Task Specific Uncertainty Quantification

(A case study in human/AI collaboration)

Aaron Roth

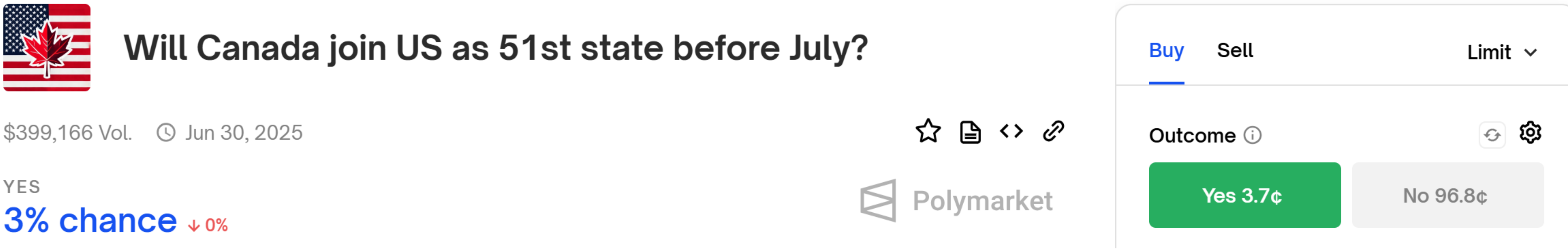
Penn

How can we make sense of probabilities?

- “What is the probability that if I flip a fair coin 16 times I get exactly 9 heads?”
  - We have a mathematical model that maps well onto reality; we can compute this in closed form.
  - We can also conduct the experiment repeatedly and empirically estimate.

How can we make sense of probabilities?

- “What is the probability that Canada will become the 51st state before July?”
  - If we posit a probabilistic model of the universe, this is perhaps philosophically coherent, but it is not a repeatable event; we can’t get empirical estimates.



Will Canada join US as 51st state before July?

$399,166 Vol. Jun 30, 2025

YES

3% chance ↓ 0%

What we want (but can’t get)

- A common scenario:
  - There is (we pretend) a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ (features x labels)
  - We see $x \in \mathcal{X}$, but then need to act before we see $y \in \mathcal{Y}$
    - e.g. we might want to predict (something about) y, or act in a scenario in which y is payoff relevant.
  - If, given x, we knew $\Pr[y|x]$, this would be a sufficient statistic for many downstream tasks
    - e.g. we could find $\arg\max_{a \in \mathcal{A}} \mathbb{E}[u(a, y)|x]$ for any utility function u.
- But real probabilities are generally inaccessible.

We can condition on other events as well.

Given a collection $\mathcal{E}$ of events $E(x, v)$, $f: X \rightarrow [0,1]^d$ has $\mathcal{E}$-bias bounded by α if:

$$||\mathbb{E}[f(x) - y|E(x, f(x))]|| \leq \alpha$$

- “Real probabilities” are unbiased subject to *every* possible conditioning event.
- But maybe for specific tasks, we only needed *some* of the conditional properties of real probabilities in order to use them as probabilities.

The Sequential Prediction Setting

- A context space X
    - Features relevant to the prediction task
- A convex prediction/outcome space $S \subset \mathbb{R}^d$
    - E.g. the probability simplex over outcomes.
- In rounds $t = 1, \ldots, T$:
    - The learner observes some context $x_t \in X$.
    - The learner produces a prediction $\hat{s}_t \in S$
    - The learner observes outcome $s_t \in S$

Efficiently Making $\mathcal{E}$-Unbiased Predictions

Theorem: For any set of events $\mathcal{E}$ and any $\alpha > 0$, there is an online prediction algorithm that can make d-dimensional adversarial predictions over T rounds such that their worst-case $\mathcal{E}$-bias is at most α for:

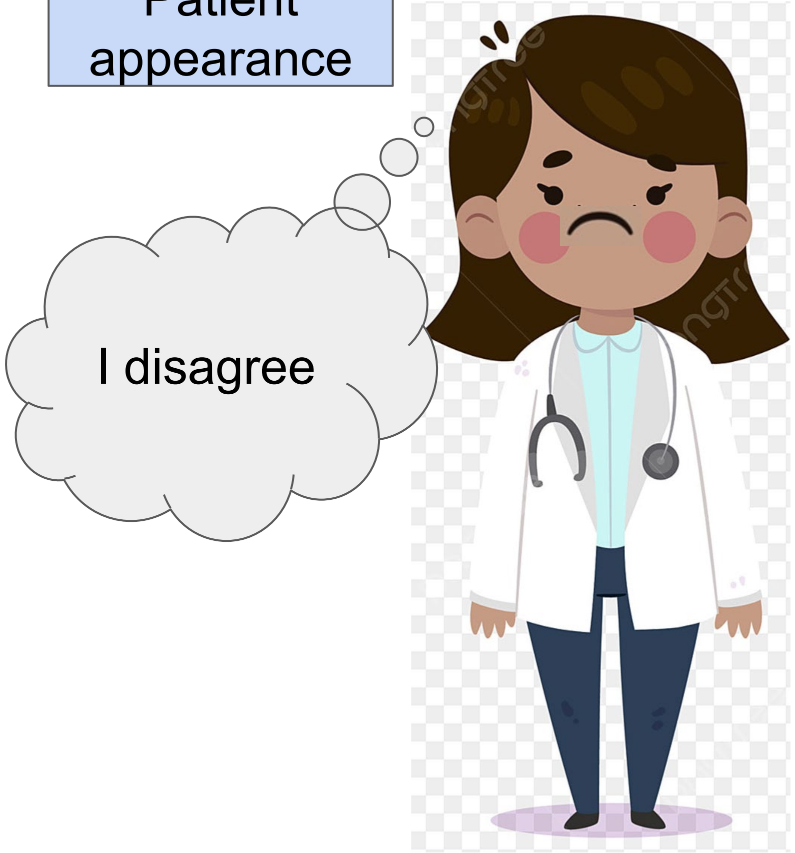
$$\alpha \leq \sqrt{\log(d|\mathcal{E}|T) + T}$$

The per-round running time is polynomial in d, $|\mathcal{E}|$, and T.

Motivation: Humans use AI to help make decisions

Patient disposition

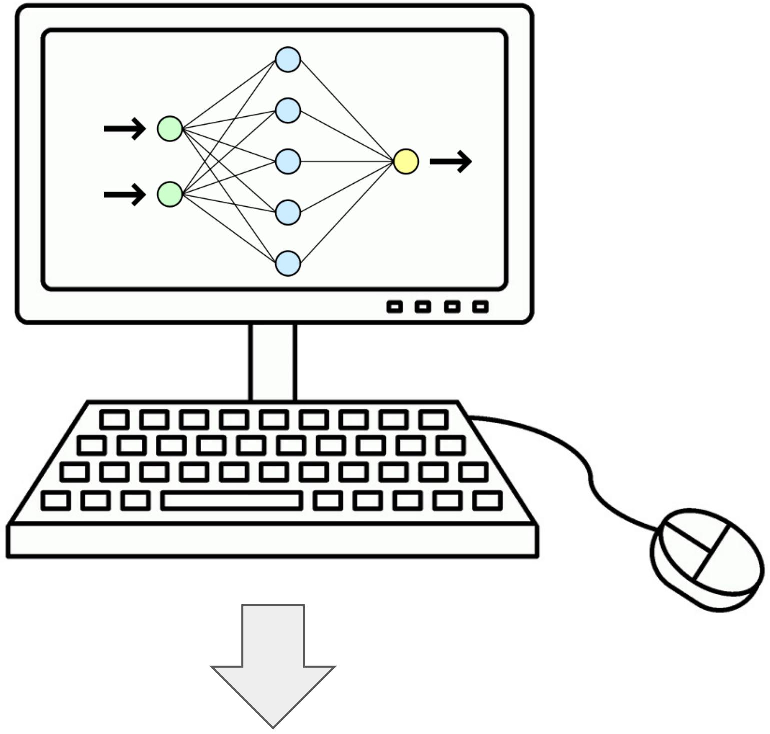
Patient appearance



Past diagnoses

Blood type

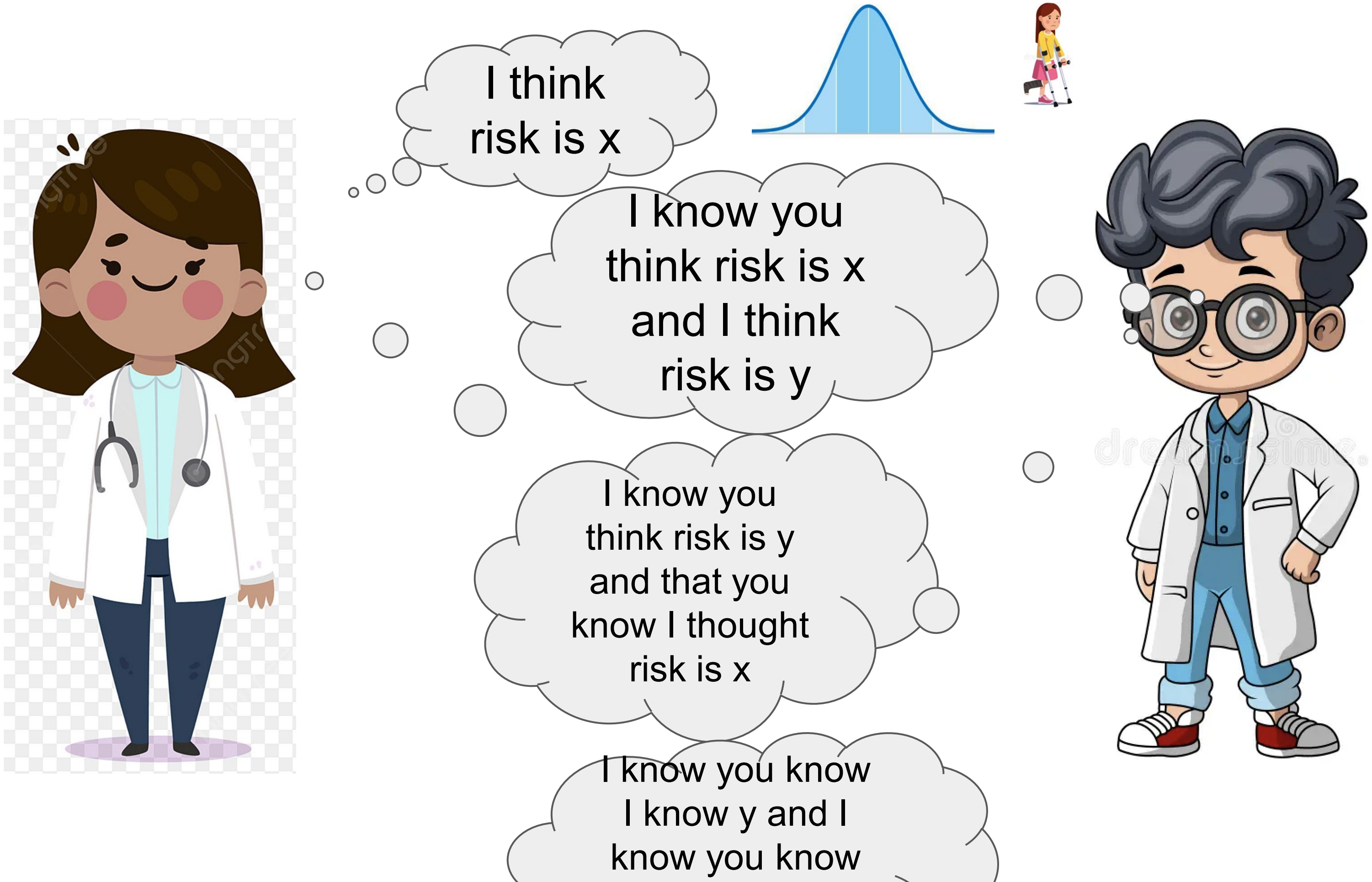
Training data



Risk of operation: 20%

Aumann’s Agreement Theorem

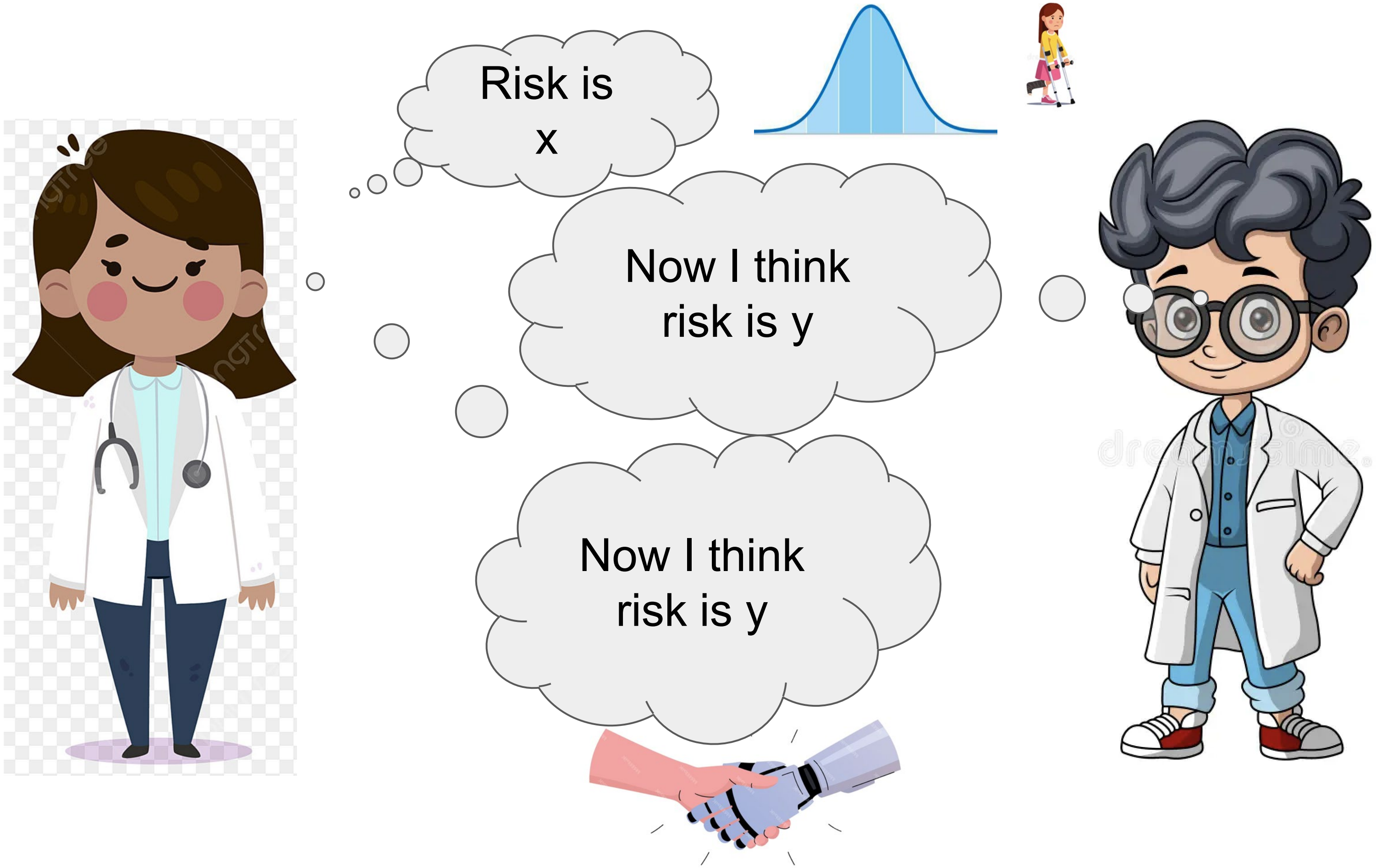
[Aum76] If two bayesian agents have a correct prior and have *common knowledge* of each other’s posterior expectation, they have the same posterior expectation


I think risk is x
I know you think risk is x and I think risk is y
I know you think risk is y and that you know I thought risk is x
I know you know I know y and I know you know

Upshot:
bayesian agents with common knowledge can’t “agree to disagree”

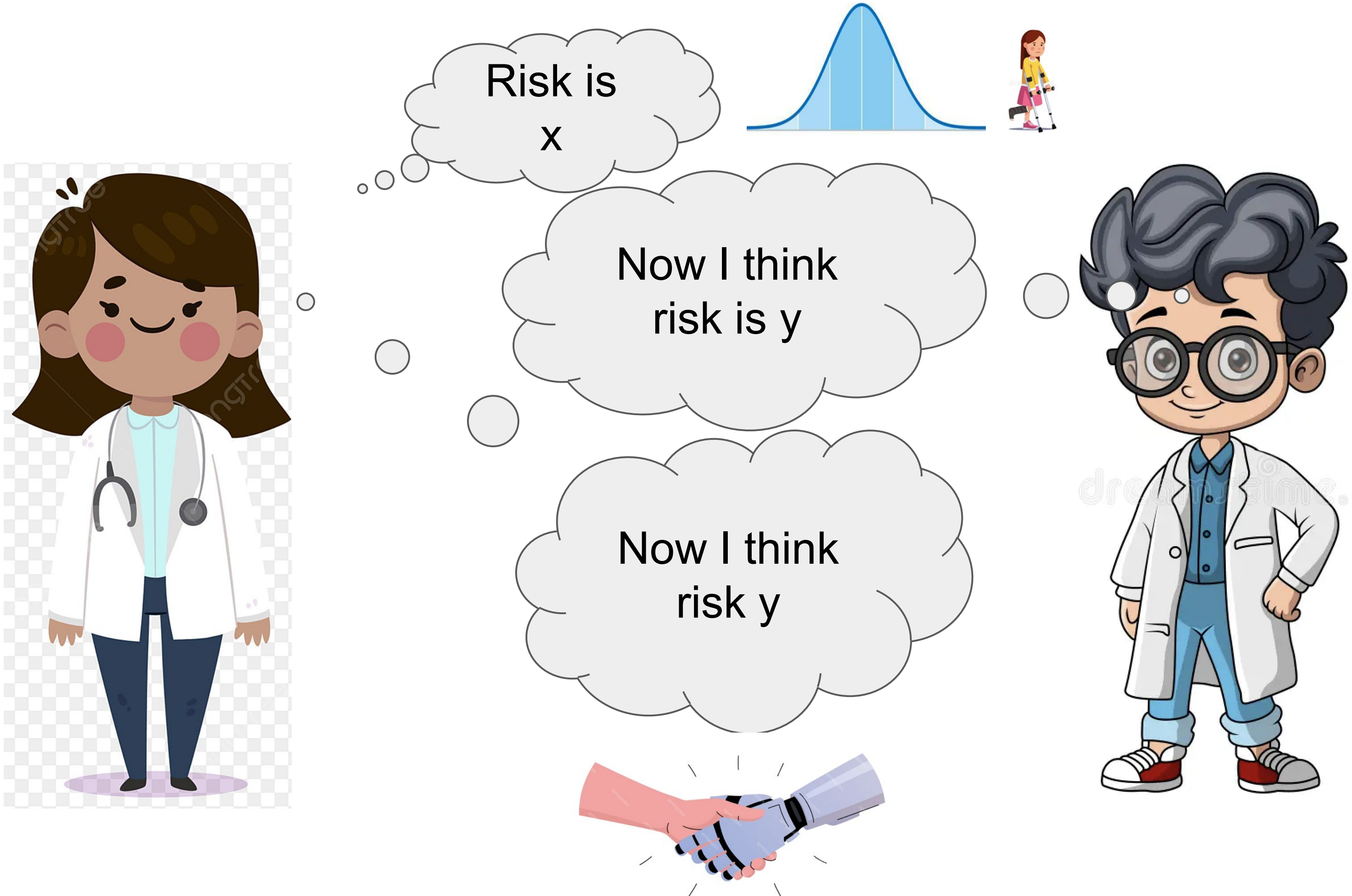
Agreement Dynamics

[GP82] If the underlying state space is finite, agreement happens in a finite number of rounds, if expectation is exchanged in each round



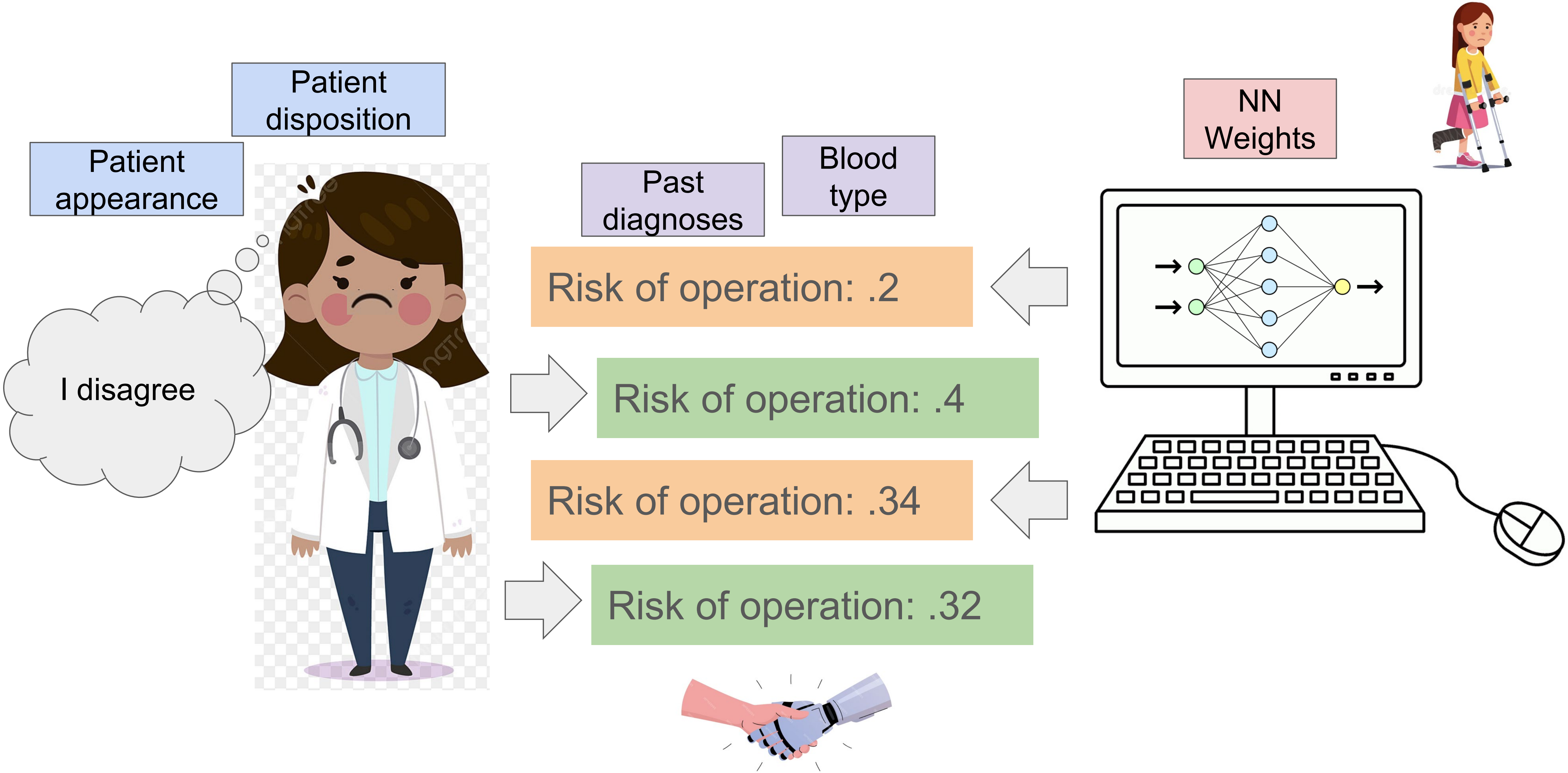
Agreement Protocols

[Aar05] If the underlying state space is finite and the predictions are 1-dimensional, then two **bayesian** agents can reach (eps, delta) agreement in $\frac{1}{\epsilon^2\delta}$ rounds



- With probability 1-delta over the draw of the true state from the prior:
- The agents will have expectations that differ by at most epsilon

Agreement with Full Bayesian Rationality with $\epsilon = 0.02$



Patient disposition

Patient appearance

I disagree

Past diagnoses

Blood type

NN Weights

Risk of operation: .2

Risk of operation: .4

Risk of operation: .34

Risk of operation: .32

Agreement via Perfect Bayesian Rationality: Drawbacks

- Why do we have a common prior?
- Intractable: feature space is huge and arbitrarily correlated with label
  - Unrealistic to assume of a human
  - Intractable to implement for a model
- Unclear how it generalizes beyond 1 dimensional expectations.

Q: Can we relax the behavioral assumption while still maintaining strong agreement convergence guarantees?

A: Yes! With conditional calibration!

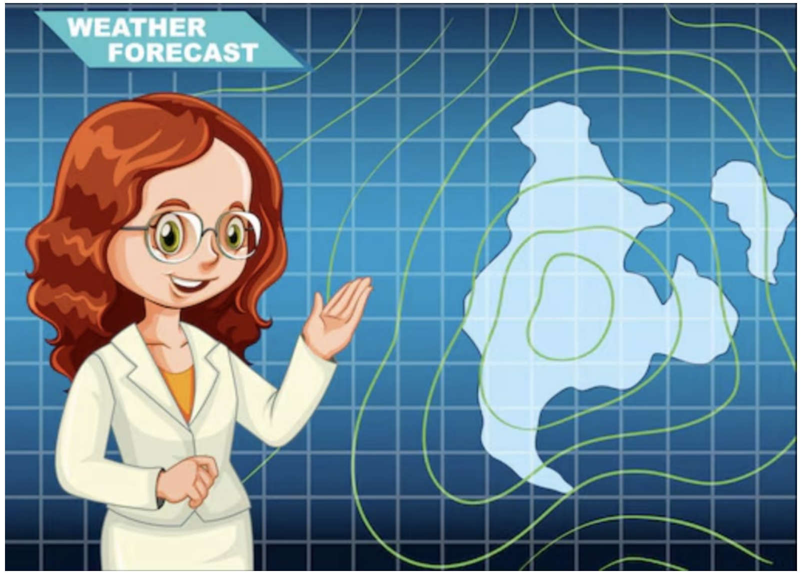
Overview of Results:

- **The right kind of conditional calibration: Conversation calibration**
  - Implied by correctly specified Bayesian rationality
  - But efficiently enforceable in adversarial settings (only small number of conditioning events) and only accuracy improving.
- **Conversation calibration -> fast agreement convergence** in a repeated setting
  - And the longer that conversations go, the *more accurate* the final predictions are
- If the two agents both really *are* Bayesian, we **recover** Aaronson ‘04 results in the one-shot setting, and **generalize** them to multiple dimensions and action feedback

What is Calibration?

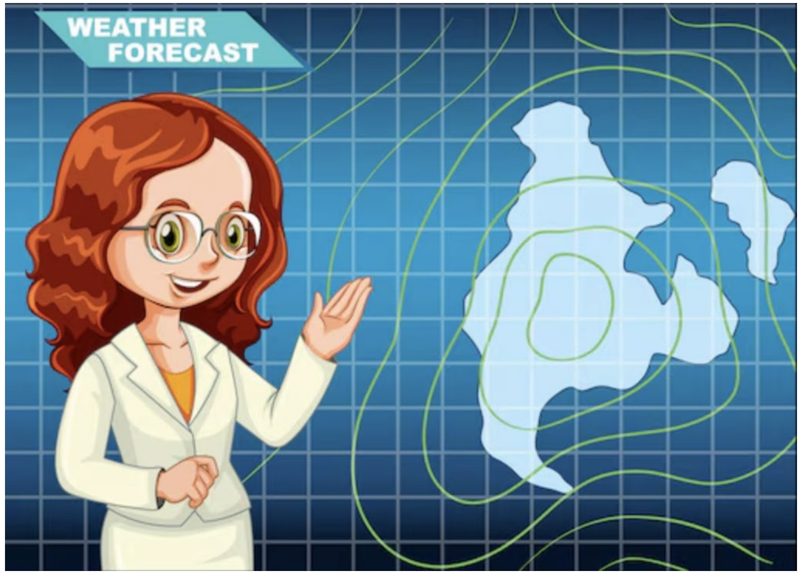
70%
rain

🌧



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?

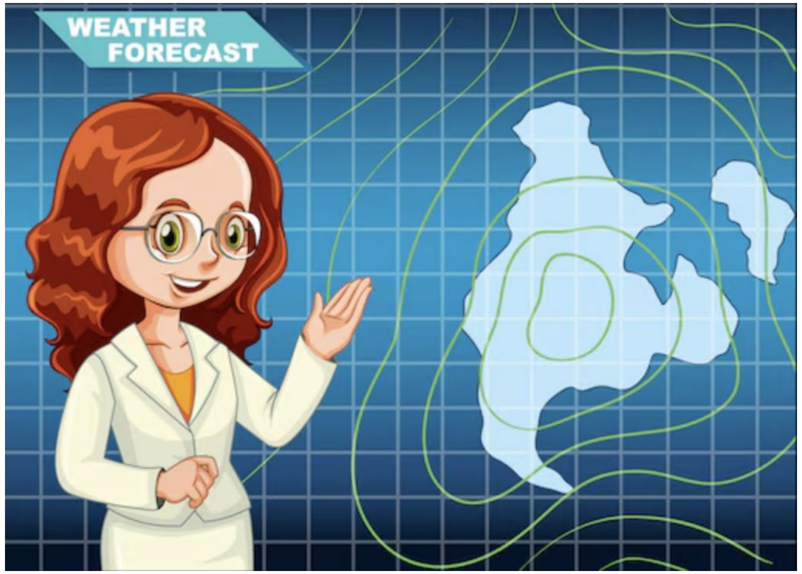
70% rain
🌧

70% rain
☀

70% rain
🌧

70% rain
☀

70% rain
🌧



70% rain
🌧

70% rain
🌧

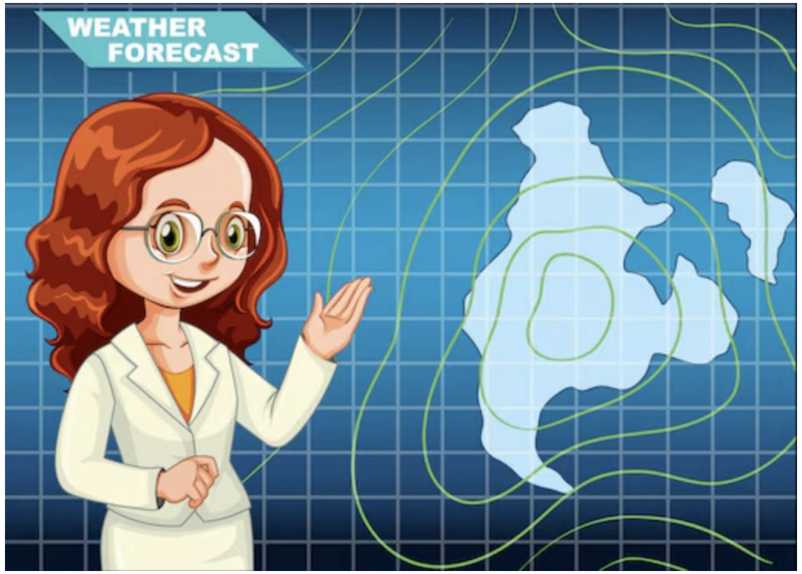
70% rain
🌧

70% rain
☀

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
🌧	🌧	☀	☀	☀	🌧	☀	🌧	☀	🌧



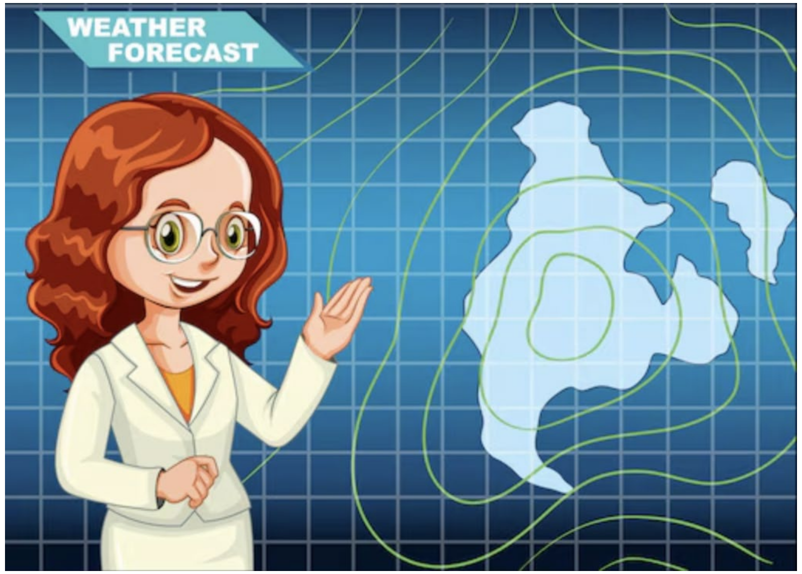
50% rain	70% rain	20% rain	70% rain	20% rain	50% rain
☀	🌧	☀	🌧	🌧	🌧

50% rain	70% rain	10% rain	20% rain	70% rain	70% rain
☀	🌧	☀	🌧	☀	🌧

What is Calibration?

20% rain

🌧



20% rain

☀

20% rain

🌧

20% rain

☀

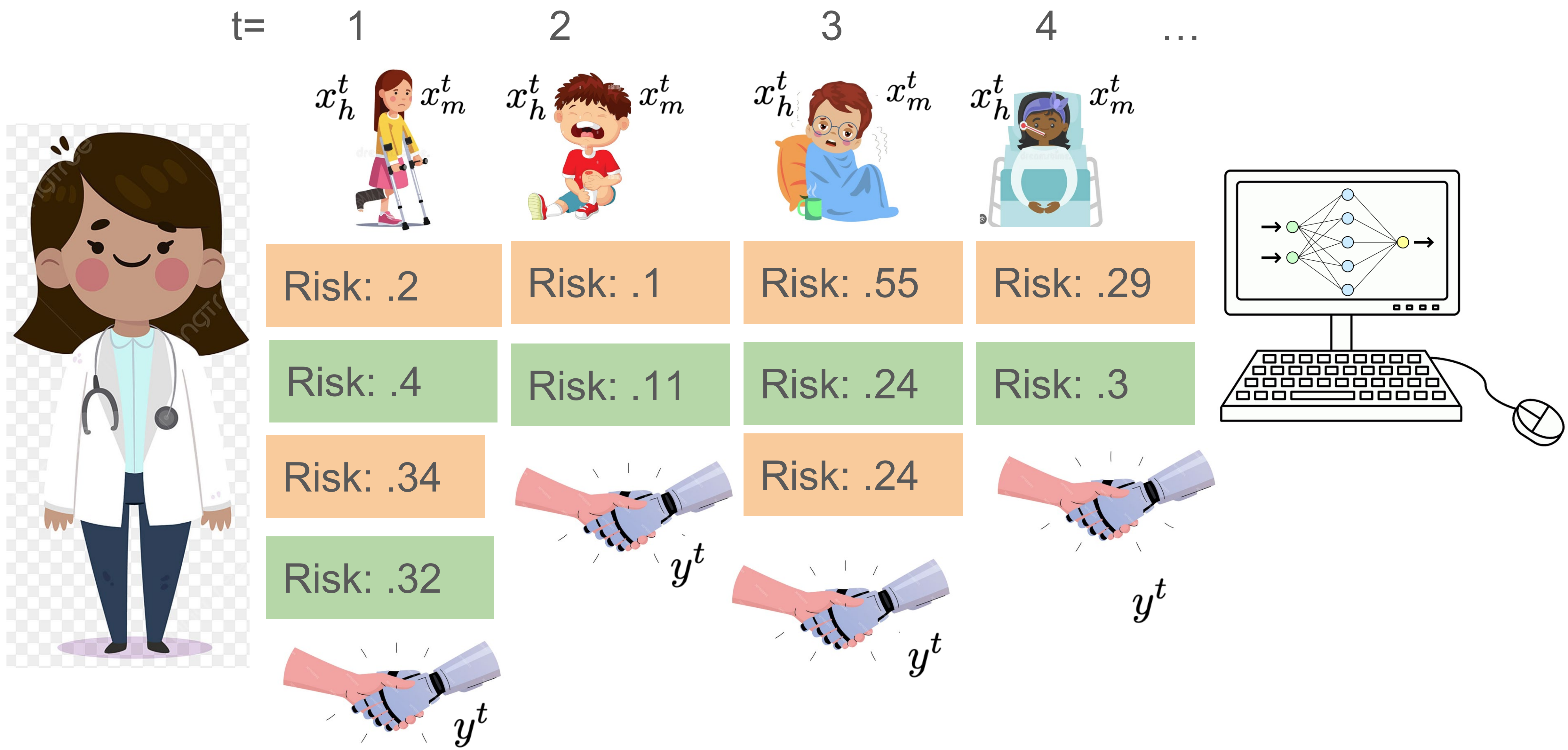
20% rain

🌧

20% rain

🌧

Sequential Agreement Protocol



t= 1

x_h^t x_m^t

Risk: .2

Risk: .4

Risk: .34

Risk: .32

y^t

2

x_h^t x_m^t

Risk: .1

Risk: .11

y^t

3

x_h^t x_m^t

Risk: .55

Risk: .24

Risk: .24

y^t

4

x_h^t x_m^t

Risk: .29

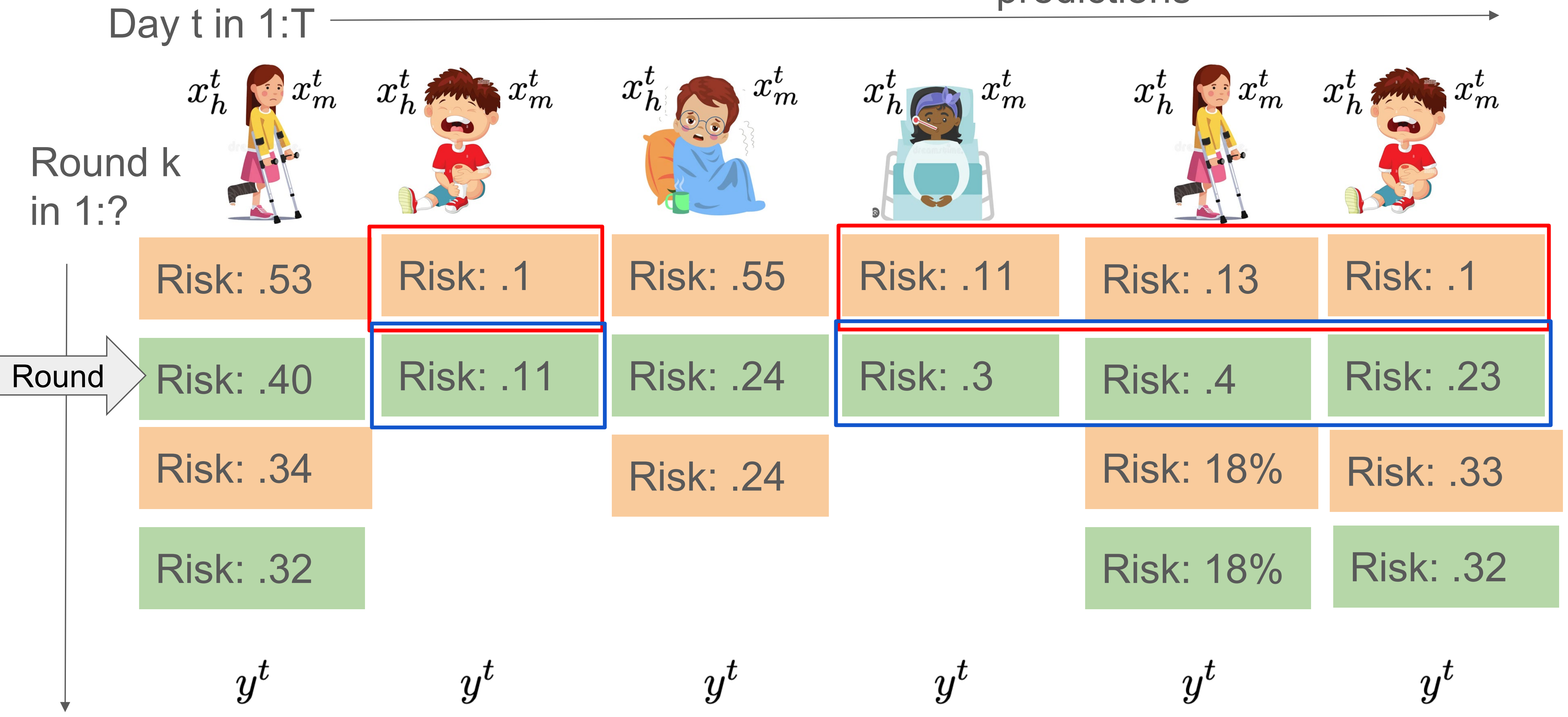
Risk: .3

y^t

…

Perfect Conversation Calibration

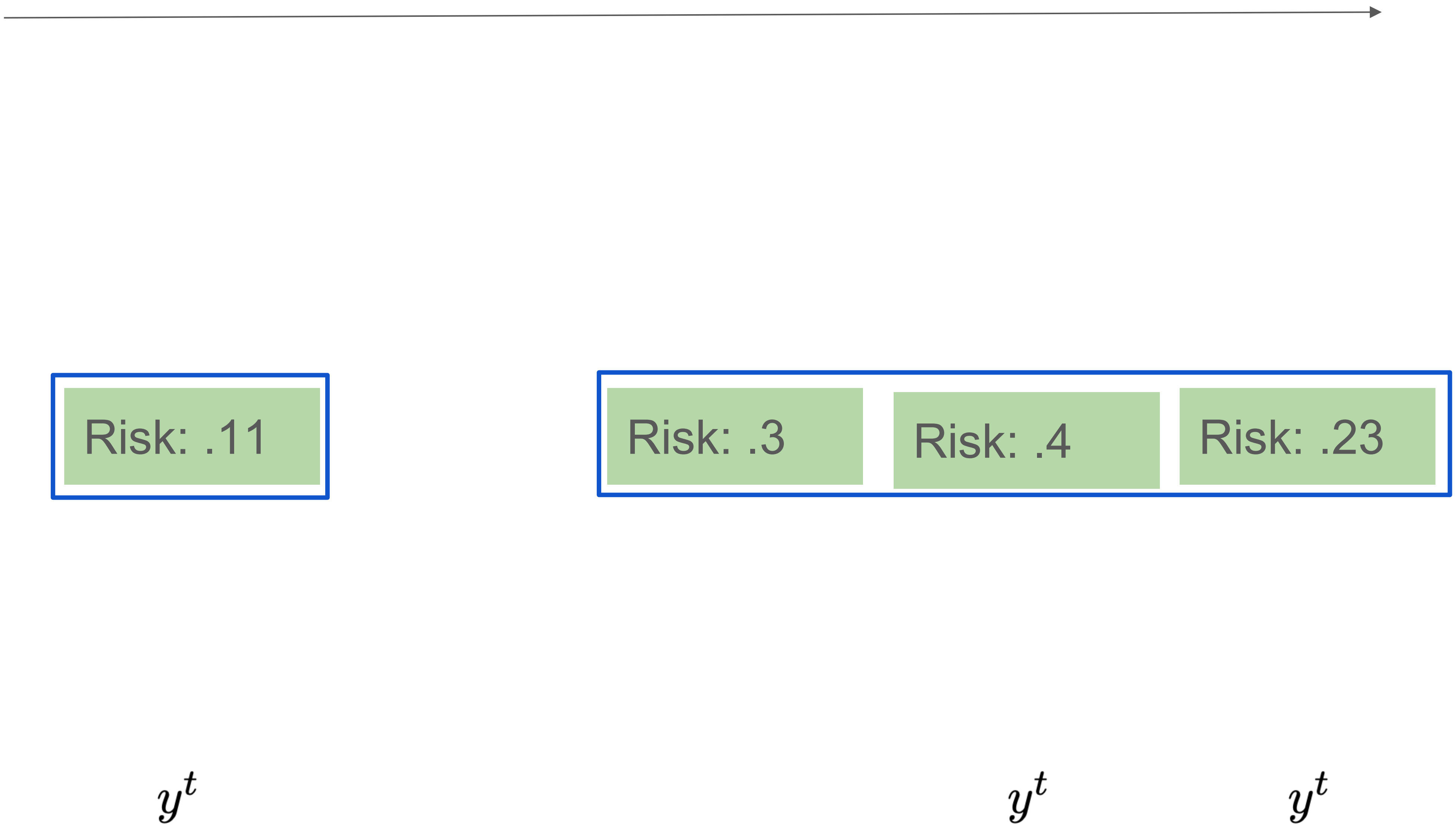
Parameterized by *bucketing* of the other agent's predictions



Risk: .53	Risk: .1	Risk: .55	Risk: .11	Risk: .13	Risk: .1
Risk: .40	Risk: .11	Risk: .24	Risk: .3	Risk: .4	Risk: .23
Risk: .34		Risk: .24		Risk: 18%	Risk: .33
Risk: .32				Risk: 18%	Risk: .32
y^t	y^t	y^t	y^t	y^t	y^t

Perfect Conversation Calibration

Day t in 1:T



Risk: .11

Risk: .3

Risk: .4

Risk: .23

y^t

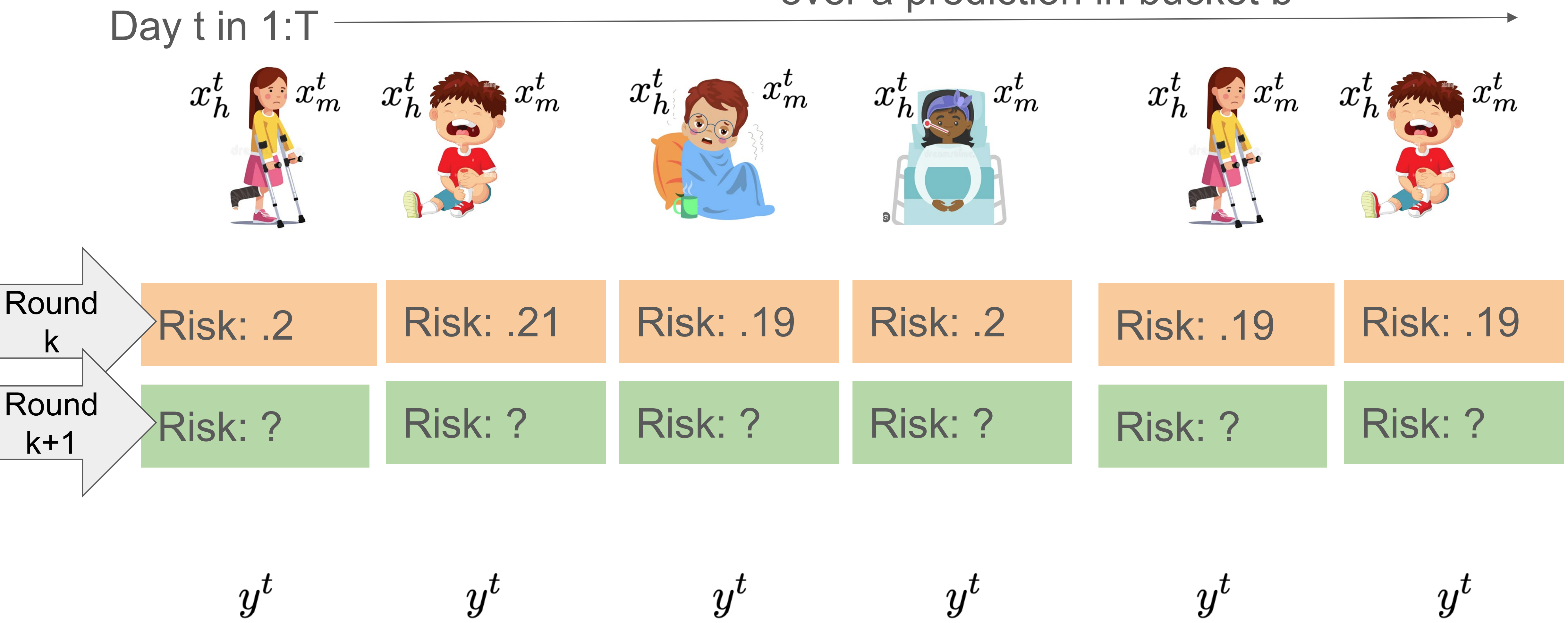
y^t

y^t

In each round, *perfect calibration* on subsequences defined by the other agent's previous prediction (bucketed)

Proof Sketch

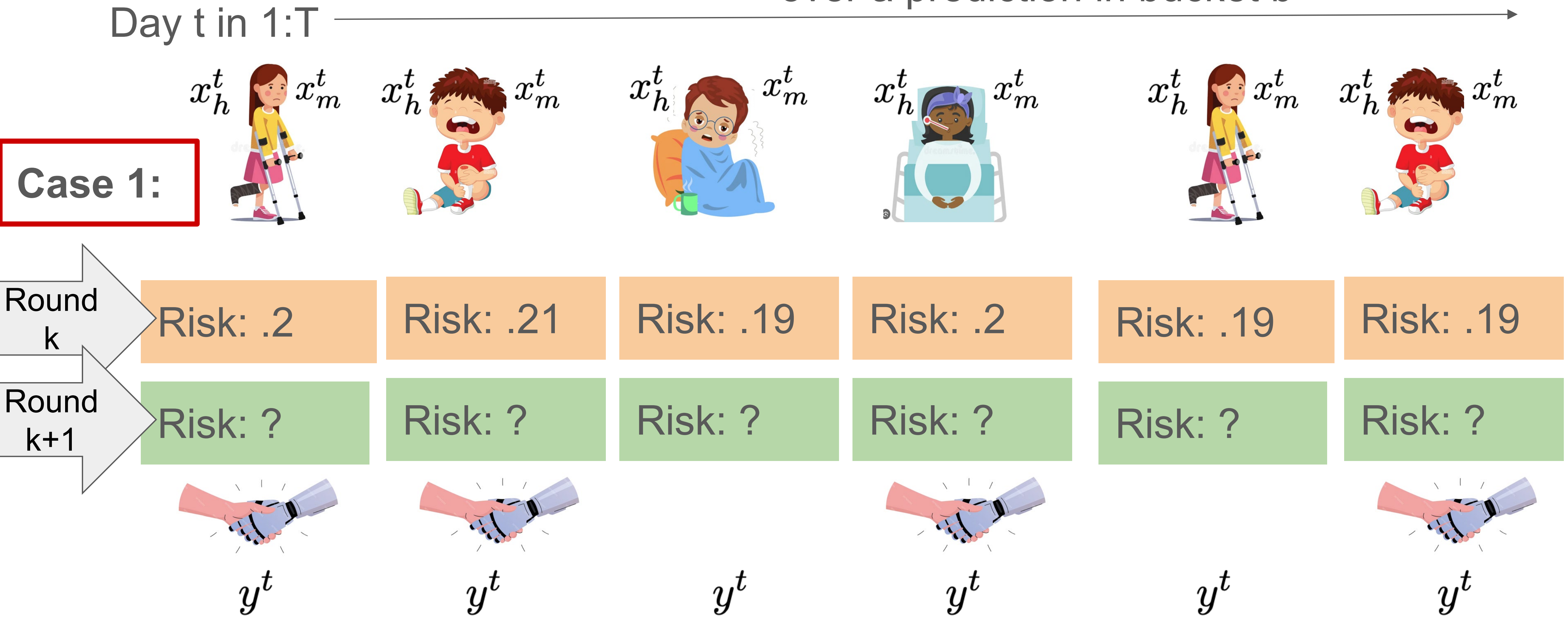
Consider the subsequence of days that make it to round k where the model sends over a prediction in bucket b


Day t in 1:T

Round k	Risk: .2	Risk: .21	Risk: .19	Risk: .2	Risk: .19	Risk: .19
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 in bucket b



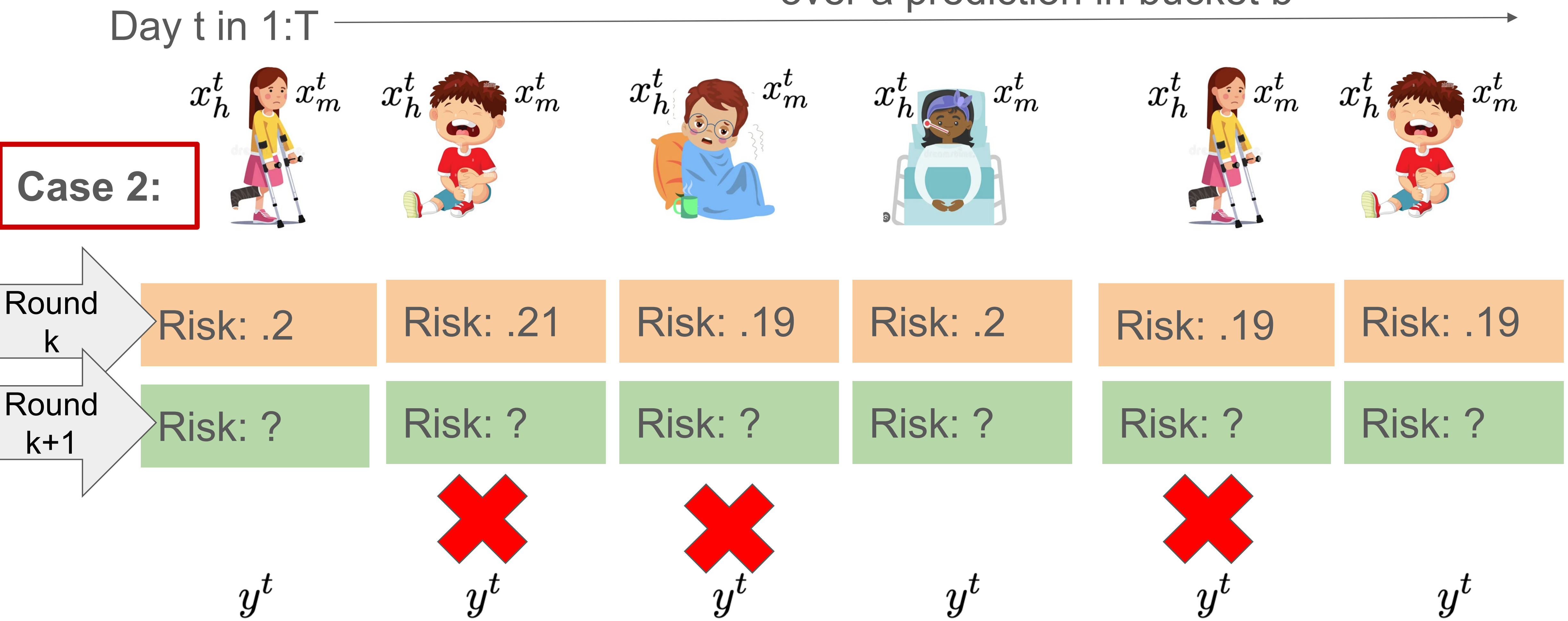
Day t in 1:T

Case 1:

Round k	Risk: .2	Risk: .21	Risk: .19	Risk: .2	Risk: .19	Risk: .19
Round k+1	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?

Proof Sketch

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



Day t in 1:T

Case 2:

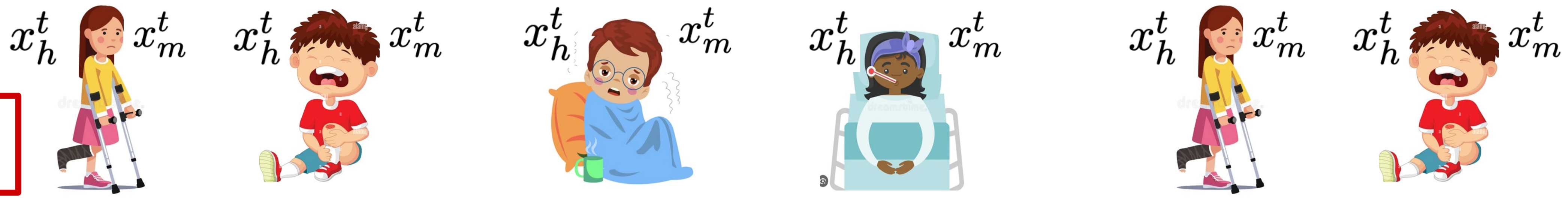
	x_h^t x_m^t	x_h^t x_m^t	x_h^t x_m^t	x_h^t x_m^t	x_h^t x_m^t	x_h^t x_m^t
Round k	Risk: .2	Risk: .21	Risk: .19	Risk: .2	Risk: .19	Risk: .19
Round k+1	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?	Risk: ?
		✗	✗		✗	
	y^t	y^t	y^t	y^t	y^t	y^t

Proof Sketch

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

Day t in 1:T →

Case 2:



Round k	Risk: .2	Risk: .21	Risk: .19	Risk: .2	Risk: .19	Risk: .19
Round k+1	Risk: ?	Risk: 0.3	Risk: 0.1	Risk: ?	Risk: 0.54	Risk: ?
		✗	✗		✗	
	y^t	y^t	y^t	y^t	y^t	y^t

(One) Key Idea:

- If sequence 2 is calibrated conditional on sequence 1, then sequence 2 has at least as low squared error
- Furthermore, if sequence 2 is substantially different than sequence 1, it has *substantially lower* squared error

Round k	Risk: .2	Risk: .21	Risk: .19	Risk: .2	Risk: .19	Risk: .19
Round k+1	Risk: ?	Risk: 0.3	Risk: 0.1	Risk: ?	Risk: 0.54	Risk: ?

Approximate Conversation Calibration leads to fast agreement: Proof Sketch

- **Perfect Conversation Calibration of both agents -> from round i to i+1 either at least half the rounds terminate (agree), or squared error goes down by** $\epsilon^2\delta$

- Approximate Conversation Calibration of both agents does not change this by much

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:

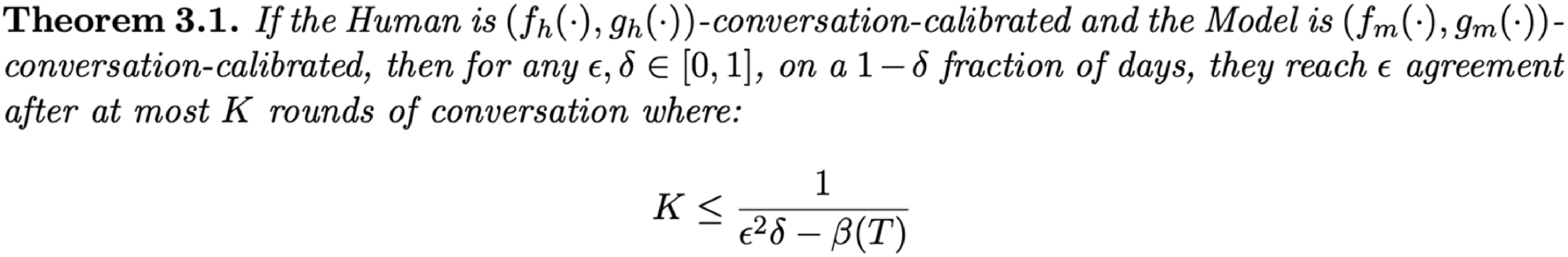
$$\text{Total \# of rounds we can disagree} = \text{(roughly)} \frac{1}{\begin{array}{c}\text{decrease in squared error whenever}\\ \text{less than half of all days terminate}\end{array}}$$

Pulling it all together, informally:

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

Pulling it all together, informally:

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



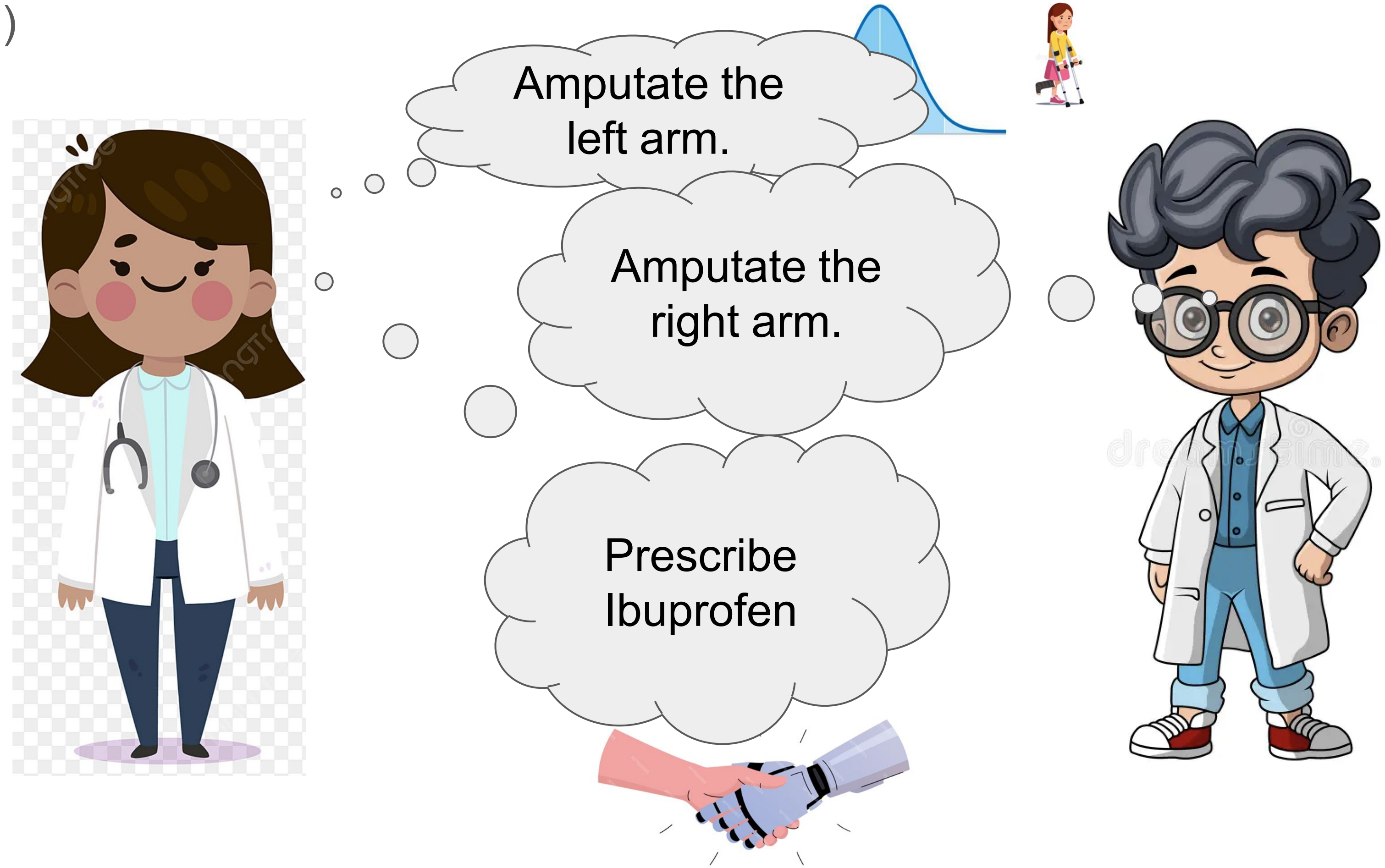
Theorem 3.1. *If the Human is $(f_h(\cdot), g_h(\cdot))$-conversation-calibrated and the Model is $(f_m(\cdot), g_m(\cdot))$-conversation-calibrated, then for any $\epsilon, \delta \in [0,1]$, on a $1-\delta$ fraction of days, they reach ϵ agreement after at most K rounds of conversation where:*

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

Extensions

Can extend to high dimensional spaces where communication/agreement is about the best response action.

Still tractable, agreement to ϵ-approximate best response happens after $\frac{1}{\epsilon\delta}$ rounds. (Independent of ambient dimension)



Amputate the left arm.

Amputate the right arm.

Prescribe Ibuprofen

Thanks!

High-Dimensional Prediction for Sequential Decision Making
Georgy Noarov, Ramya Ramalingam, Aaron Roth, Stephan Xie

Forecasting for Swap Regret for All Downstream Agents
Aaron Roth, Mirah Shi (EC 2024)

Tractable Agreement Protocols
Natalie Collina, Surbhi Goel, Varun Gupta, Aaron Roth (STOC 2025)