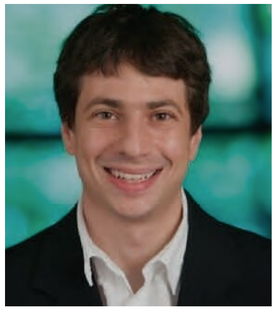
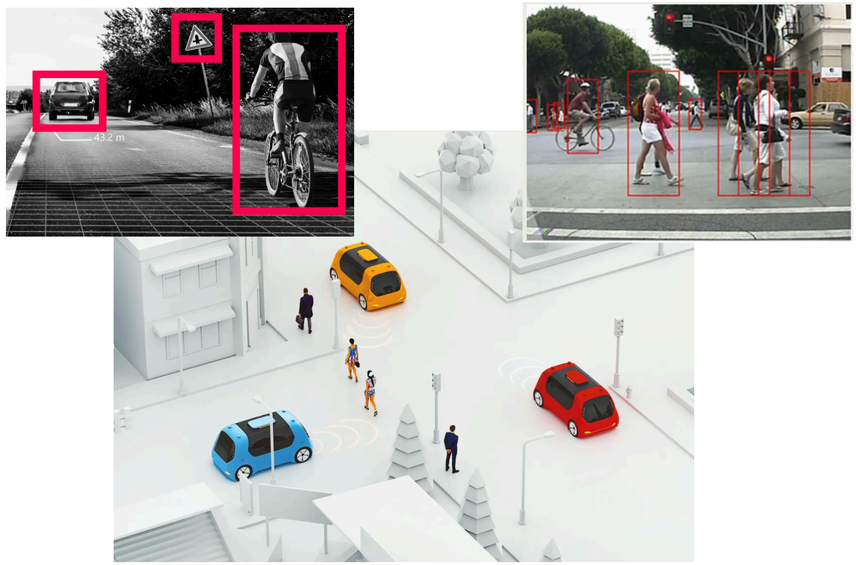
How to Optimally Quantify Uncertainty for Risk-Averse Agents

Shayan Kiyani
University of Pennsylvania

Joint work with: George Pappas, Aaron Roth, Hamed Hassani


Predictions Shape Decisions



Predictions ⟶ Actions

Can we trust the predictions, and hence, the subsequent actions?

The Need for Precise Uncertainty Quantification (UQ)

Tesla's robotaxi push hinges on 'black box' AI gamble

By **Norihiko Shirouzu** and **Chris Kirkham**

October 10, 2024 3:41 PM EDT · Updated 5 months ago

REUTERS

Incorrect AI Advice Influences Diagnostic Decisions

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

RSNA Radiological Society of North America

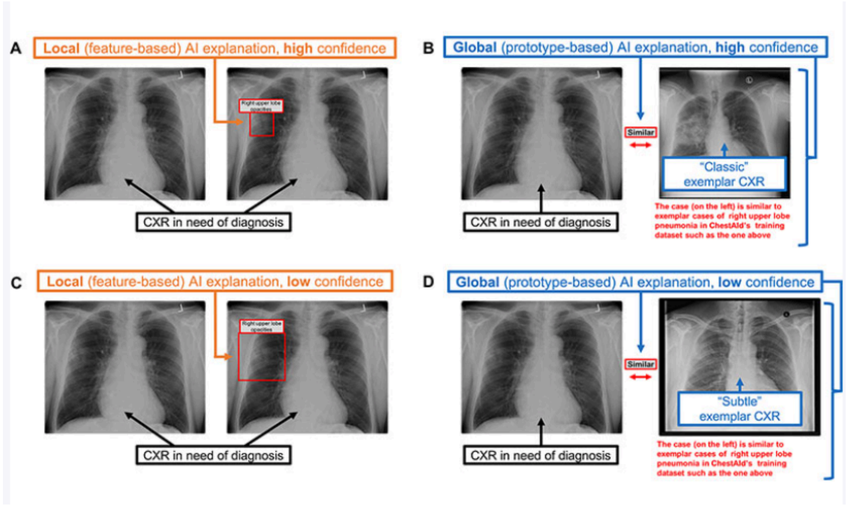
An inside conversation:

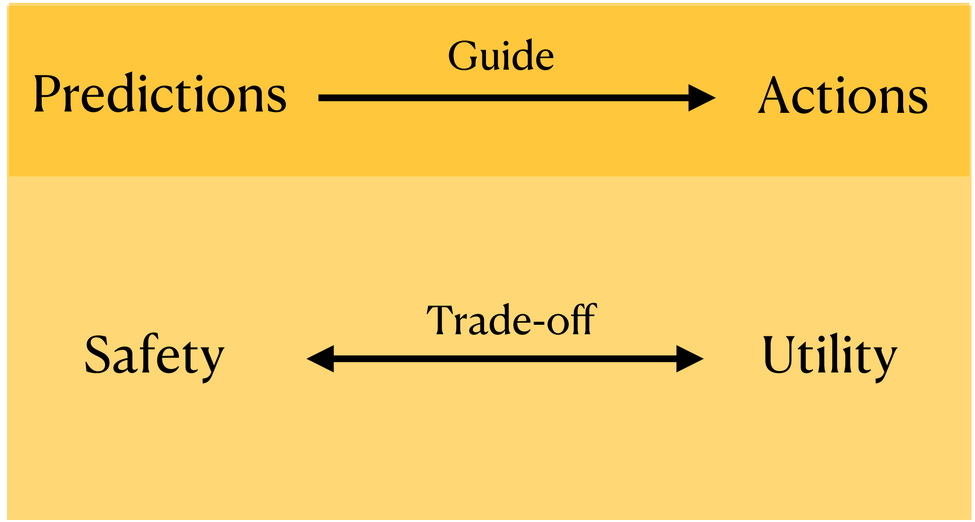
Former Tesla Engineer: "Our end-to-end ML system is a complete black box—we have no insight into how it makes predictions, so if something goes wrong, we can't pinpoint the issue."

AI Researcher: "Right, and while computer vision is generally accurate, studies show it still misses about 3% of objects, which could mean failing to detect a pedestrian."

Former Tesla Engineer: "Exactly. There's always the risk of missing one of the infinite 'edge cases' on the road, making safety hard to guarantee."

Trust in AI is a Double-Edged Sword





On the one hand:

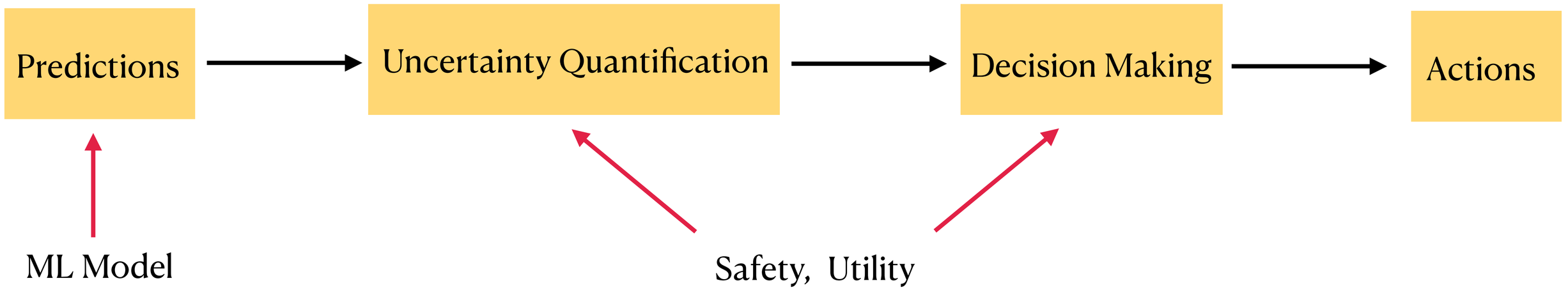
Do nothing! → Ultimate safety

→ Zero utility

On the other hand:

Ignore safety → Maximize utility

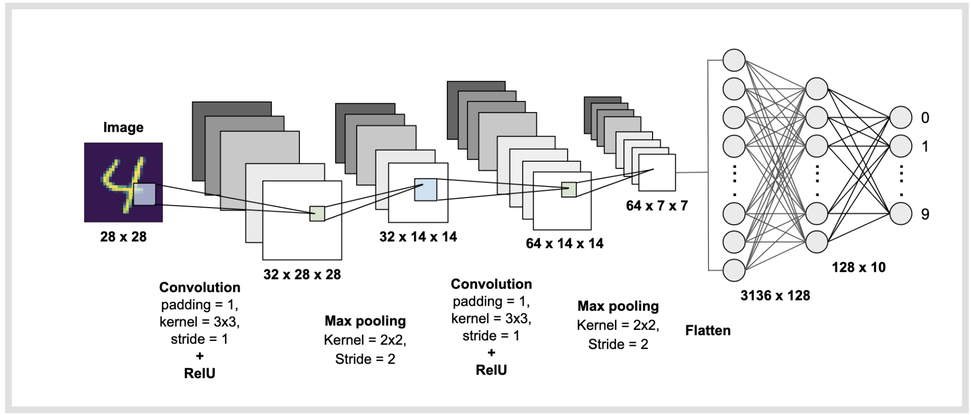
→ Disastrous outcomes



What is the optimal interface between prediction and action that allows for navigating the trade-off between safety and utility in high stakes applications?

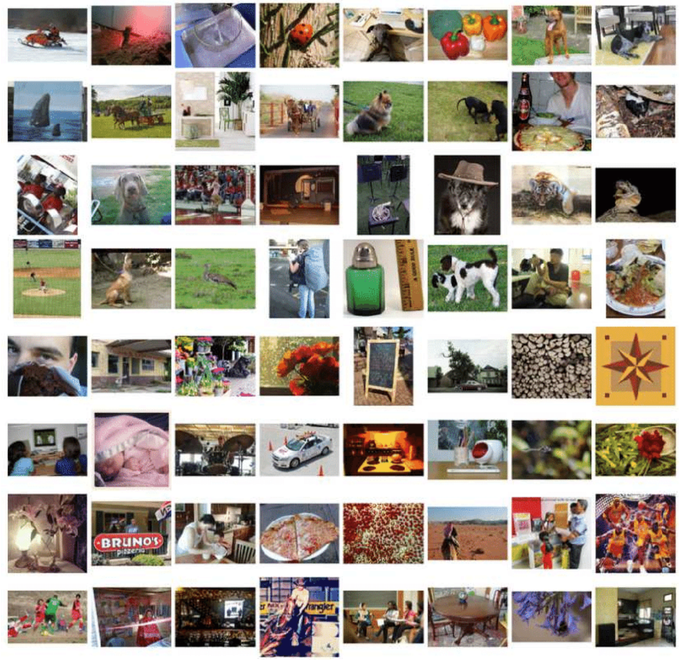
Conformal Prediction – A Promising Framework

Modern ML Paradigm



Model (predictor)

typically of large size

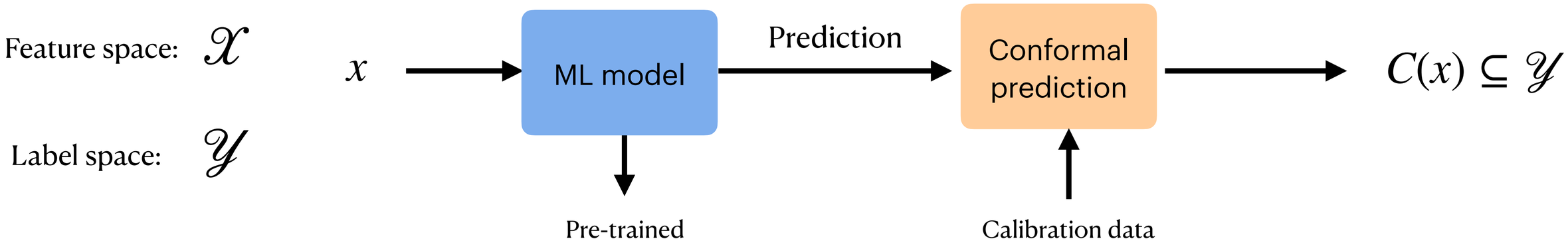


training data (image, text, ..)

Black-box and distribution-free uncertainty quantification

Conformal Prediction (CP)

From Predictions to Prediction Sets



$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

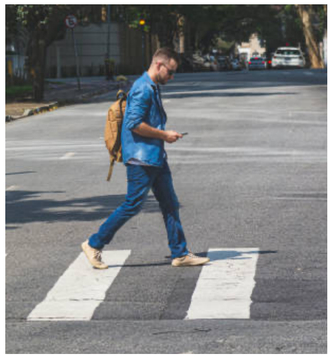
$(X_{\text{test}}, Y_{\text{test}}) \sim \mathcal{D}$

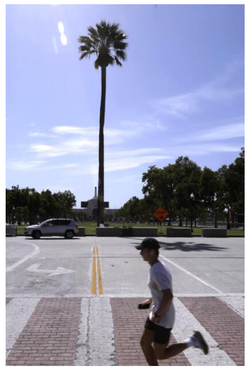
User-specified value; e.g. $1 - \alpha = 0.9$

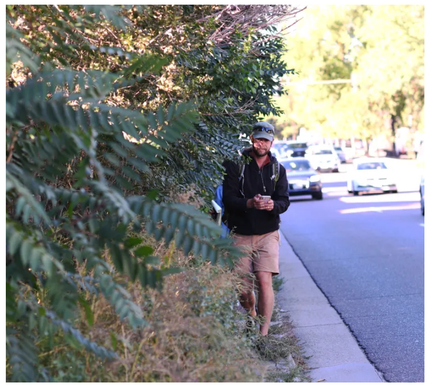
- This is called a marginal guarantee

CP as a framework for UQ

x







{ human 0.99 }

{ human , tree, bin
0.90 0.08 0.02 }

{ Tree, human, trash can
0.70 0.20 0.05 }

$C(x)$ { human } { human } { tree , human }

✅ ✅ ✅

$|C(x)|$ 1 1 2

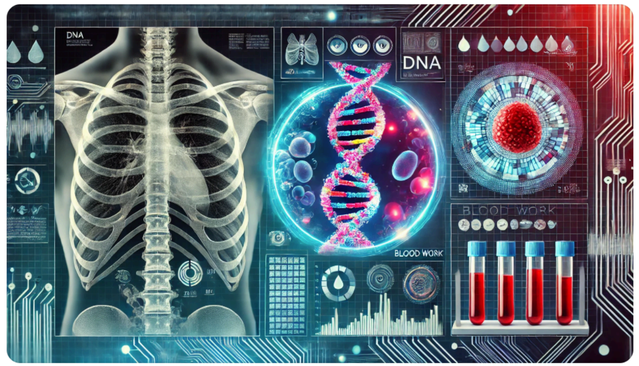
size

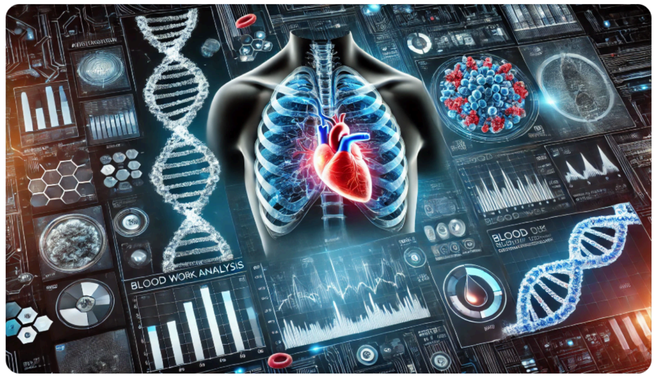
difficulty of the input ↑ models' uncertainty about the label ↑ size of the prediction set ↑

Revisiting CP as a framework for UQ

Clinical medicine

x





$C(x)$ {Cold, Flu, Covid, Allergies} {Cold, Lung cancer}

$|C(x)|$ 4 2

UQ should be decision informed!

Autonomous vehicles

x

$C(x)$ {Tree, Traffic light, Street sign, Utility pole} {Pedesterian, Tree}

$|C(x)|$ 4 2

[Related work: Risk-controlling CP , Angelopoulos et al '22, Lindermann et al '23]

What are prediction sets good for?

1) What kind of downstream decision making process make prediction sets the correct notion of UQ?

"Well designed" prediction sets are a sufficient statistic for risk averse decision makers who wish to optimize their value at risk.

2) What is the optimal policy that a risk averse decision maker should use to map prediction sets to actions?

A simple max min policy is an optimal map from prediction sets to actions.

3) How can we derive prediction sets that are optimal for such decision makers?

We will drive an explicit characterization of these sets over expectation and also provide a finite sample approximation of them.

Decision Making Pipeline

	$\mathcal{X}$	$\mathcal{Y}$	$\mathcal{A}$
	Patients data including X-ray, bloodwork, etc.	Diagnosis	Treatment
	Sensory inputs including LiDAR, vision, etc.	Position of objects, pedestrian, cars, etc, in the next second.	Car inputs
	Historical trends, news, etc.	State of the stock market for tomorrow.	Portfolio design
⋮			

Instead, relying on (uncertain) predictions.

Utility map: $u(a, y) \in \mathbb{R}$ → It captures the preference of decision maker.

The order of things: $X \in \mathcal{X} \to a \in \mathcal{A} \to u(a, Y)$

Risk Averse Decision Making

Action policy

$$a(\cdot) : \mathcal{X} \to \mathcal{A}$$

Utility certificate

$$\nu(\cdot) : \mathcal{X} \to \mathbb{R}$$

Risk Averse Decision Policy Optimization (RA-DPO):

Safety guarantee:

$$\underset{a(\cdot),\, \nu(\cdot)}{\text{maximize}} \quad \mathbb{E}_X\Big[\nu(X)\Big]X)] \geq 1 - \alpha$$

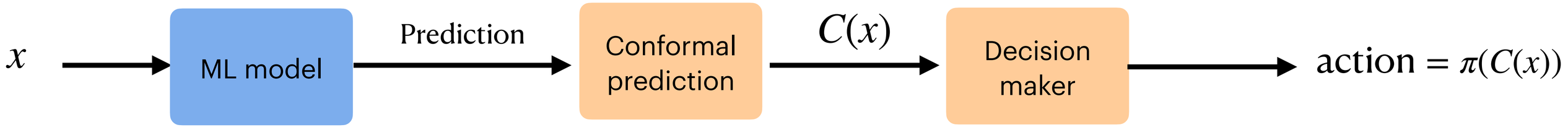
$$\text{subject to}$$

Value at risk!

Risk Averse Decision Making

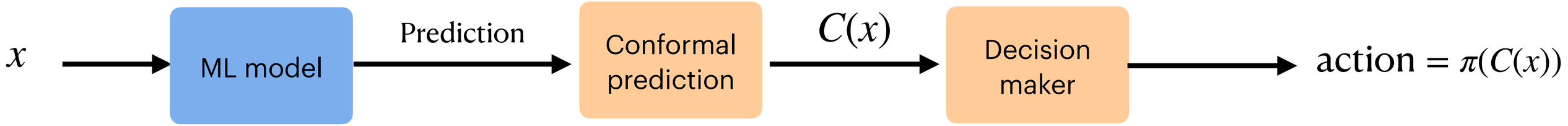
A prediction set perspective

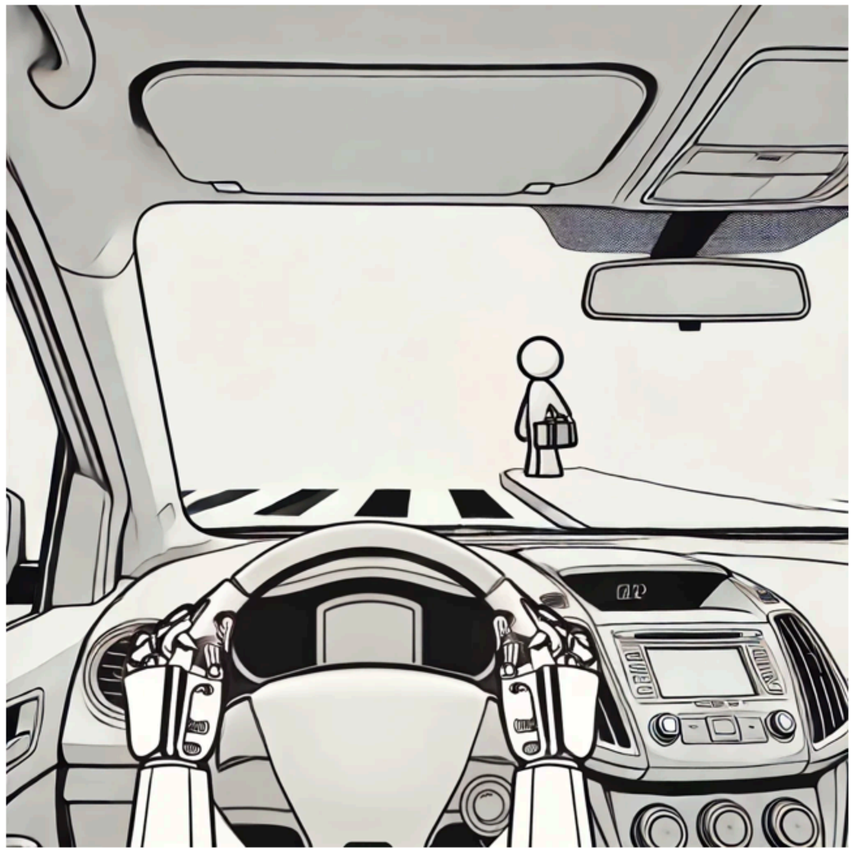
$$\forall x \in \mathcal{X}, \quad C(x) \subseteq \mathcal{Y} \qquad \Pr_{(X,Y)}[Y \in C(X)] \geq 1 - \alpha \qquad \text{Goal: } \pi(.) : 2^{\mathcal{Y}} \to \mathcal{A}$$



What policy π do humans typically choose?

Risk Averse Decision Making





- The pedestrian seems to be standing still

- But there is a small chance that they walk

$$C(x) = \{\text{standing}, \text{walking}\}$$

- To be safe, we act according to the worst-case

$$\pi^*(x) = \arg\max_{a \in \mathcal{A}} \min_{y \in C(x)} u(a, y)$$

We can show that this is provably optimal!

Risk Averse Decision Making

A prediction set perspective

$$\forall x \in \mathcal{X}, \quad C(x) \subseteq \mathcal{Y} \qquad \Pr_{(X,Y)}[Y \in C(X)] \geq 1 - \alpha \qquad \text{Goal: } \pi(.) : 2^{\mathcal{Y}} \to \mathcal{A}$$

Let Ω be the set of all distributions satisfying the above inequality.

RA-DPO:

$$\nu^*(\pi, p) = \underset{\nu(.)}{\text{Maximize}} \quad \mathbb{E}_{X \sim p(x)}\Big[\nu(X)\Big]$$

$$\underset{a(\cdot),\nu(\cdot)}{\text{maximize}} \quad \mathbb{E}_X\big[\nu(X)\big],$$

$$\text{subject to} \quad \Pr\Big[u(a(X), Y)\Big] \qquad \text{subject to} \quad \Pr_{X,Y \sim p(x,y)}[u(\pi(C(X)), Y) \geq \nu(X)] \geq 1 - \alpha$$

$$\underset{\pi}{\text{Maximize}} \ \underset{p \in \Omega}{\text{Minimize}} \quad \nu^*(\pi, p)$$

Proposition: Let $\pi^*(x)$ be the optimal solution to the above objective. Then we have,

$$\pi^*(x) = \arg\max_{a \in \mathcal{A}} \min_{y \in C(x)} u(a, y)$$

Risk Averse Decision Making

A prediction set perspective

Risk averse action policy: $a_{\mathrm{RA}}(C(x)) = \arg\max_{a \in \mathcal{A}} \min_{y \in C(x)} u(a, y)$

Risk averse utility certificate: $\nu_{\mathrm{RA}}(C(x)) = \max_{a \in \mathcal{A}} \min_{y \in C(x)} u(a, y)$

$$\Pr_{(X,Y)}[Y \in C(X)] \geq 1 - \alpha$$

Safety Guarantee: $\Pr_{(X,Y)}[u(a_{\mathrm{RA}}(X), Y) \geq \nu_{\mathrm{RA}}(X)] \geq 1 - \alpha$

RA-DPO:

$$\begin{aligned} \underset{a(\cdot),\, \nu(\cdot)}{\text{maximize}} \quad & \mathbb{E}_X\Big[\nu(X)\Big], \\ \text{subject to} \quad & \Pr\Big[u(a(X), Y) \geq \nu(X)\Big] \geq 1 - \alpha, \end{aligned}$$

Risk Averse Confor

Maximiz $\underset{C(.)}{}$

RA-CPO:

$$\begin{aligned} \underset{C(.)}{\text{Maximize}} \quad & \mathbb{E}_X\left[\max_{a \in \mathcal{A}} \min_{y \in C(X)} u(a, y)\right] \\ \text{subject to} \quad & \Pr[Y \in C(X)] \geq 1 - \alpha \end{aligned}$$

$, y)\Big]$

subject to $\Pr[Y \in C(X)] \geq 1 - \alpha.$

Theorem: RA-DPO and RA-CPO are equivalent.

What are prediction sets good for?

RA-DPO:

$$\underset{a(\cdot),\nu(\cdot)}{\text{maximize}} \quad \mathbb{E}_X\Big[\nu(X)\Big],$$

$$\text{subject to} \quad \Pr\Big[u(a(X), Y) \geq \nu(X)\Big] \geq 1-\alpha,$$

$\equiv$

RA-CPO:

$$\underset{C(.)}{\text{Maximize}} \quad \mathbb{E}_X\left[\max_{a\in\mathcal{A}} \min_{y\in C(X)} u(a,y)\right]$$

$$\text{subject to} \quad \Pr[Y \in C(X)] \geq 1-\alpha$$

1) What kind of downstream decision making process make prediction sets the correct notion of UQ?

✓ "Well designed" prediction sets are a sufficient statistic for risk averse decision makers who wish to optimize their value at risk.

2) What is the optimal policy that a risk averse decision maker should use to map prediction sets to actions?

✓ A simple max min policy is an optimal map from prediction sets to actions.

3) How can we derive prediction sets that are optimal for such decision makers?

We will drive an explicit characterization of these sets over expectation and also provide a finite sample approximation of them.

?

How to Optimize Prediction Sets?

RA-CPO:

$$\underset{C(.)}{\text{Maximize}} \quad \mathbb{E}_X \left[\max_{a \in \mathcal{A}} \min_{y \in C(X)} u(a, y) \right]$$

$$\text{subject to} \quad \Pr[Y \in C(X)] \geq 1 - \alpha$$

No convexity or concavity property!

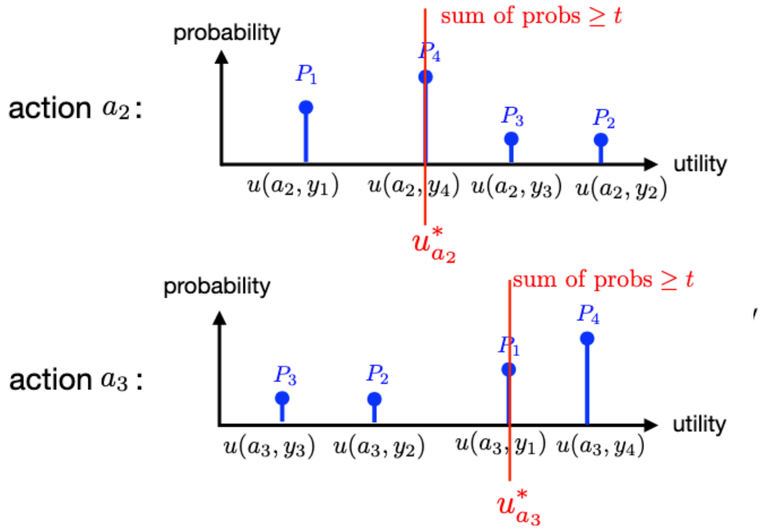
A different parametrization:

Conditional coverage assignment: $t(x) = \Pr\left[Y \in C(X) \,|\, X = x\right]$

Optimal risk averse utility for a fixed coverage assignment:

$$\boldsymbol{\theta}(x, t) = \max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a, Y) \mid X = x\right] = \max\{u^*_{a_1}, u^*_{a_2}, u^*_{a_3}\} = u^*_{a_3}$$

$$\boldsymbol{a}(x, t) = \arg\max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a, Y) \mid X = x\right] = a_3$$



How to Optimize Prediction Sets?

RA-CPO:

$$\underset{C(.)}{\text{Maximize}} \quad \mathbb{E}_X\left[\max_{a\in\mathcal{A}} \min_{y\in C(X)} u(a,y)\right]$$

$$\text{subject to} \quad \Pr[Y \in C(X)] \geq 1-\alpha$$

$$\equiv$$

$$\underset{t:\mathcal{X}\to[0,1]}{\text{maximize}} \quad \mathbb{E}_X\left[\boldsymbol{\theta}(X, t(X))\right]$$

$$\text{subject to:} \quad \mathbb{E}_X\left[t(X)\right] \geq 1-\alpha.$$

Having t^* we have,

$$C^*(x) = \left\{ y \in \mathcal{Y} : \; u(\boldsymbol{a}(x, t^*(x)), y) \geq \boldsymbol{\theta}(x, t^*(x)) \right\}.$$

How to Optimize Prediction Sets?

What the risk averse agent cares about

RA-DPO:

$$\underset{a(\cdot),\nu(\cdot)}{\text{maximize}} \quad \mathbb{E}_X\Big[\nu(X)\Big],$$

$$\text{subject to} \quad \Pr\Big[u(a(X), Y) \geq \nu(X)\Big] \geq 1-\alpha,$$

$\equiv$

An equivalent conformal optimization

RA-CPO:

$$\underset{C(.)}{\text{Maximize}} \quad \mathbb{E}_X\left[\max_{a\in\mathcal{A}} \min_{y\in C(X)} u(a,y)\right]$$

$$\text{subject to} \quad \Pr[Y \in C(X)] \geq 1-\alpha$$

$\equiv$

An equivalent parametrization

$$\underset{t:\mathcal{X}\to[0,1]}{\text{maximize}} \quad \mathbb{E}_X\Big[\boldsymbol{\theta}(X, t(X))\Big]$$

$$\text{subject to:} \quad \mathbb{E}_X\Big[t(X)\Big] \geq 1-\alpha.$$

What is t^*?

No convexity or concavity property!

A one dimensional characterization:

$$\boldsymbol{g}(x,\beta) = \arg\max_{s\in[0,1]} \Big\{\boldsymbol{\theta}(x,s) + \beta s\Big\}$$

Theorem: Let $t^*(x)$ be the optimal solution. Then, there exists a $\beta^* \geq 0$ such that,

$$t^*(x) = \boldsymbol{g}(x,\beta^*)\,.$$

Further, β^* is a solution to the following equation $\mathbb{E}_X[\boldsymbol{g}(X,\beta^*)] = 1-\alpha$

Finite Sample Algorithm

Essential components

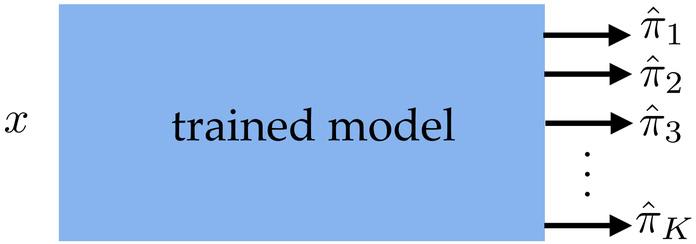
$$\boldsymbol{\theta}(x,t) = \max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a,Y) \mid X = x\right] \qquad \boldsymbol{a}(x,t) = \arg\max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a,Y) \mid X = x\right]$$

$$\boldsymbol{g}(x,\beta) = \arg\max_{s \in [0,1]} \left\{\boldsymbol{\theta}(x,s) + \beta s\right\} \qquad t^*(x) = \boldsymbol{g}(x,\beta^*) \qquad \mathbb{E}_X[\boldsymbol{g}(X,\beta^*)] = 1-\alpha$$

$$C^*(x) = \left\{y \in \mathcal{Y}:\ u(\boldsymbol{a}(x,t^*(x)),y) \geq \boldsymbol{\theta}(x,t^*(x))\right\}$$

The only quantity that we have to approximate: $\text{quantile}_{1-t}\left[u(a,Y) \mid X = x\right]$

Approximate by: $\text{quantile}_{1-t}\left[u(a,Y) \mid Y \sim \hat{\pi}(x)\right]$



$$\boldsymbol{\theta}(x,t) \quad \boldsymbol{g}(x,\beta) \quad \boldsymbol{a}(x,t) \longrightarrow \hat{\boldsymbol{\theta}}(x,t) \quad \hat{\boldsymbol{g}}(x,\beta) \quad \hat{\boldsymbol{a}}(x,t)$$

Finite Sample Algorithm

Essential components

$$\boldsymbol{\theta}(x,t) = \max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a,Y) \mid X = x\right] \qquad \boldsymbol{a}(x,t) = \arg\max_{a \in \mathcal{A}} \text{quantile}_{1-t}\left[u(a,Y) \mid X = x\right]$$

$$\boldsymbol{g}(x,\beta) = \arg\max_{s \in [0,1]} \left\{\boldsymbol{\theta}(x,s) + \beta s\right\} \qquad t^*(x) = \boldsymbol{g}(x,\beta^*) \qquad \mathbb{E}_X[\boldsymbol{g}(X,\beta^*)] = 1-\alpha$$

$$C^*(x) = \left\{y \in \mathcal{Y} : \; u(\boldsymbol{a}(x,t^*(x)),y) \geq \boldsymbol{\theta}(x,t^*(x))\right\}$$

$\boldsymbol{\theta}(x,t)$ $\boldsymbol{g}(x,\beta)$ $\boldsymbol{a}(x,t)$ ⟶ $\hat{\boldsymbol{\theta}}(x,t)$ $\hat{\boldsymbol{g}}(x,\beta)$ $\hat{\boldsymbol{a}}(x,t)$

$$\hat{C}(x;\beta) = \left\{y \in \mathcal{Y} : \; u\left(\hat{\boldsymbol{a}}\left(x, \hat{\boldsymbol{g}}(x,\beta)\right), y\right) \geq \hat{\boldsymbol{\theta}}\left(x, \hat{\boldsymbol{g}}(x,\beta)\right)\right\}$$

We calibrate β to get the correct coverage!

Risk Averse Calibration (RAC)

Algorithm:

1: **Input:** Miscoverage level α, Calibration samples $\{(X_i, Y_i)\}_{i=1}^n$, Test covariate X_{test}.

2: **for each** $y \in \mathcal{Y}$**:**

$$\hat{\beta}_y = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}}\, \beta \quad \text{subject to:} \quad \frac{1}{n+1}\Big\{\sum_{i=1}^{n} [Y_i \in \hat{C}(X_i; \beta)] + \mathbf{1}[y \in \hat{C}(X_{\text{test}}; \beta)]\Big\} \geq 1 - \alpha.$$

3: **Output:**

$$C_{\text{RAC}}(X_{\text{test}}) = \left\{y \in \mathcal{Y} \mid y \in \hat{C}(X_{\text{test}}; \hat{\beta}_y)\right\}.$$

Theorem: Assume that the calibration samples $\{(X_i, Y_i)\}_{i=1}^n$ and $\{(X_i, Y_i)\}_{i=1}^n$ are exchangeable. Then, we have

$$\Pr\left[Y_{\text{test}} \in C_{\text{RAC}}(X_{\text{test}})\right] \geq 1 - \alpha,$$

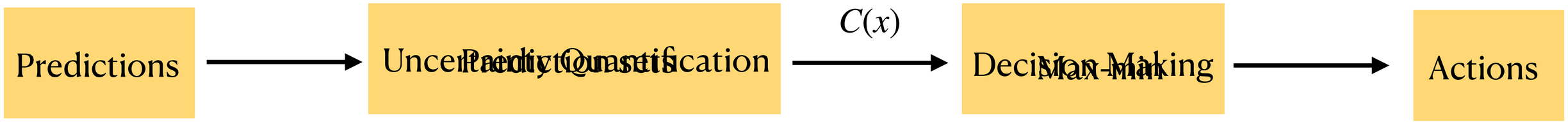
over the randomness of the test and calibration data.

$$\Pr\left[u\big(a_{\text{RA}}\left(C_{\text{RAC}}(X_{\text{test}})\right), Y_{\text{test}}\big) \geq \nu_{\text{RA}}(C_{\text{RAC}}(X_{\text{test}}))\right] \geq 1 - \alpha$$

Importantly,

Predictive model performance ↑ $\quad \text{quantile}_{1-t}\left[u(a, Y) \mid Y \sim f_x\right] \longrightarrow \text{quantile}_{1-t}\left[u(a, Y) \mid X = x\right]$ $\quad$ RAC utility performance ↑

Risk-averse decision making (safety vs utility):



$$t^*(x) = \boldsymbol{g}(x, \beta^*)$$

$$\pi^*(x) = \arg\max_{a \in \mathcal{A}} \min_{y \in C(x)} u(a, y)$$

$$C^*(x) = \left\{ y \in \mathcal{Y} : \; u(\boldsymbol{a}(x, t^*(x)), y) \geq \boldsymbol{\theta}(x, t^*(x)) \right\}$$

Thank You!

[Related work: Foster & Vohra ‘98 , Zhao et al ’21, Noarov et al ’23]

Clinical Medicine

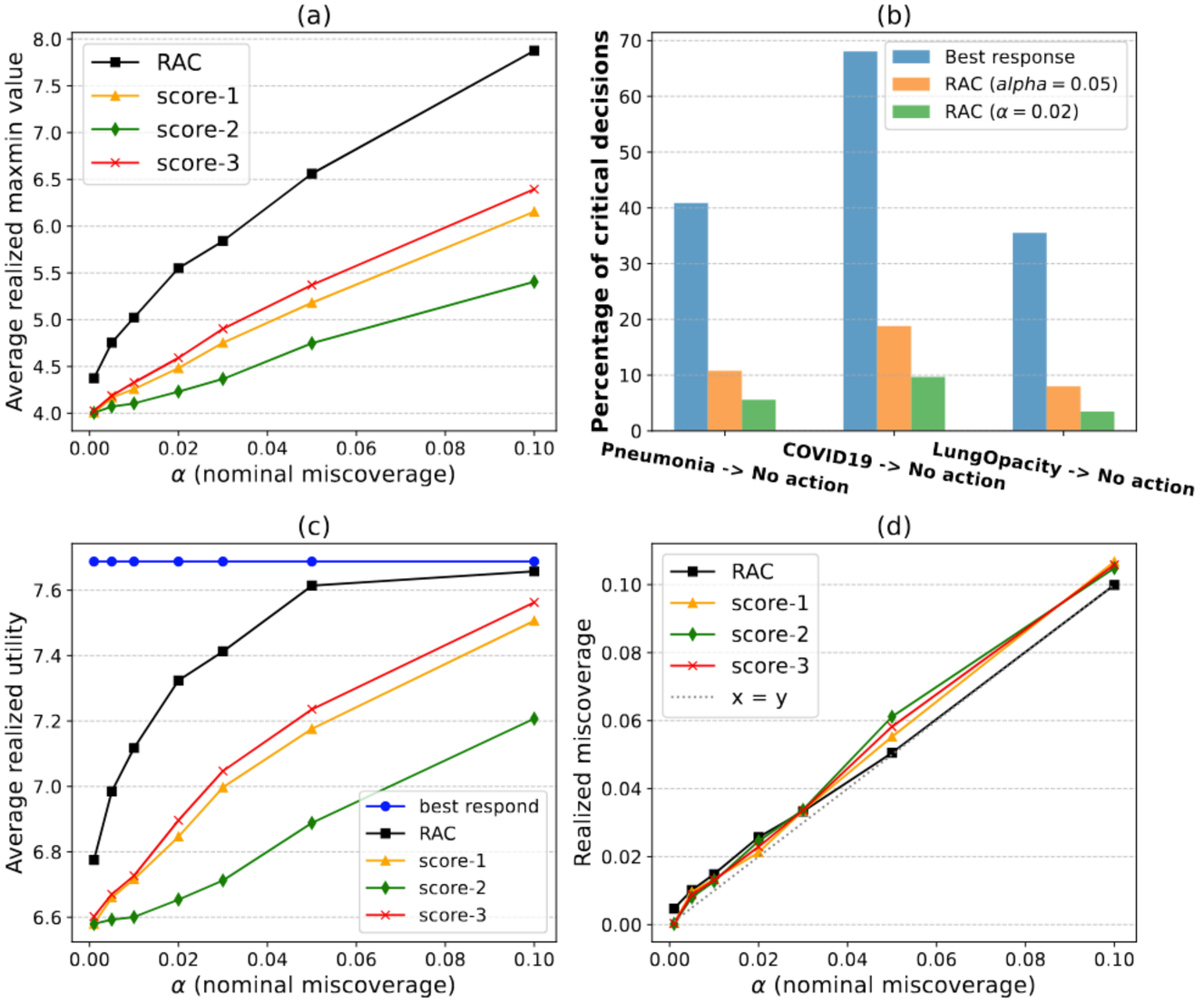
Labels: Normal, COVID,
Lung opacity,
Pneumonia.

Actions: 'No action',
Antibiotics, Quarantine,
Additional Testing

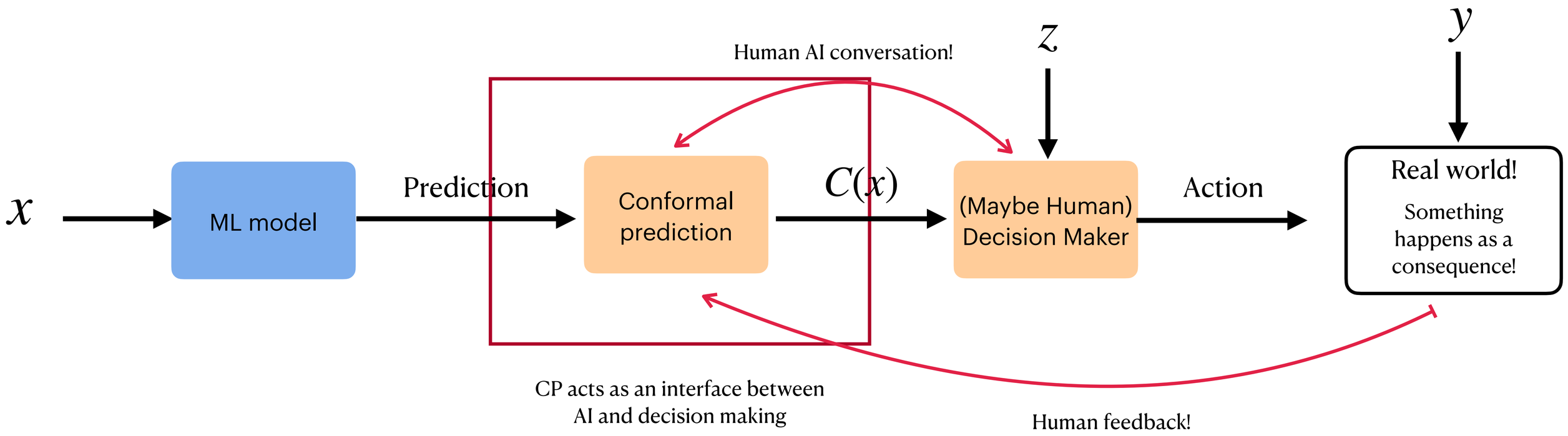
True Label	No Action	Antibiotics	Quarantine	Additional Testing
Normal (0)	10	2	2	4
Pneumonia (1)	0	10	3	7
COVID-19 (2)	0	3	10	8
Lung Opacity (3)	1	4	4	10

How to design utility? Preference functions?
How to handle multiple sequential decisions?

Working with Penn Medicine ...



Human AI Collaborative Decision Making



How can Humans be helpful?!

They have domain knowledge! ——-> z

They can evaluate\determine what is an ideal outcome!

Interactive agreement protocols

Feedbacks!