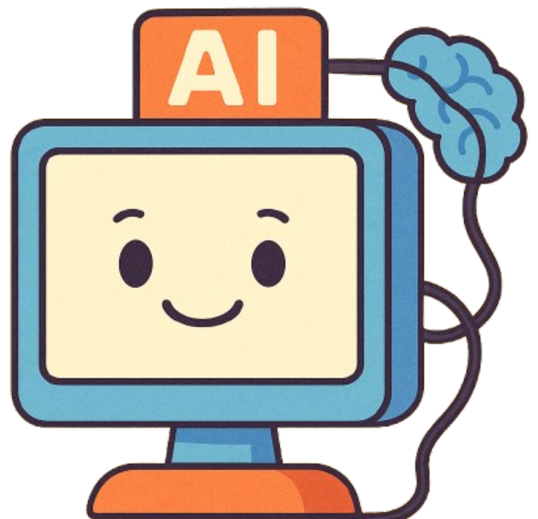
$$\approx \frac{1}{\epsilon^2\delta}$$

EXTENSIONS - MULTI-DIMENSIONAL

- Marginal conversation-calibration on each coordinate
- Agree when predictions on all dimensions within ϵ
- Guarantee that error in at least one dimension will go down by $\epsilon^2\delta/d$
- Total squared error is $d \implies$ agreement happens in $\approx \dfrac{d^2}{\epsilon^2\delta}$





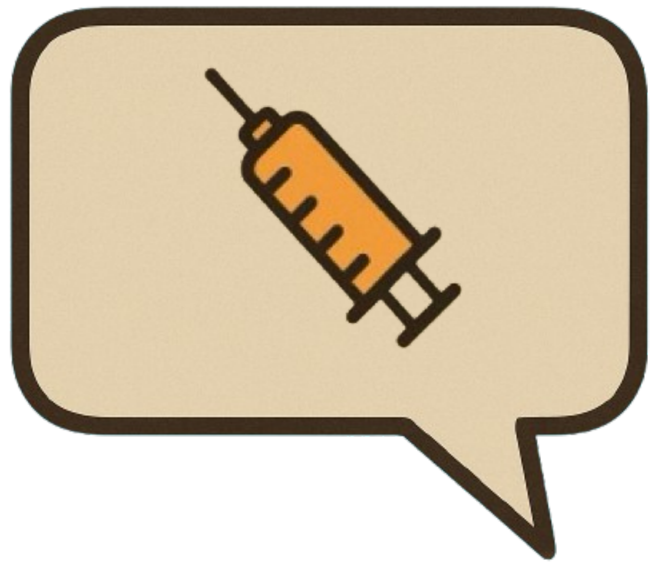
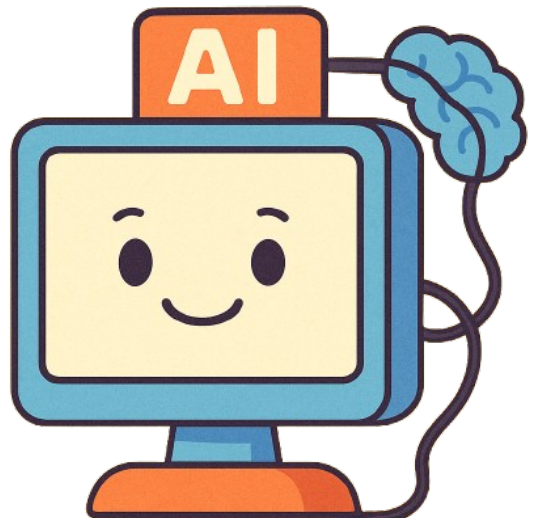


EXTENSIONS - ACTION FEEDBACK

- Extend to best-response action feedback via decision-conversation-calibration (defined based on utilities)
- If no agreement then the other party can improve utility by $\epsilon\delta$ Utility is linear
- So we get to agreement happens in $\approx \dfrac{1}{\epsilon\delta}$ rounds

Size of action set shows up in $\beta(T)$ so we need T to be large enough before this kicks in



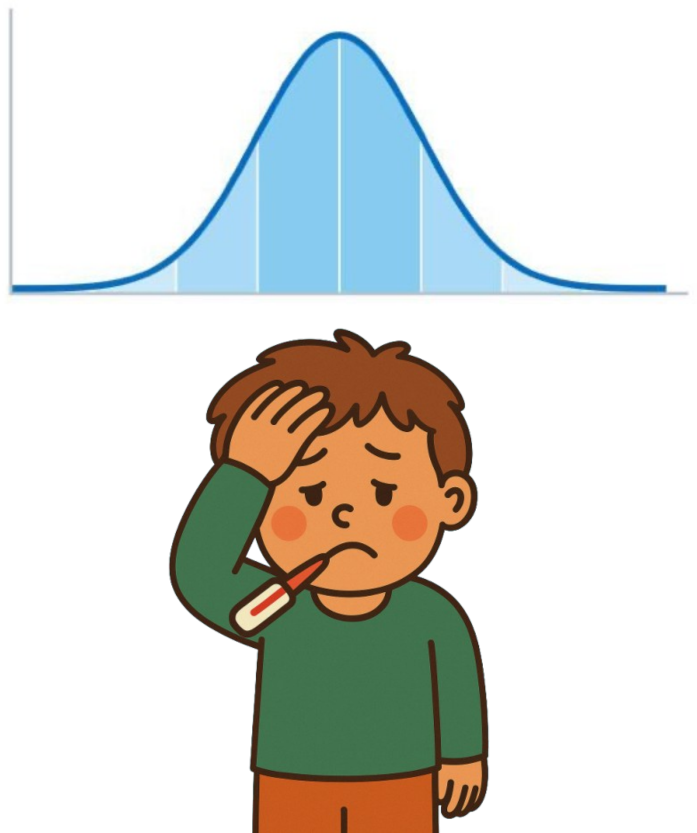




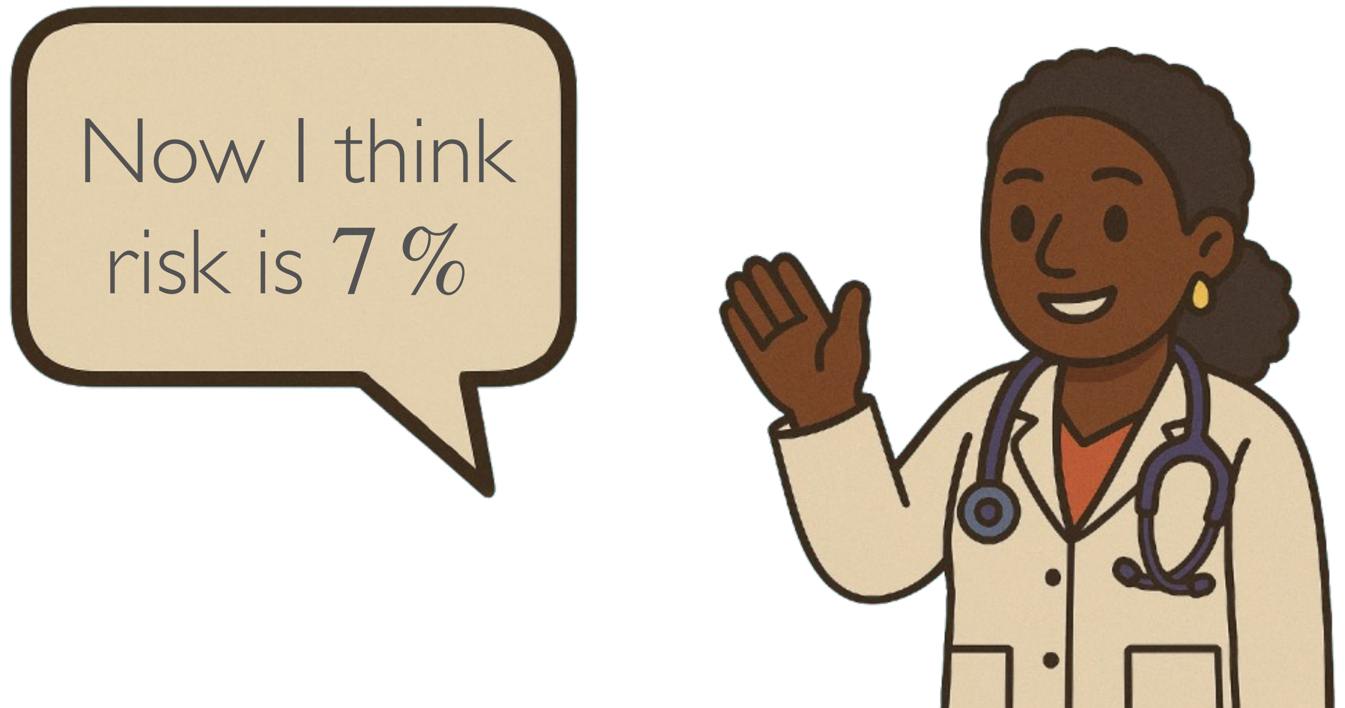
REDUCTION TO ONE-SHOT

- Assume first day is the observed state and the other days the observation is drawn i.i.d. from the prior
- Bayesians are approximately conversation-calibrated

$T \to \infty$

- By our theorem, $1 - \delta$ fraction rounds will reach agreement in $1/\epsilon^2\delta$ rounds
- But Bayesians don't need history of other rounds, so we can permute the rounds
- Therefore, probability first round reaches agreement in $1/\epsilon^2\delta$ rounds is $1 - \delta$





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 features?
  - Well, not always

When can we guarantee 'information aggregation' without sharing features?

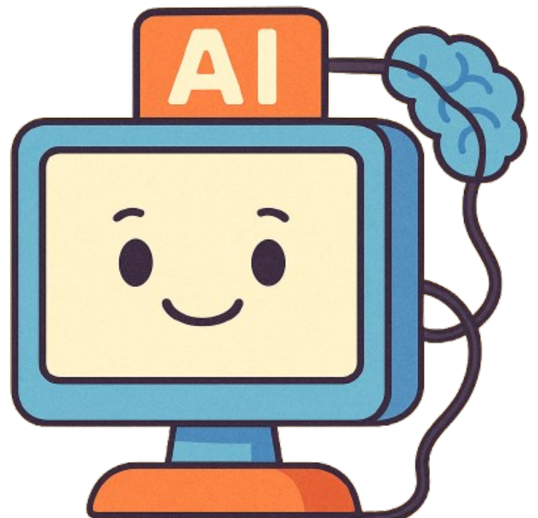
Risk is $50\,\%$

Risk is $50\,\%$

$$y = x_h + x_m \mod 2$$

$x_h \in \{0,1\}$

$x_m \in \{0,1\}$



Part 3:

Information Aggregation via Agreement





Natalie Collina
UPenn

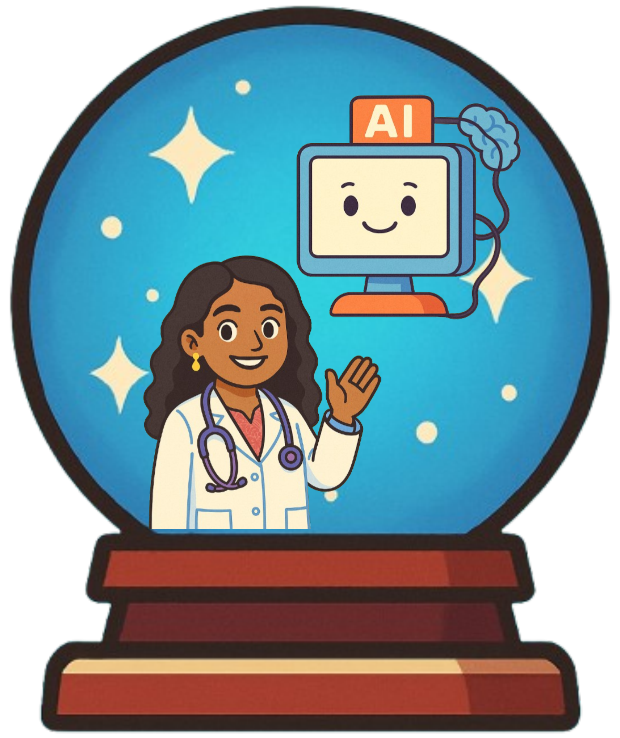
Ira Globus-Harris
UPenn → Cornell

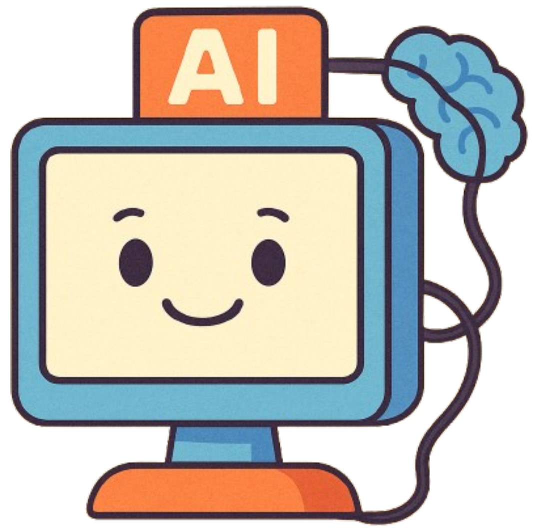
Varun Gupta
UPenn → Vector Institute

Aaron Roth
UPenn

Mirah Shi
UPenn

COLLABORATIVE PREDICTION: SETUP





$$\mathscr{F}_J \subseteq \{f_J : \mathscr{X}_h \times \mathscr{X}_m \to \mathscr{Y}\}$$

$$\mathscr{F}_h \subseteq \{f_h : \mathscr{X}_h \to \mathscr{Y}\}$$

$$\mathscr{F}_m \subseteq \{f_m : \mathscr{X}_m \to \mathscr{Y}\}$$

Goal: We want the agreed upon predictions to have low regret w.r.t. function class $\mathscr{F}_J$ defined on the joint features $x = (x_h, x_m)$

Not possible always (recall parity), so when is it true?

COLLABORATIVE PREDICTION: DEFINITIONS

Weak-learning (recall boosting):

- For all distributions,

  Bounded linear predictors satisfy this with $w(\gamma) = \Theta(\gamma^2)$

  - If there is some $f_J \in \mathcal{F}_J$ that improves over the constant predictor by γ
  - Then there exists either $f_h \in \mathcal{F}_h$ over the human's features or $f_m \in \mathcal{F}_m$ over the AI's features that also improves over the constant predictor by $w(\gamma)$

[Kong-Schoenebeck'23, Frongillo et al.'23] studied assumptions that do guarantee agreement implies information aggregation for Bayesians. Ours are strictly weaker!

COLLABORATIVE PREDICTION: MAIN RESULT

Conversation Multi-Calibration

Multi-calibration $\Longleftrightarrow$ *no-swap regret*

For AI, for all even rounds k, values $p, p' \in [0,1]$, and $f_m \in \mathcal{F}_m$

$$\sum_{t=1}^{T} \mathbb{1}[\hat{y}_m^{(t)}(k) = p]\mathbb{1}[\hat{y}_h^{(t)}(k-1) = p']f_m(x_m)(p - y^{(t)}) = 0$$

Over many days

when AI predicts p

after human predicts p'

AI's predictions are correct on average even when checked against a different rule f_m

Theorem [Colina-GlobusHarris-G-Gupta-Roth-Shi'25]

If both the human and AI are (approximately) conversation multi-calibrated with respect to $\mathcal{F}_h$ and $\mathcal{F}_m$ respectively and $(\mathcal{F}_h, \mathcal{F}_m, \mathcal{F}_J)$ satisfy weak-learnability then agreement* implies low regret with respect to $\mathcal{F}_J$.

**A few caveats*

COLLABORATIVE PREDICTION: HIGH-LEVEL PROOF

Proof sketch:

- We run the protocol till the end of K rounds
- We show that there is a round k where the fraction of disagreements are small
- At this round across days, the predictions have low-swap regret with $\mathcal{F}_h \cup \mathcal{F}_m$
- Using weak-learning guarantee on the level-sets, we get that this should imply low external regret to $\mathcal{F}_J$
- Running for more rounds breaks the low-swap regret condition, but regret cannot increase by much

Take-aways:

- **Simple interaction works:** Exchanging only predictions or actions (no raw features!) can drive effective collaboration
- **Tractable conditions suffice:** We don't need unrealistic Bayesian assumptions; efficiently checkable conditions like *conversation calibration/swap regret* suffice
- **Agreement ⇒ Aggregation:** Under a natural "weak learning" condition, protocols guaranteeing fast agreement *also* achieve information aggregation
- **Provides a practical path:** Offers efficient algorithms to build systems where humans and AI provably make better decisions together

Thank you for listening!