Human-AI Collaborative Uncertainty Quantification

Sima Noorani

University of Pennsylvania

November 17th, 2025

Joint work with: Shayan Kiyani, George Pappas, Hamed Hassani



Penn
University of Pennsylvania

AI has become powerful

AI has become powerful

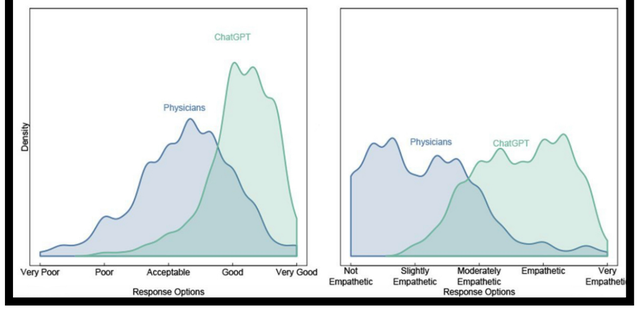
Menu

UC San Diego

UC SAN DIEGO **TODAY**

Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions

While AI won't replace your doctor, the JAMA Internal Medicine paper suggests physicians working together with technologies like ChatGPT may revolutionize medicine



AI has become powerful

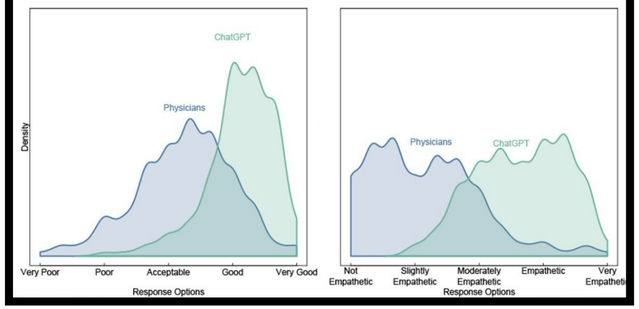
Menu

UC San Diego

UC SAN DIEGO **TODAY**

Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions

While AI won't replace your doctor, the JAMA Internal Medicine paper suggests physicians working together with technologies like ChatGPT may revolutionize medicine



EXPLORE TIME SUBSCRIBE

JUL 2, 2025 10:11 AM ET

Microsoft's AI Is Better Than Doctors at Diagnosing Disease

HEALTH AI

Kilito Chan—Getty Images

by Alice Park
SENIOR CORRESPONDENT

Medicine may be a combination of art and science, but Microsoft just showed that much of both can be learned

AI has become powerful

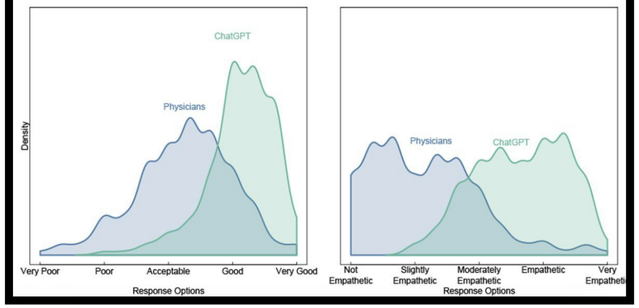
Menu

UC San Diego

UC SAN DIEGO **TODAY**

Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions

While AI won't replace your doctor, the JAMA Internal Medicine paper suggests physicians working together with technologies like ChatGPT may revolutionize medicine



EXPLORE TIME SUBSCRIBE

JUL 2, 2025 10:11 AM ET

Microsoft's AI Is Better Than Doctors at Diagnosing Disease

HEALTH AI

Kilito Chan—Getty Images

by Alice Park
SENIOR CORRESPONDENT

Medicine may be a combination of art and science, but Microsoft just showed that much of both can be learned

Stanford University

Stanford Report

January 25th, 2017 | 6 min read

Science & Engineering

Deep learning algorithm does as well as dermatologists in identifying skin cancer

In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.

It's scary enough making a doctor's appointment to see if a strange mole could be

AI has become powerful

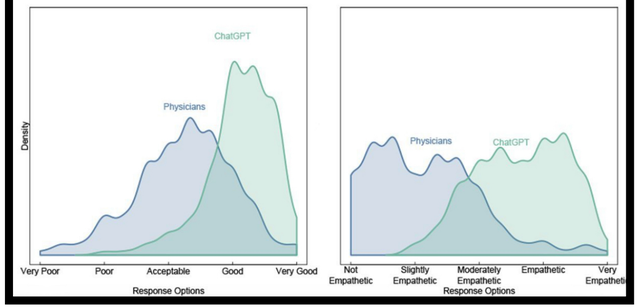
Menu

UC San Diego

UC SAN DIEGO **TODAY**

Study Finds ChatGPT Outperforms Physicians in High-Quality, Empathetic Answers to Patient Questions

While AI won't replace your doctor, the JAMA Internal Medicine paper suggests physicians working together with technologies like ChatGPT may revolutionize medicine



EXPLORE TIME SUBSCRIBE

JUL 2, 2025 10:11 AM ET

Microsoft's AI Is Better Than Doctors at Diagnosing Disease

HEALTH AI

Kilito Chan—Getty Images

by Alice Park
SENIOR CORRESPONDENT

Medicine may be a combination of art and science, but Microsoft just showed that much of both can be learned

Stanford University

Stanford Report

January 25th, 2017 | 6 min read

Science & Engineering

Deep learning algorithm does as well as dermatologists in identifying skin cancer

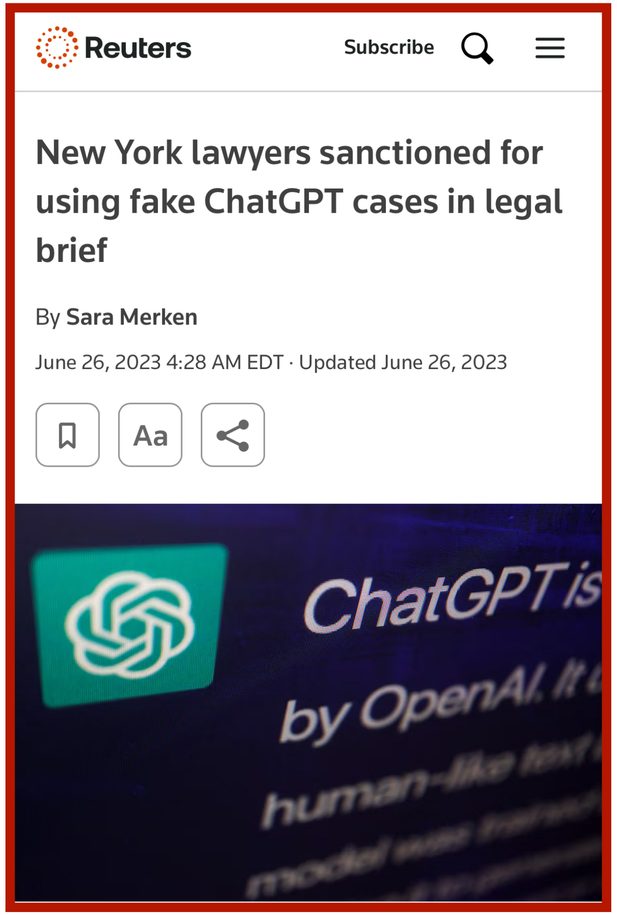
In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.

It's scary enough making a doctor's appointment to see if a strange mole could be

Question: Why not replace human decision makers with AI already?

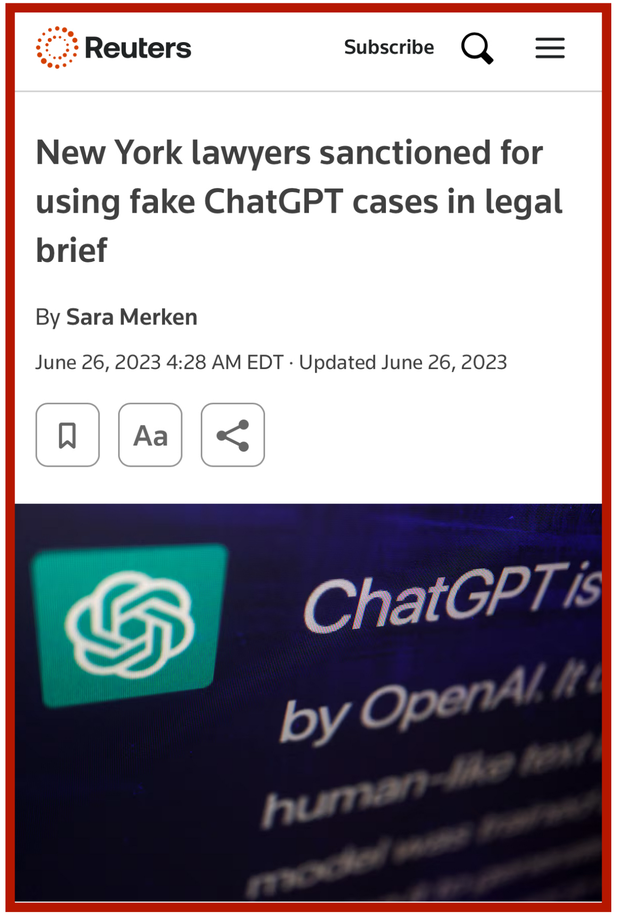
Question: Why not replace human decision makers with AI already?

Question: Why not replace human decision makers with AI already?



Question: Why not replace human decision makers with AI already?

Hallucinations



Question: Why not replace human decision makers with AI already?

Hallucinations

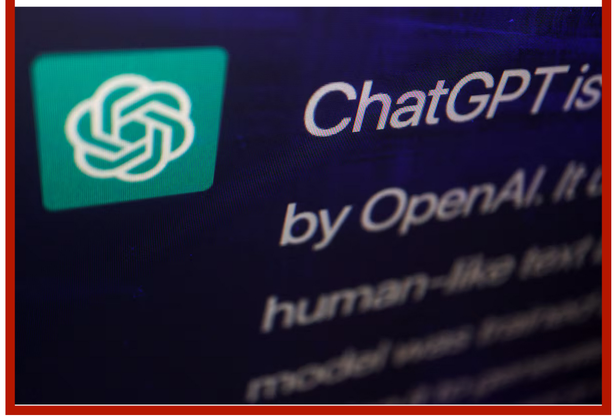
Reuters Subscribe

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By Sara Merken

June 26, 2023 4:28 AM EDT · Updated June 26, 2023

Aa



The damage AI hallucinations can do – and how to avoid them

'Even if these systems are right 80% of the time, that still means they're wrong 20% of the time,' says tech CMO Dr. Jay Anders, who describes the risks of artificial intelligence errors and outlines some protection strategies for providers.

AI & ML Intelligence | Global

By Bill Siwicki, *Managing Editor* | July 24, 2025 | 11:53 AM

Question: Why not replace human decision makers with AI already?

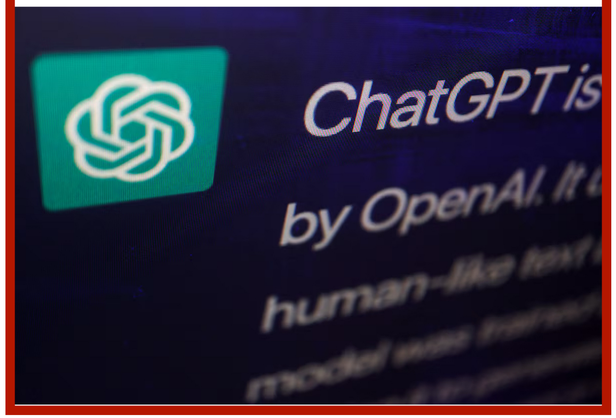
Hallucinations

Reuters Subscribe

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By **Sara Merken**

June 26, 2023 4:28 AM EDT · Updated June 26, 2023



The damage AI

AI also struggles with context. If I'm discussing a physical exam, it might introduce elements that have nothing to do with physical examinations. It loses track of what we're actually talking about.

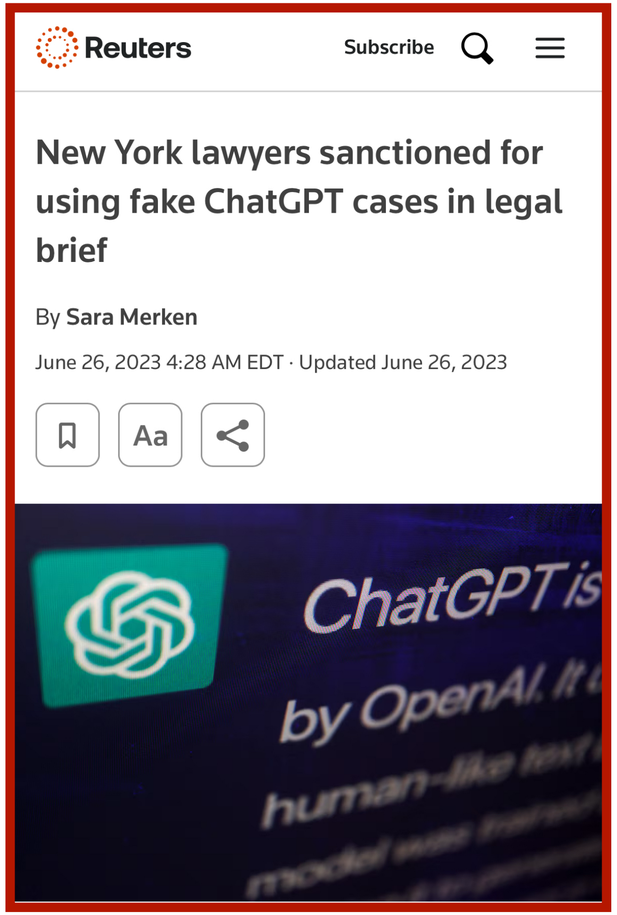
describes the risks of artificial intelligence errors and outlines some protection strategies for providers.

AI & ML Intelligence | Global

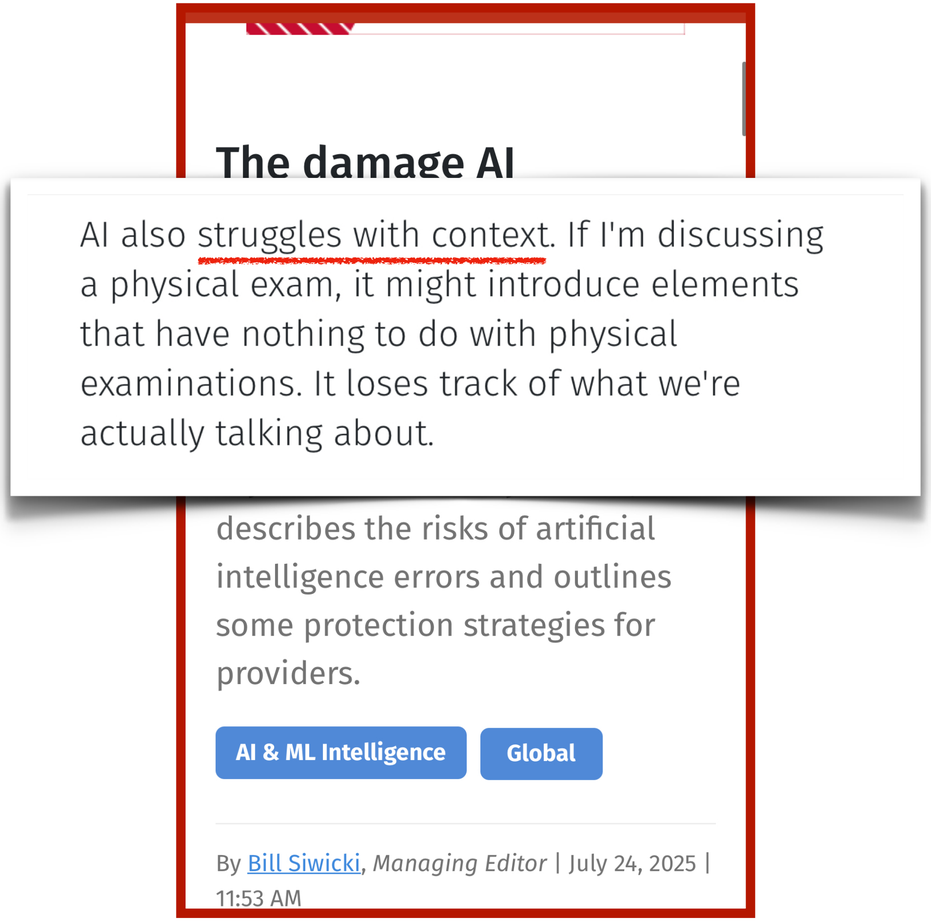
By Bill Siwicki, *Managing Editor* | July 24, 2025 | 11:53 AM

Question: Why not replace human decision makers with AI already?

Hallucinations



Struggles with long-term context



Question: Why not replace human decision makers with AI already?

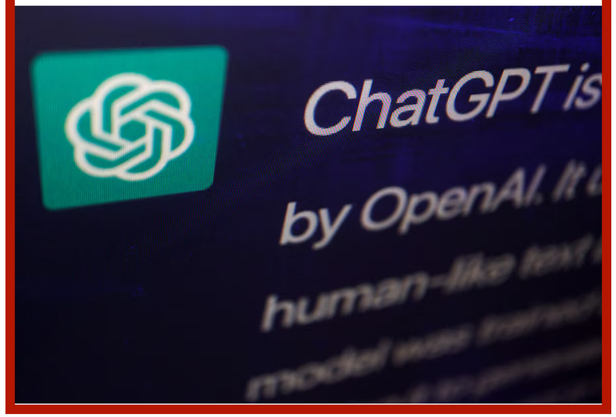
Hallucinations

Reuters Subscribe

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By **Sara Merken**

June 26, 2023 4:28 AM EDT · Updated June 26, 2023



Struggles with long-term context

The damage AI

AI also struggles with context. If I'm discussing a physical exam, it might introduce elements that have nothing to do with physical examinations. It loses track of what we're actually talking about.

describes the risks of artificial intelligence errors and outlines some protection strategies for providers.

AI & ML Intelligence | Global

By Bill Siwicki, *Managing Editor* | July 24, 2025 | 11:53 AM

AXIOS Pittsburgh

Aug 27, 2025 - News

AI is overconfident even when wrong, says report

Ryan Deto

Add Axios on Google

Illustration: Allie Carl/Axios

Question: Why not replace human decision makers with AI already?

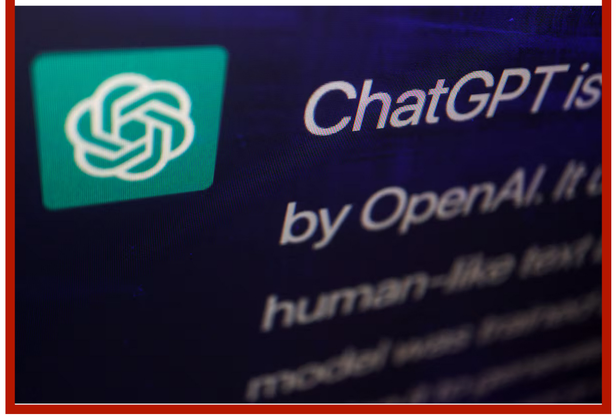
Hallucinations

Reuters Subscribe

New York lawyers sanctioned for using fake ChatGPT cases in legal brief

By **Sara Merken**

June 26, 2023 4:28 AM EDT · Updated June 26, 2023



Struggles with long-term context

The damage AI

AI also struggles with context. If I'm discussing a physical exam, it might introduce elements that have nothing to do with physical examinations. It loses track of what we're actually talking about.

describes the risks of artificial intelligence errors and outlines some protection strategies for providers.

AI & ML Intelligence | Global

By Bill Siwicki, *Managing Editor* | July 24, 2025 | 11:53 AM

Unaware when uncertain

AXIOS Pittsburgh

Aug 27, 2025 - News

AI is overconfident even when wrong, says report

Ryan Deto

Add Axios on Google

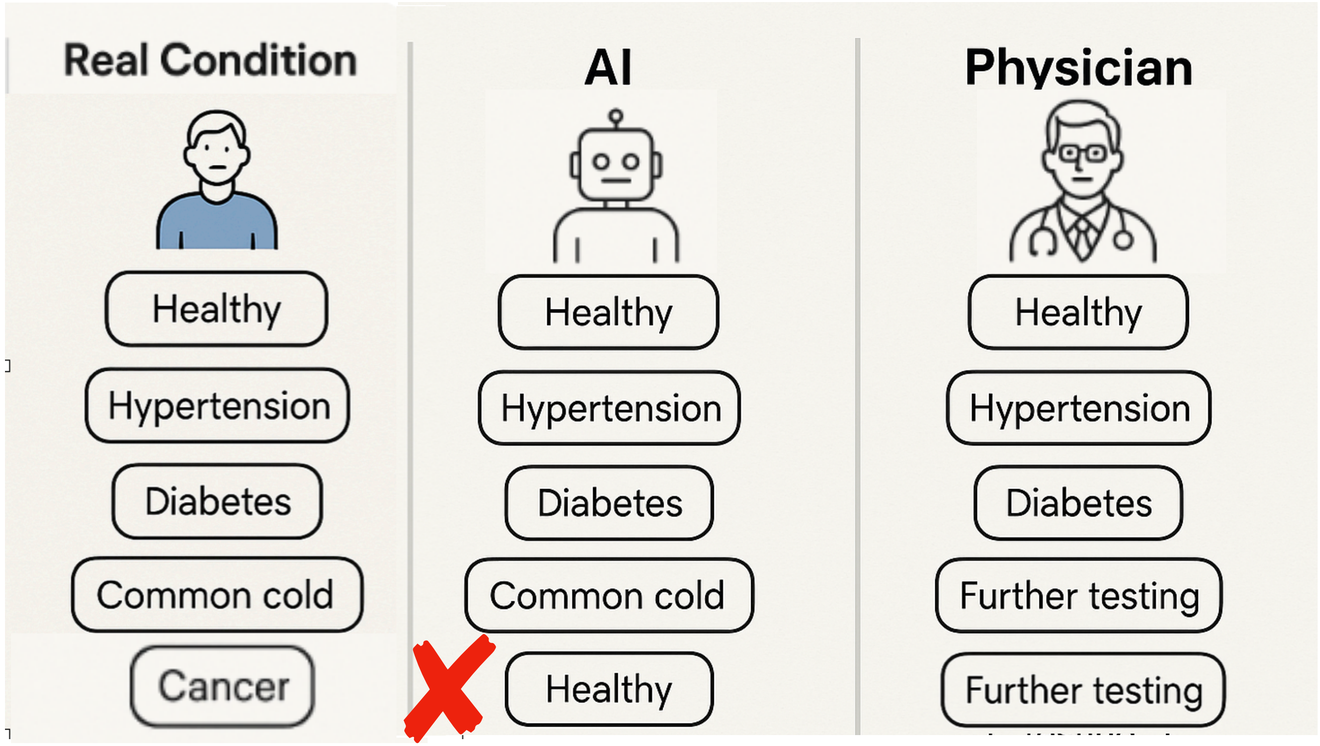
Illustration: Allie Carl/Axios

Question: Why not replace human decision makers with AI already?

Hallucinations

Struggles with long-term context

Unaware when uncertain

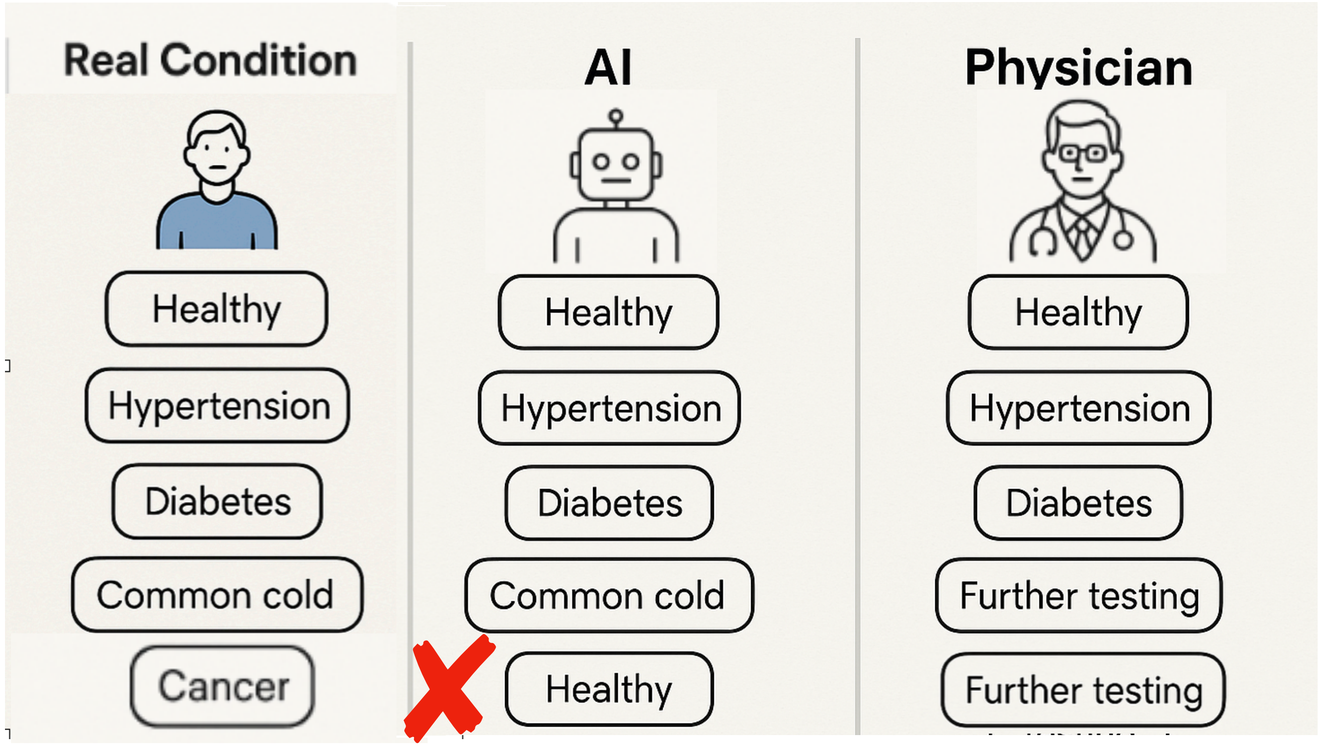


Question: Why not replace human decision makers with AI already?

Hallucinations

Struggles with long-term context

Unaware when uncertain



Reliability matters!

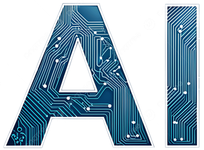
Robust decision making requires more than predictive accuracy!

Question: Why not replace human decision makers with AI already?

Question: Why not replace human decision makers with AI already?

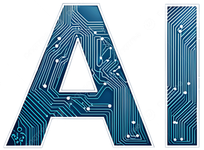
Human and AI should coexist in the decision making pipeline

Human and AI should coexist in the decision making pipeline



Handle large unstructured data

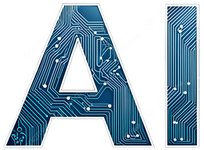
Human and AI should coexist in the decision making pipeline



Handle large unstructured data

Excel at patten extraction

Human and AI should coexist in the decision making pipeline



Handle large unstructured data

Excel at patten extraction

Offer statistical accuracy

Human and AI should coexist in the decision making pipeline

AI

Handle large unstructured data

Excel at patten extraction

Offer statistical accuracy

Domain Knowledge

Human and AI should coexist in the decision making pipeline

AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

- Domain Knowledge
- Persistent Memory

Human and AI should coexist in the decision making pipeline

AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

- Domain Knowledge
- Persistent Memory
- Reason and act in the physical world

Human and AI should coexist in the decision making pipeline

AI

- Handle large unstructured data
- Excel at patten extraction
- Offer statistical accuracy

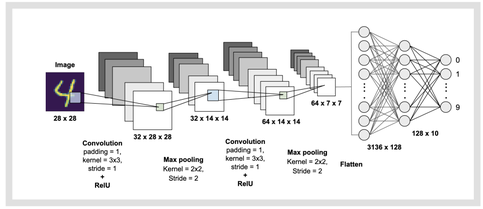
- Domain Knowledge
- Persistent Memory
- Reason and act in the physical world

Human and AI should jointly exist in the decision making process!

Decision making pipeline

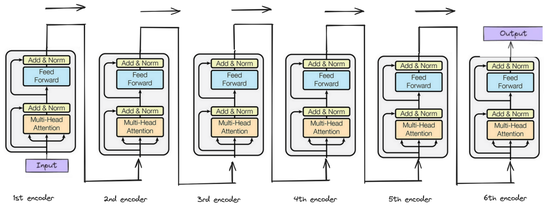
Decision making pipeline

Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions



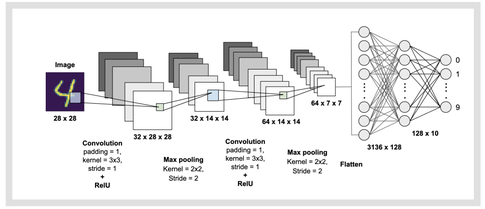
- (Heavily) Pre-trained

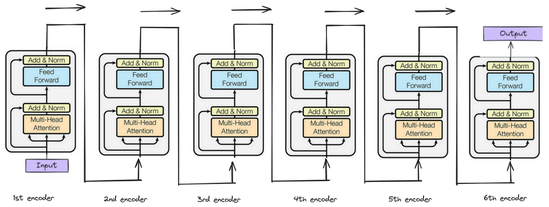
- Possibility of little tweaks using extra data (fine-tuning)



Decision making pipeline

Inputs/prompts → AI model → Predictions/Recommendations → Decision Maker → Actions





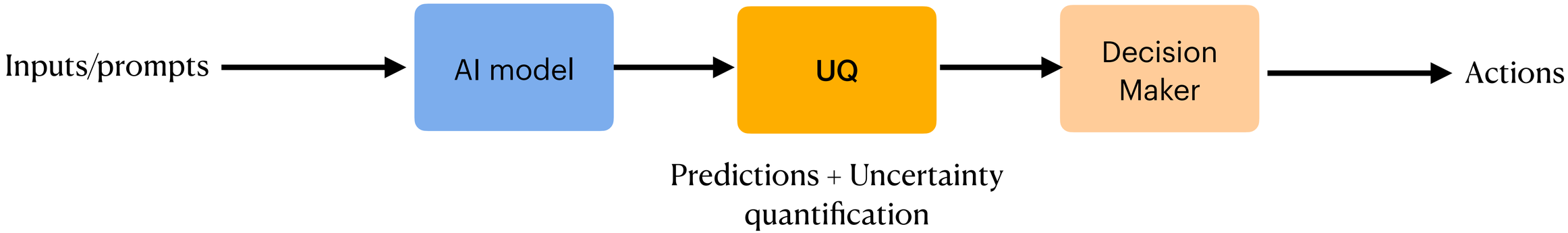
- (Heavily) Pre-trained

- Possibility of little tweaks using extra data (fine-tuning)

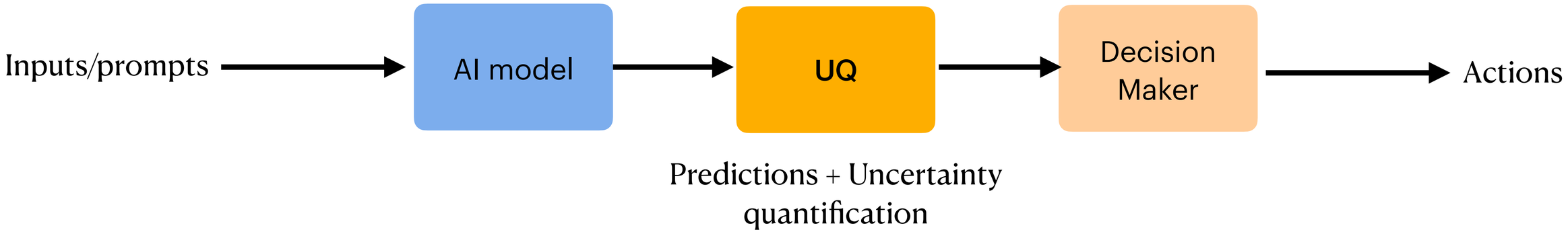
- Automated

- Humans

Robust decision making pipeline



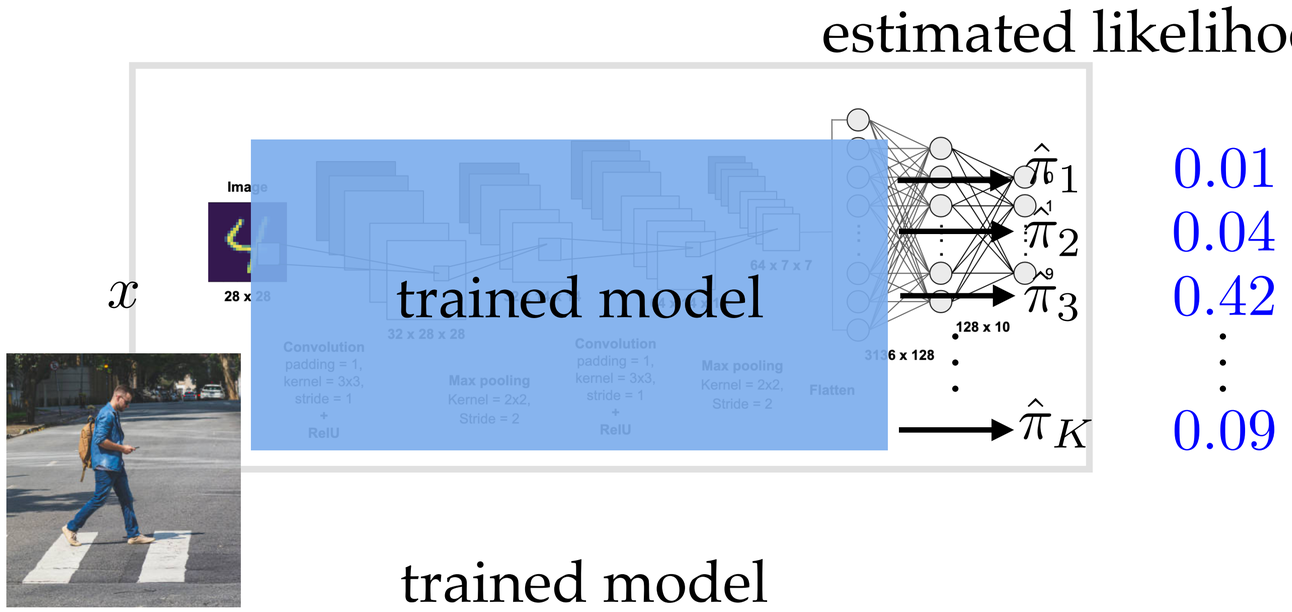
Robust decision making pipeline



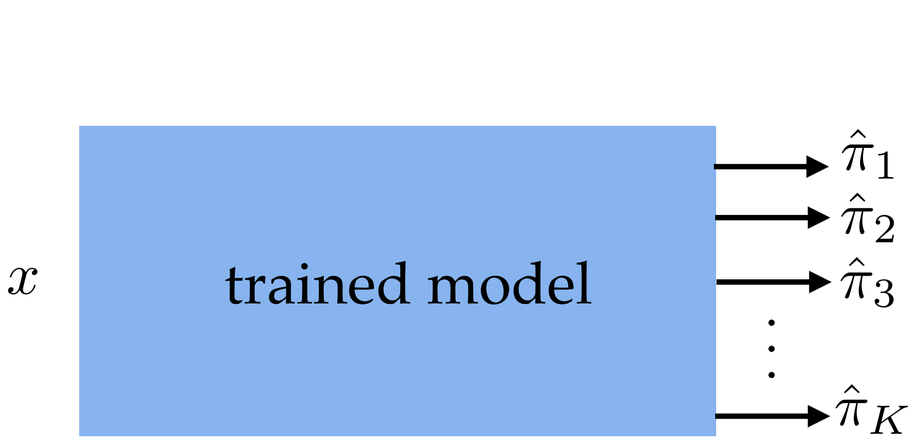
Robust Decision making requires precise Uncertainty Quantification

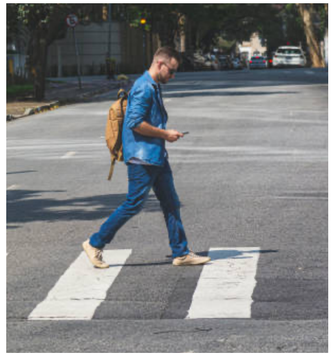
(Quick) Review of Conformal Prediction

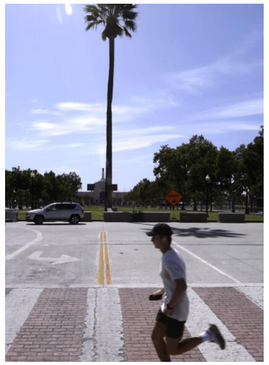
Conformal Prediction

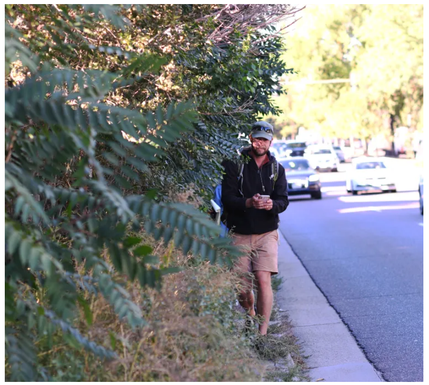


- prediction is then based on the class that has maximum likelihood









{ human 0.99 }

{ human, tree, bin
0.90 0.08 0.02 }

{ Tree, human, trash can
0.70 0.20 0.05 }

easy

less easy

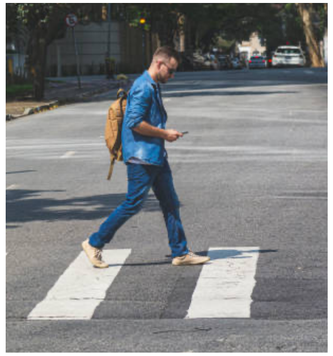
difficult

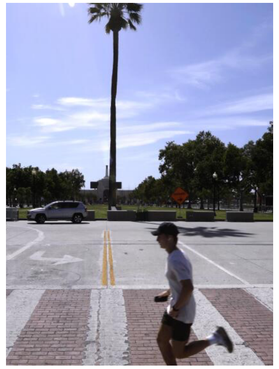
model's prediction: human ✅ human ✅ tree ❌

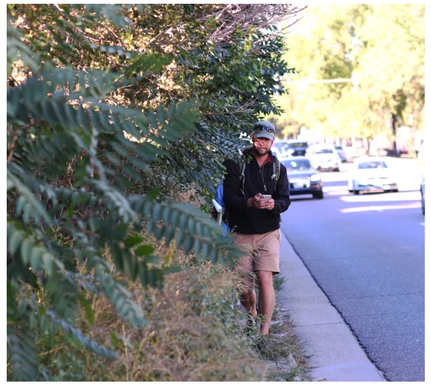
- the trained model provides estimated likelihoods (a notion of uncertainty)

- these likelihoods are informative but not always correct

x



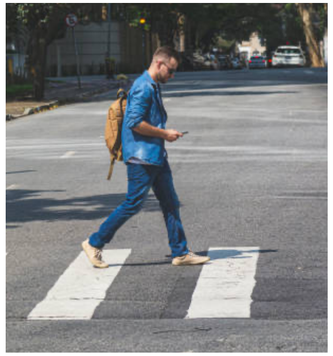


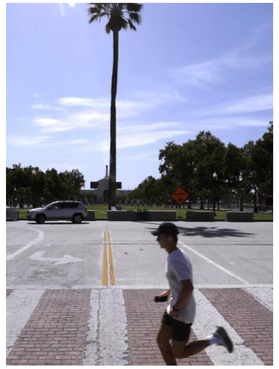


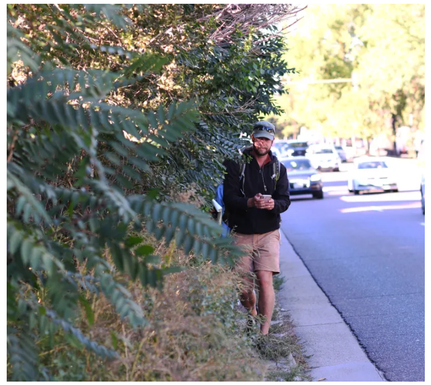
{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

model's prediction: human ✅ human ✅ tree ❌

x







{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

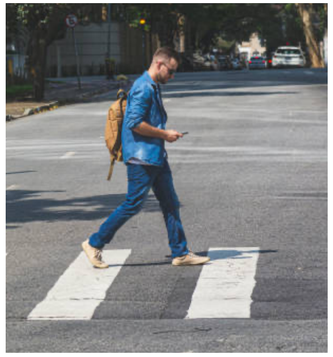
$C(x)$

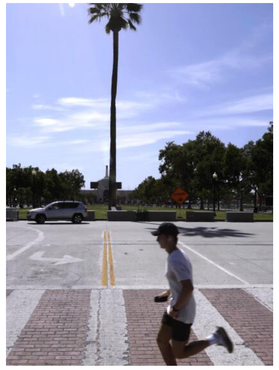
(model's conformal prediction)

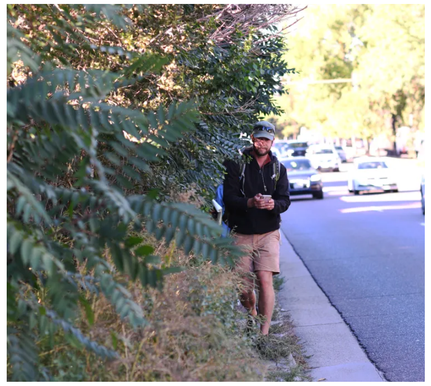
{ human } ✅ { human } ✅ { tree, human } ✅

Single prediction → set of predictions

x







{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

$C(x)$

(model's conformal prediction)

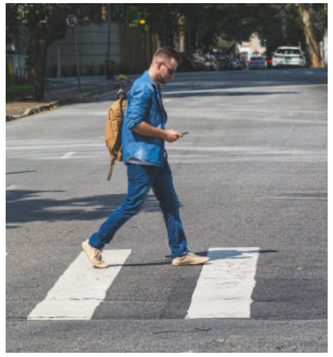
{ human } ✅ { human } ✅ { tree, human } ✅

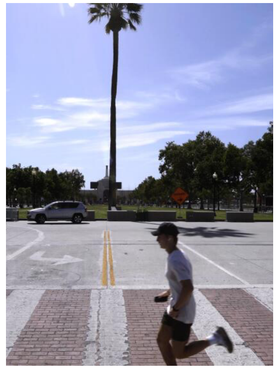
Marginal Coverage

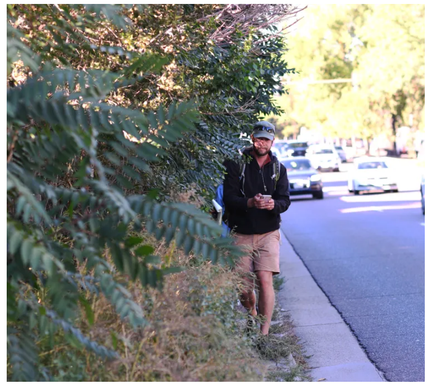
$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

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

x







{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

$C(x)$ (model's conformal prediction)

{ human } ✅ { human } ✅ { tree , human } ✅

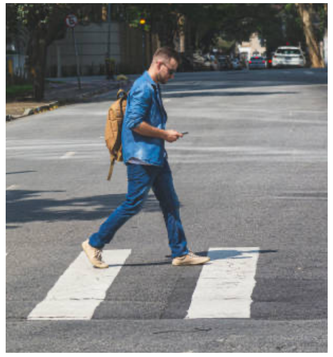
Marginal Coverage

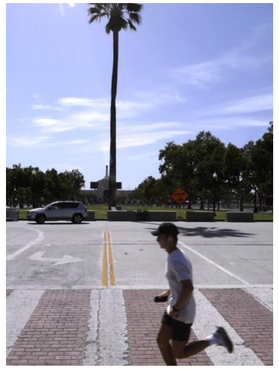
$$\Pr\{Y_{\text{test}} \in C(X_{\text{test}})\} \geq 1 - \alpha$$

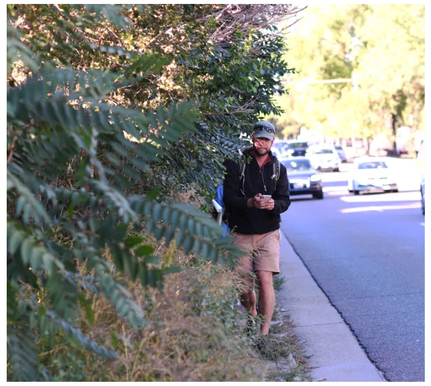
$$C(x) = \{y \in \mathcal{Y} : \quad S(x, y) \leq q\}$$

the threshold

x







{ human 0.99 } { human 0.90, tree 0.08, bin 0.02 } { Tree 0.70, human 0.20, trash can 0.05 }

$C(x)$
(model's conformal prediction)

{ human } ✅ { human } ✅ { tree, human } ✅

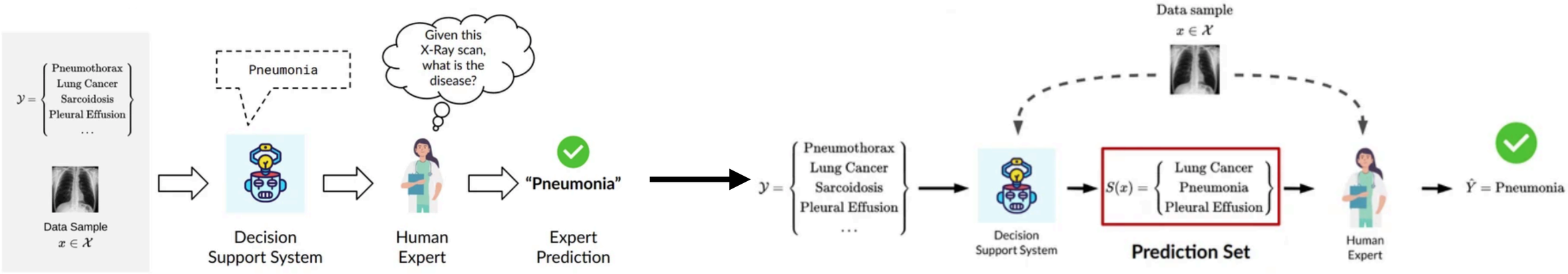
$|C(x)|$
size

1 1 2

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

Prediction sets are useful for decision makers

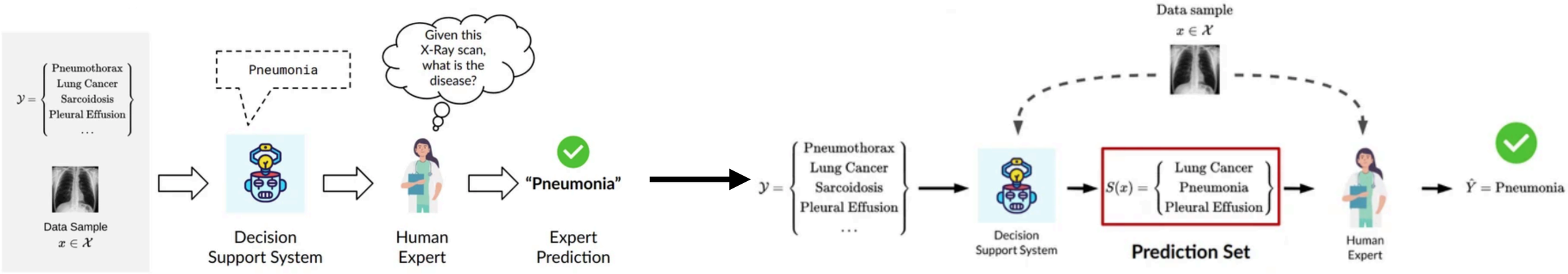
Prediction sets are useful for decision makers



Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

Prediction sets are useful for decision makers



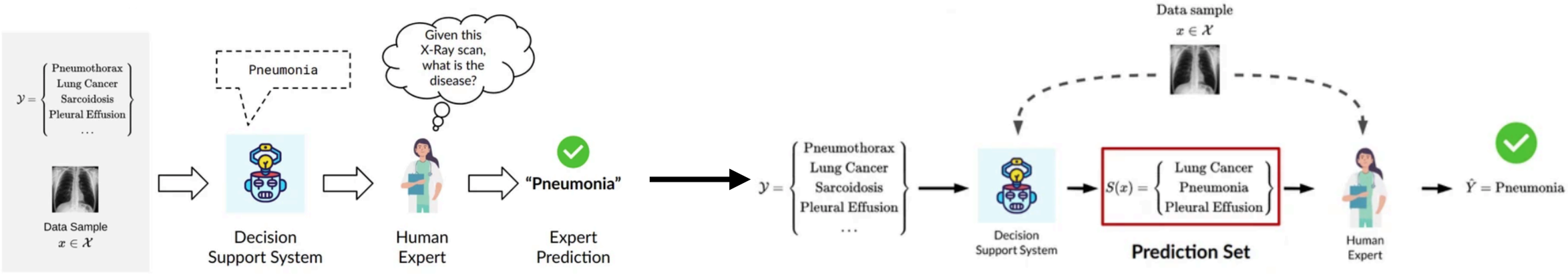
Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

"Decision-theoretic framework for evaluating predictive uncertainty as informative signals"

Hullman et. al.., *Conformal Prediction and Human Decision Making*, 2025.

Prediction sets are useful for decision makers



Straitouri et al., Improving Expert Predictions with Conformal Prediction, ICML, 2023.

E. Straitouri & M. Gomez-Rodriguez, *Designing Decision Support Systems using Counterfactual Prediction Sets*, ICML, 2024.

"Decision-theoretic framework for evaluating predictive uncertainty as informative signals"

Hullman et. al.., *Conformal Prediction and Human Decision Making*, 2025.

"Well designed" prediction sets are a sufficient statistic for risk averse decision making

S. Kiyani et Al., *Decision Theoretic Foundations for Conformal Prediction: Optimal Uncertainty Quantification for Risk-Averse Agents*, ICML, 2025.

Thus far …

Thus far …

Human and AI should coexist in the decision making pipeline

Thus far …

- ✓ **Human and AI should coexist in the decision making pipeline**

- ✓ **UQ is essential in the decision making pipeline**

UQ

Thus far …

- ✓ **Human and AI should coexist in the decision making pipeline**

- ✓ **UQ is essential in the decision making pipeline**

UQ

- ✓ **CP is a promising tool for UQ of AI decision support systems**

Thus far …

- ✓ **Human and AI should coexist in the decision making pipeline**

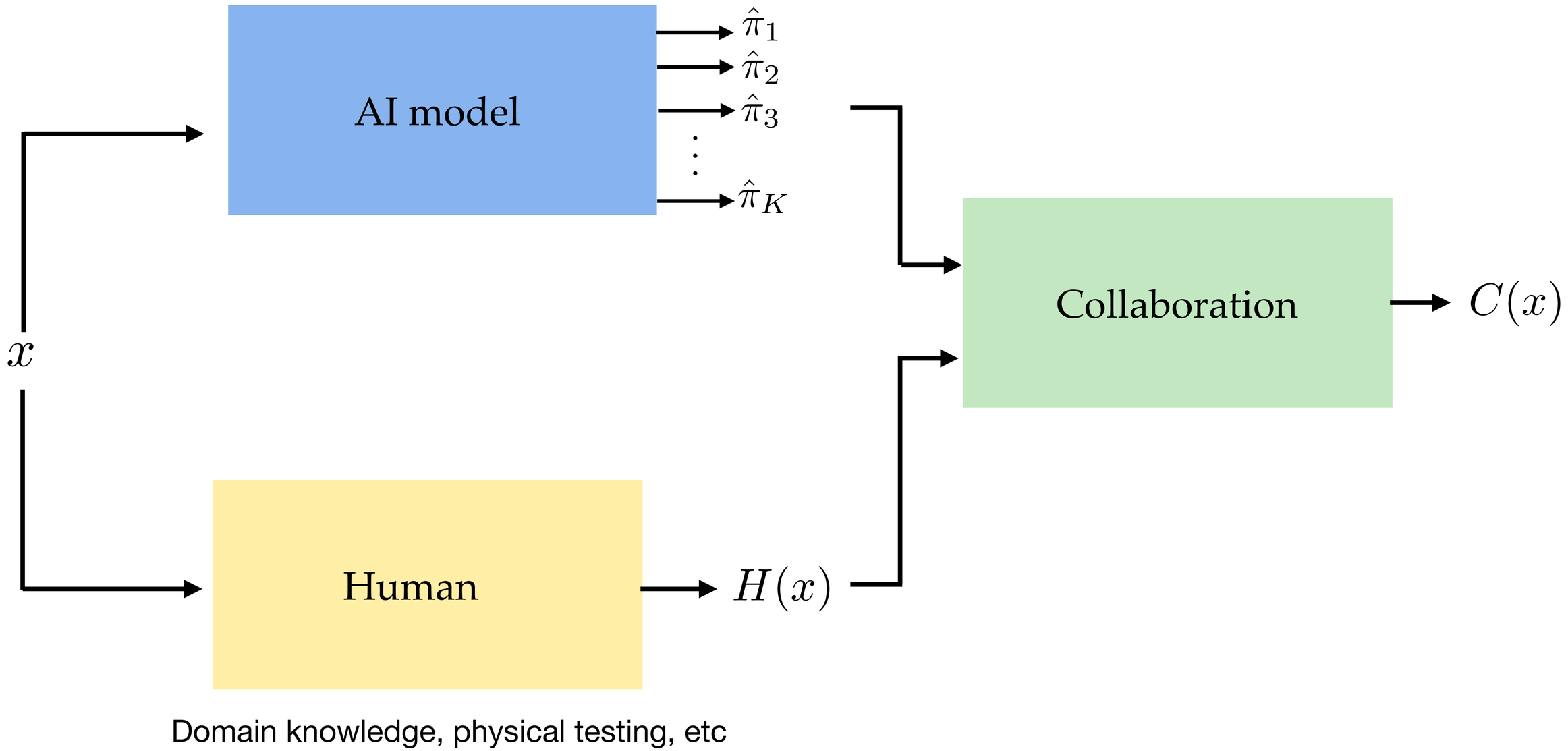
- ✓ **UQ is essential in the decision making pipeline**

UQ

- ✓ **CP is a promising tool for UQ of AI decision support systems**

What should be the principles of UQ when Human and AI are jointly in the loop?

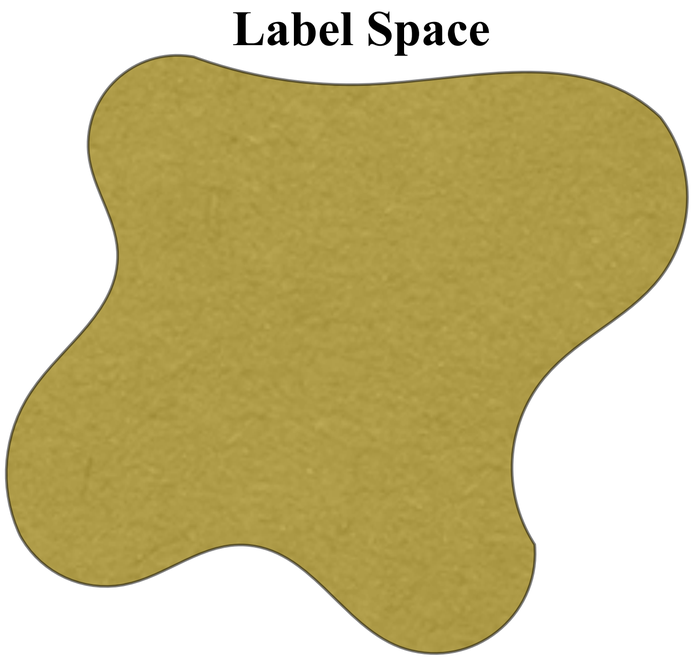
Collaborative Uncertainty Quantification pipeline



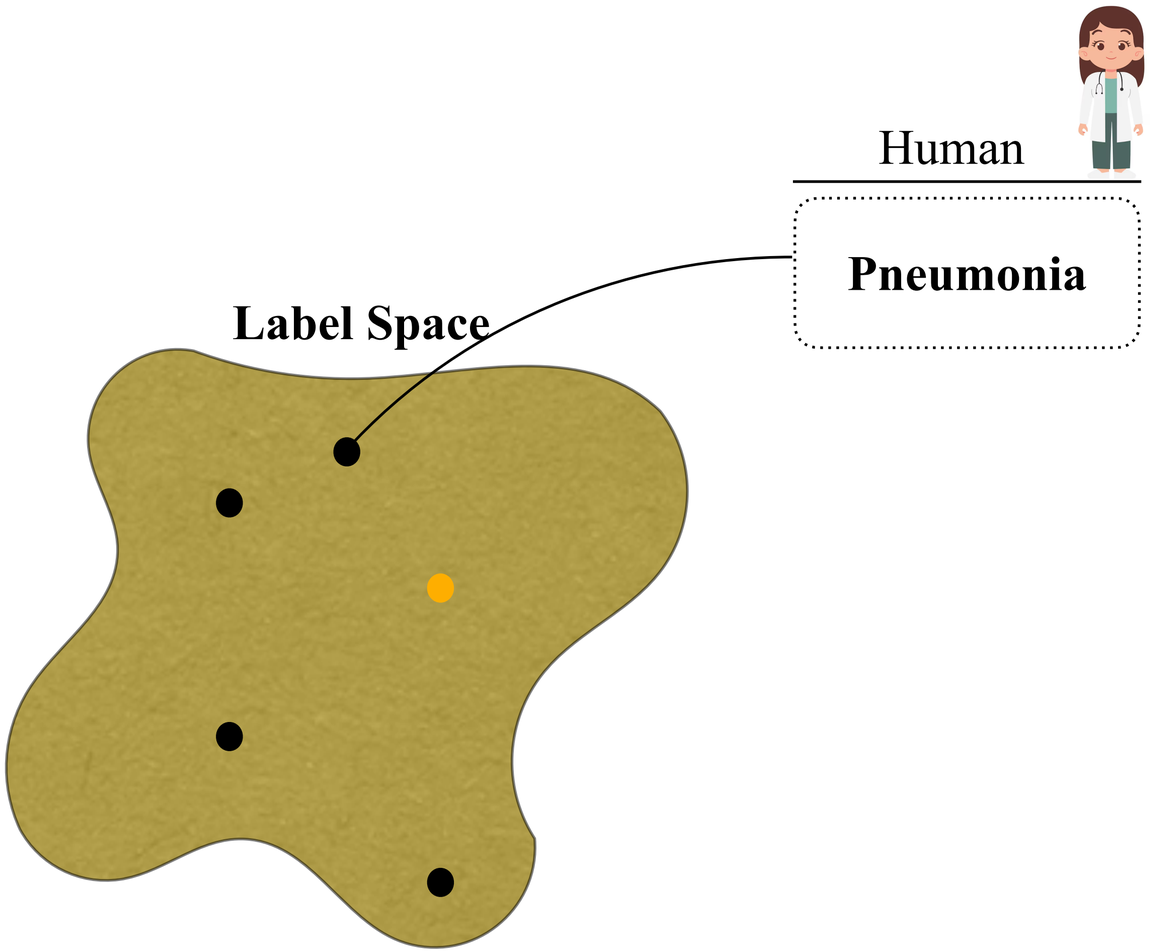
Question: what constitutes a good collaboration?

Question: what constitutes a good collaboration?

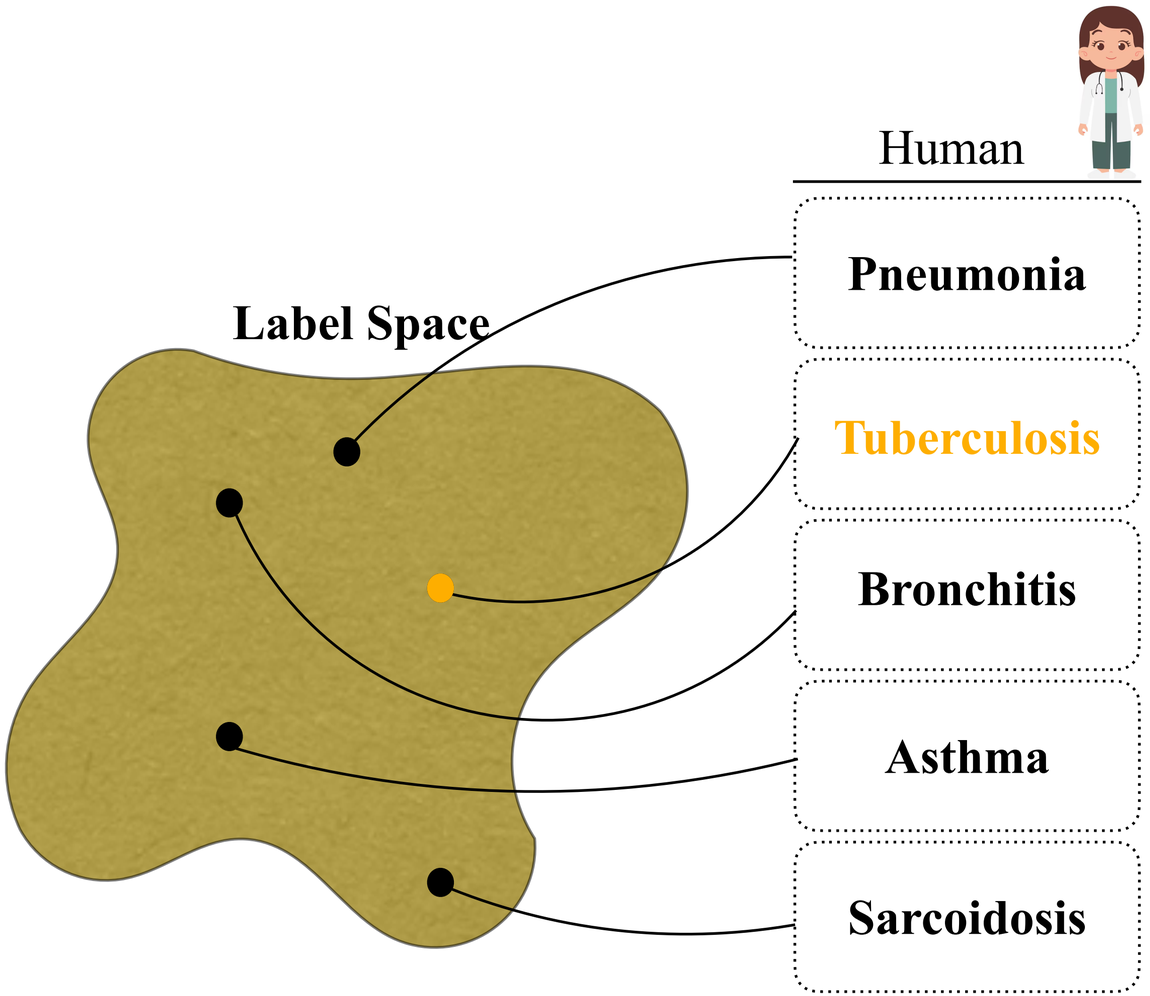
Question: what constitutes a good collaboration?



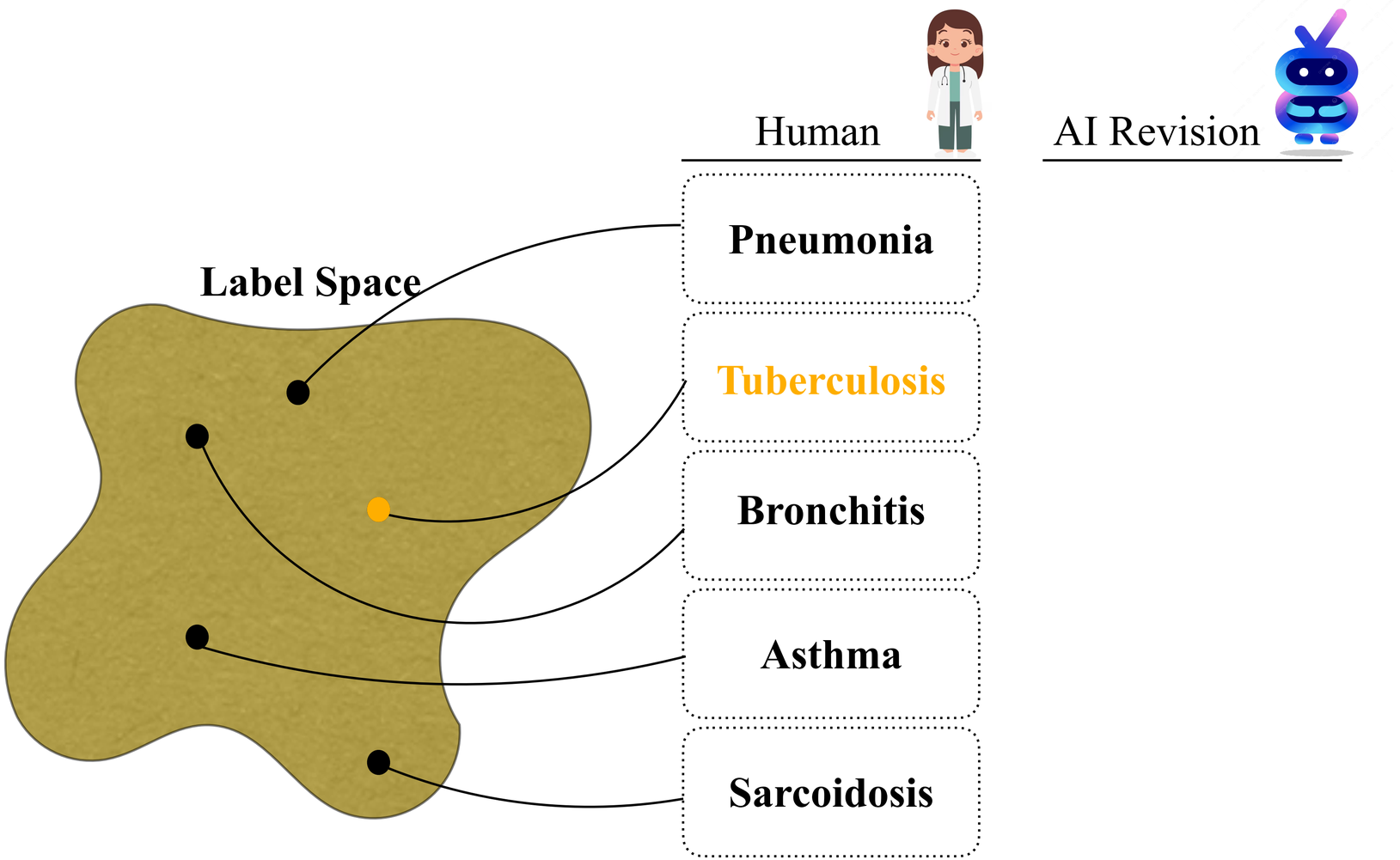
Question: what constitutes a good collaboration?



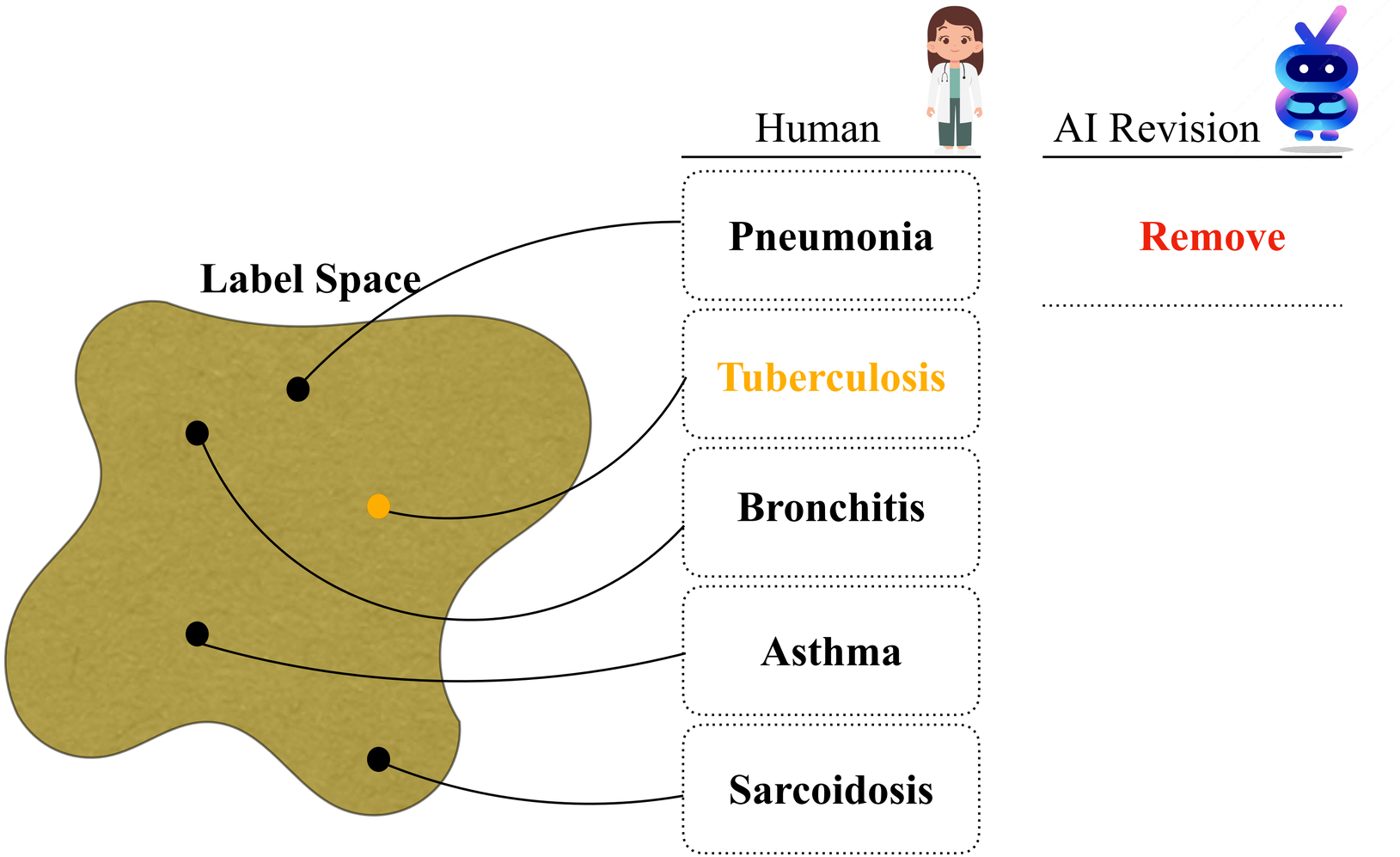
Question: what constitutes a good collaboration?



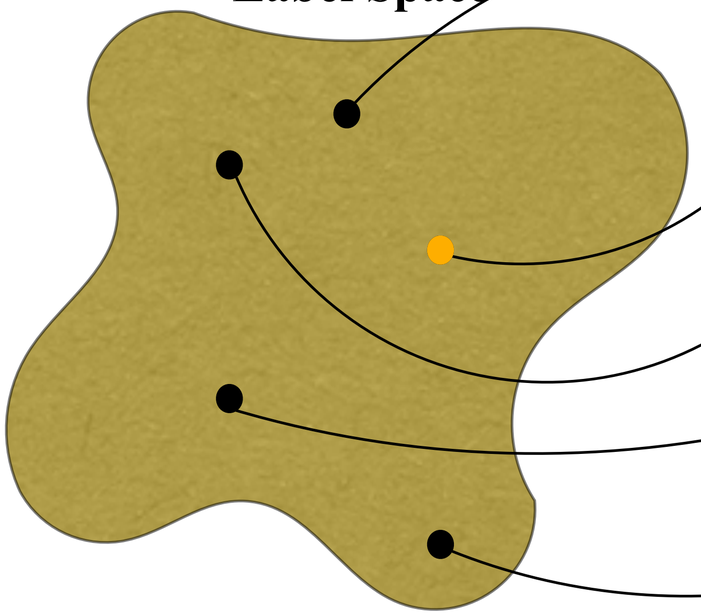
Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

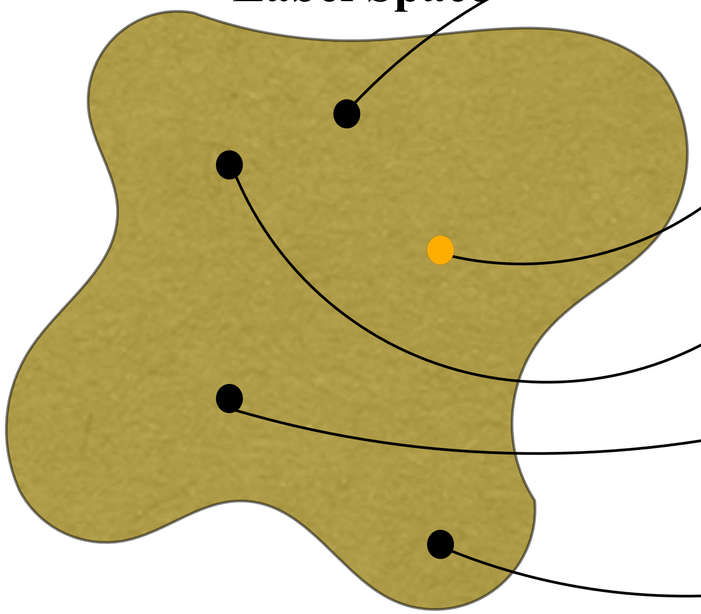


Question: what constitutes a good collaboration?



Human	AI Revision
Pneumonia	Remove
Tuberculosis	Keep
Bronchitis	Remove
Asthma	Keep
Sarcoidosis	Remove

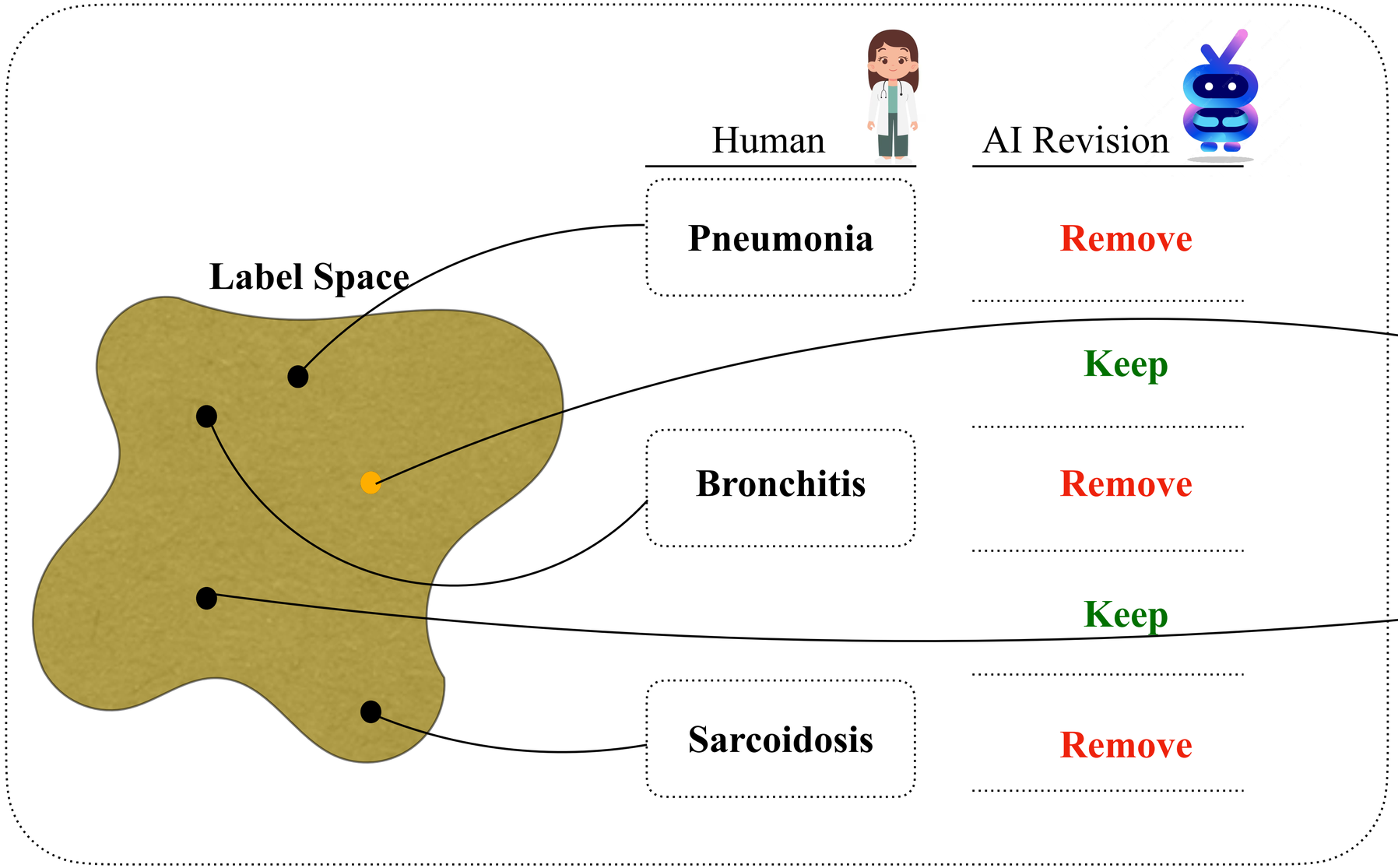
Question: what constitutes a good collaboration?



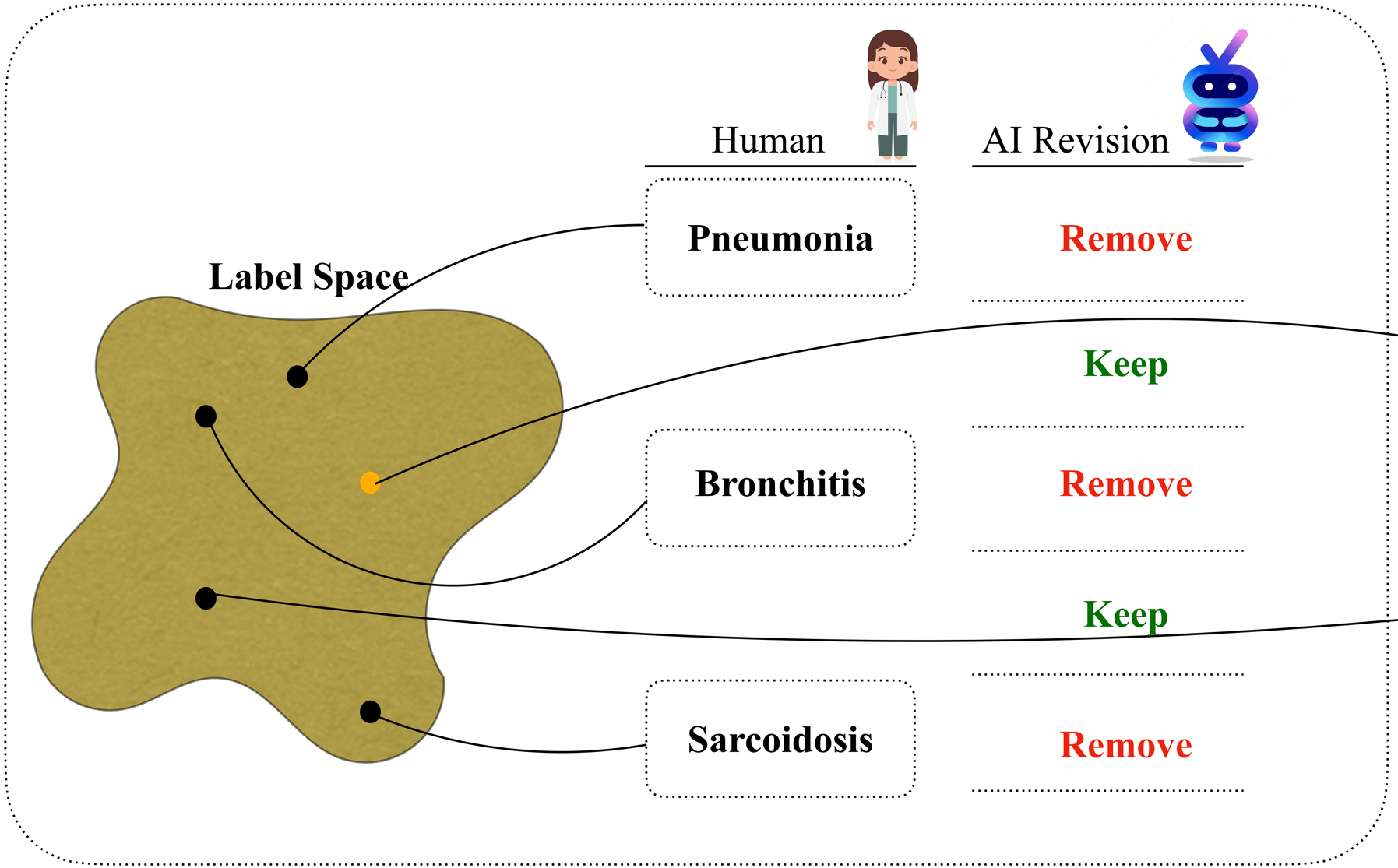
Human	AI Revision
Pneumonia	Remove
Tuberculosis	Keep
Bronchitis	Remove
Asthma	Keep
Sarcoidosis	Remove

Collaborative Proposal

Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

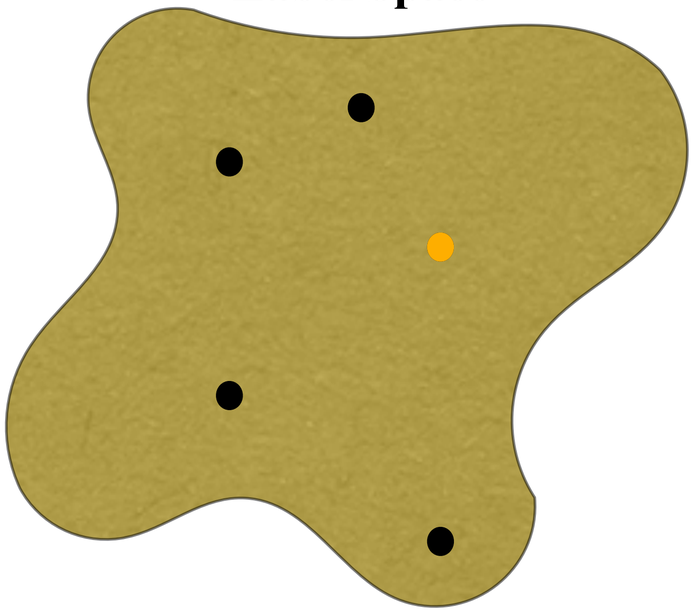


First: dont cause harm!

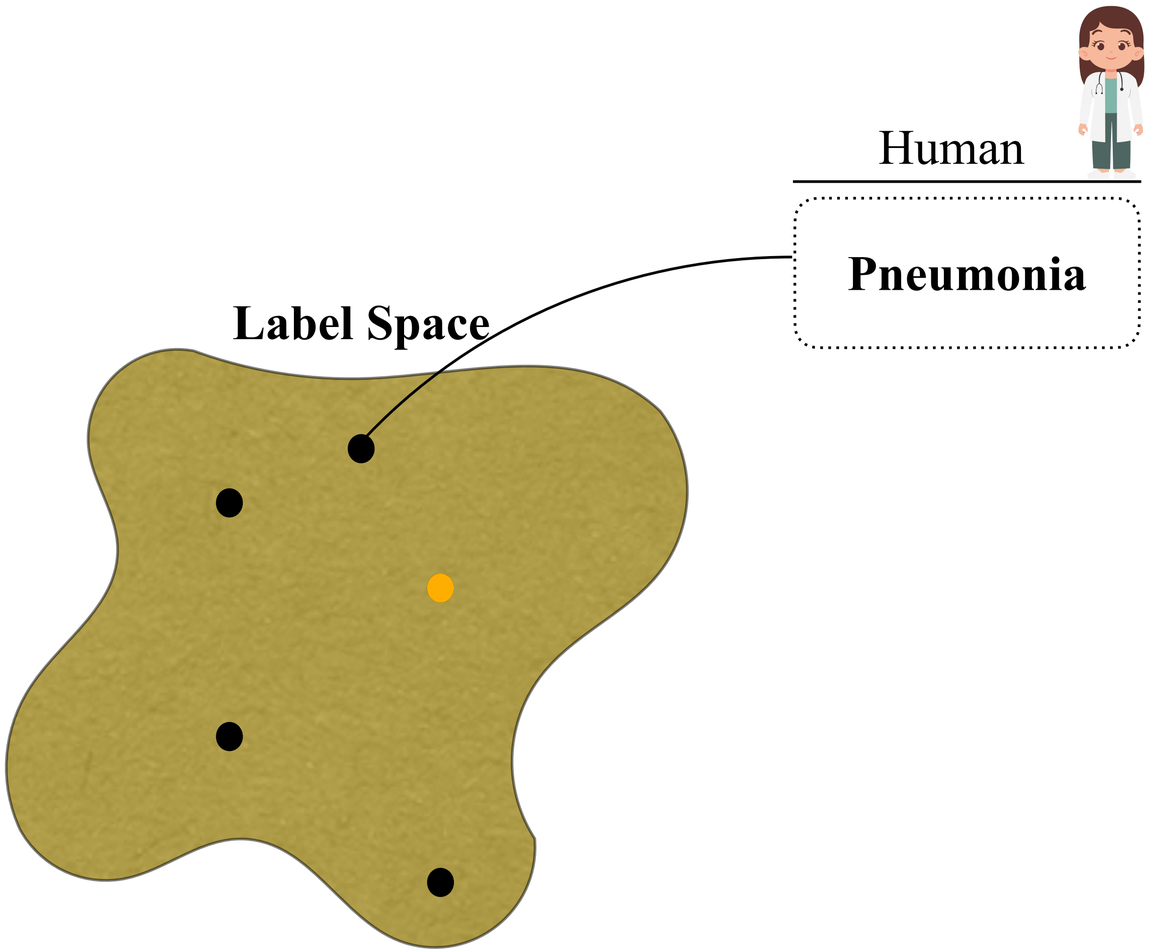
Question: what constitutes a good collaboration?

Human 

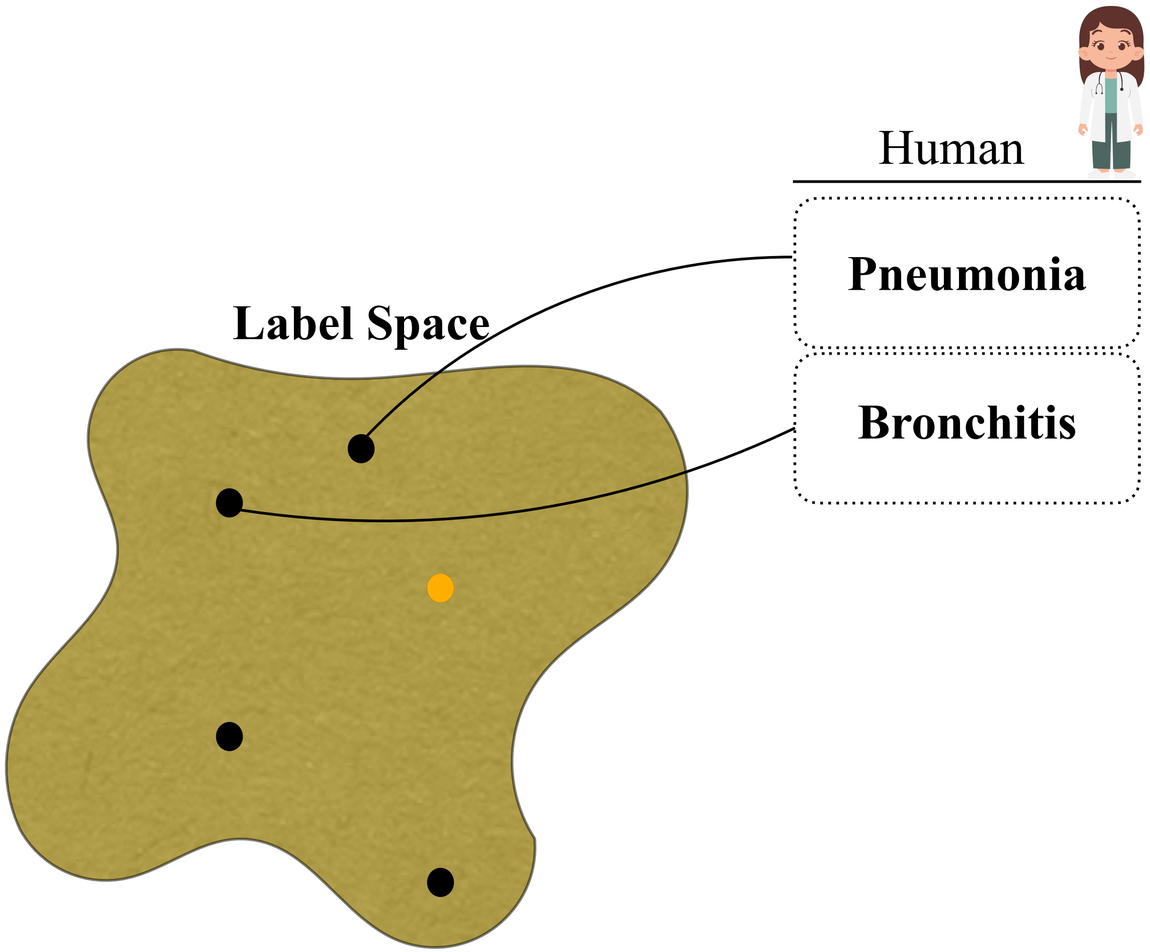
Label Space



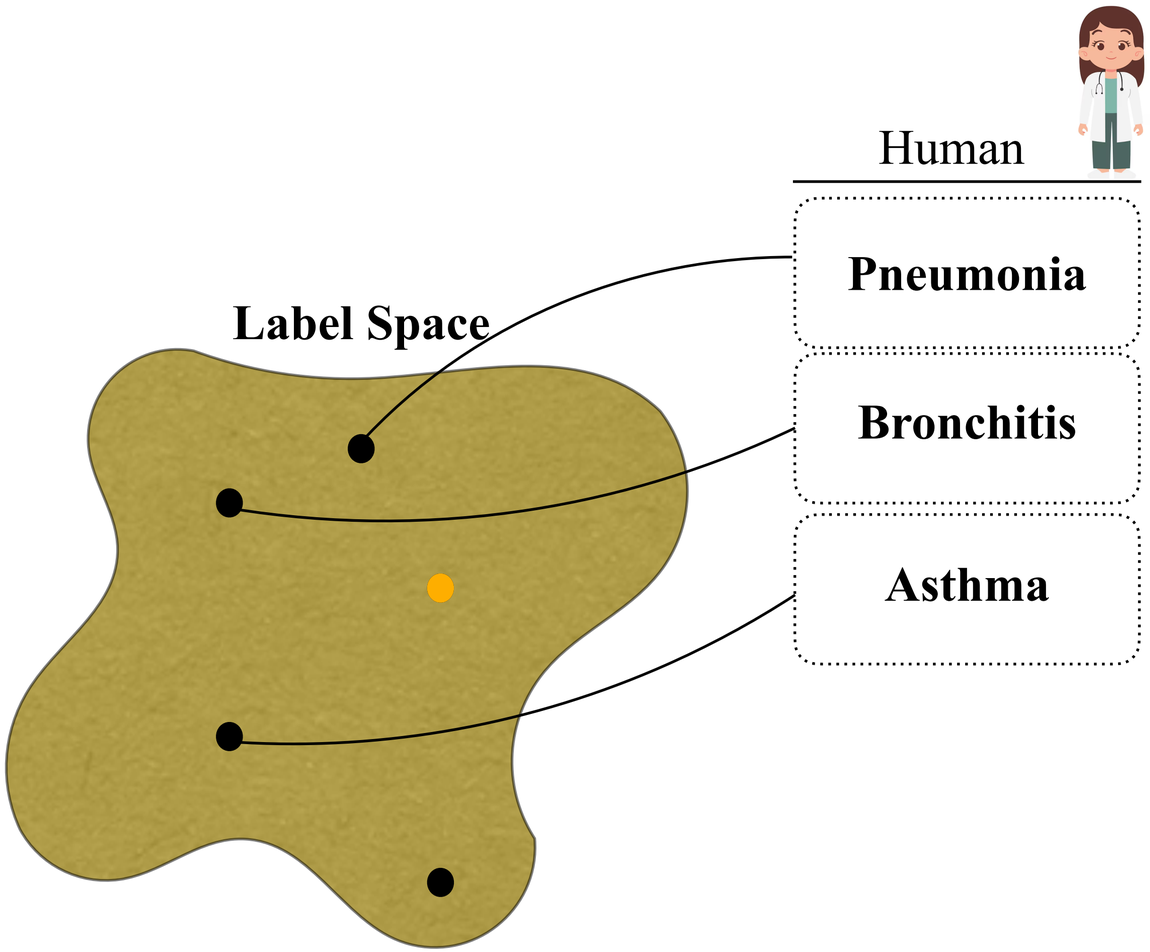
Question: what constitutes a good collaboration?



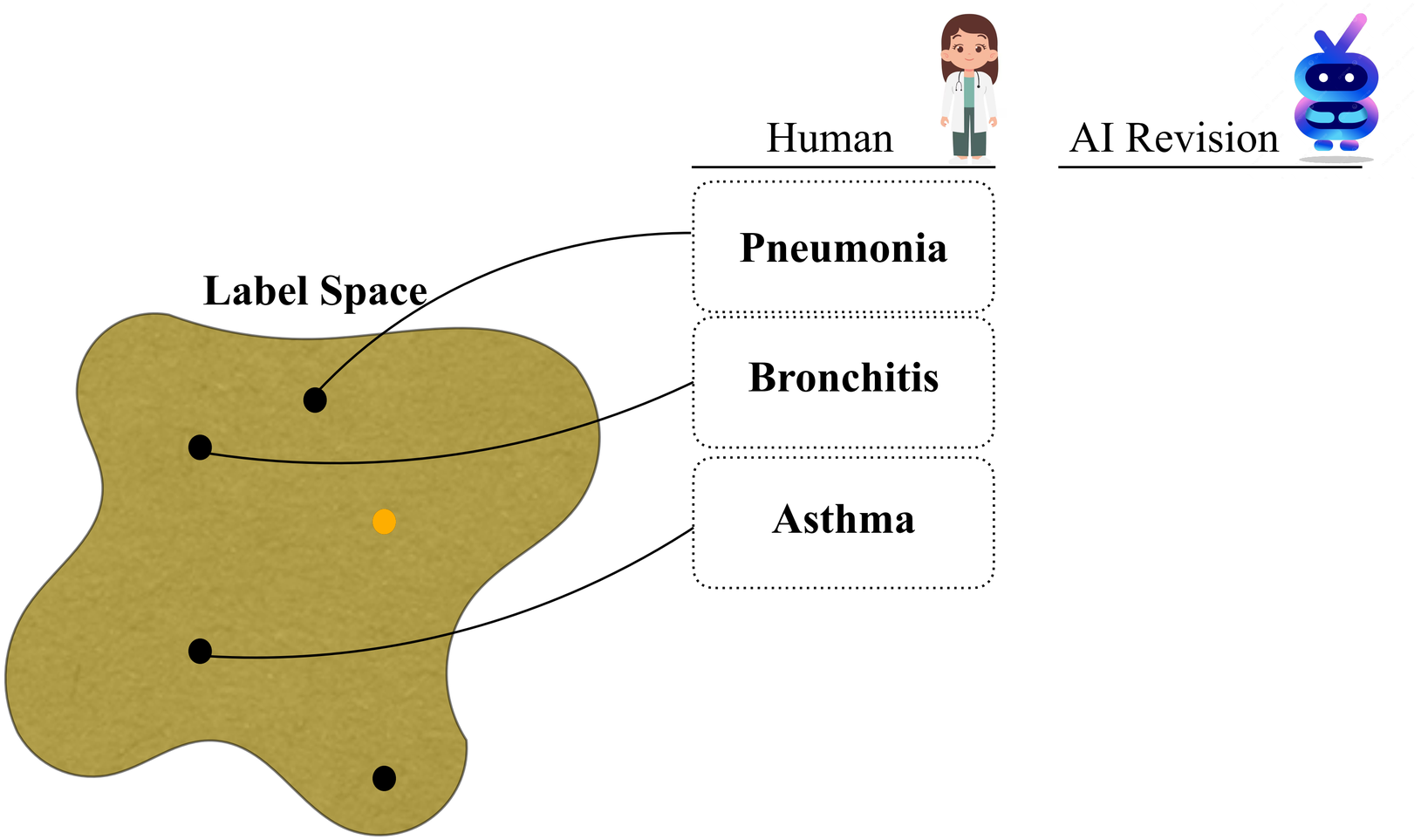
Question: what constitutes a good collaboration?



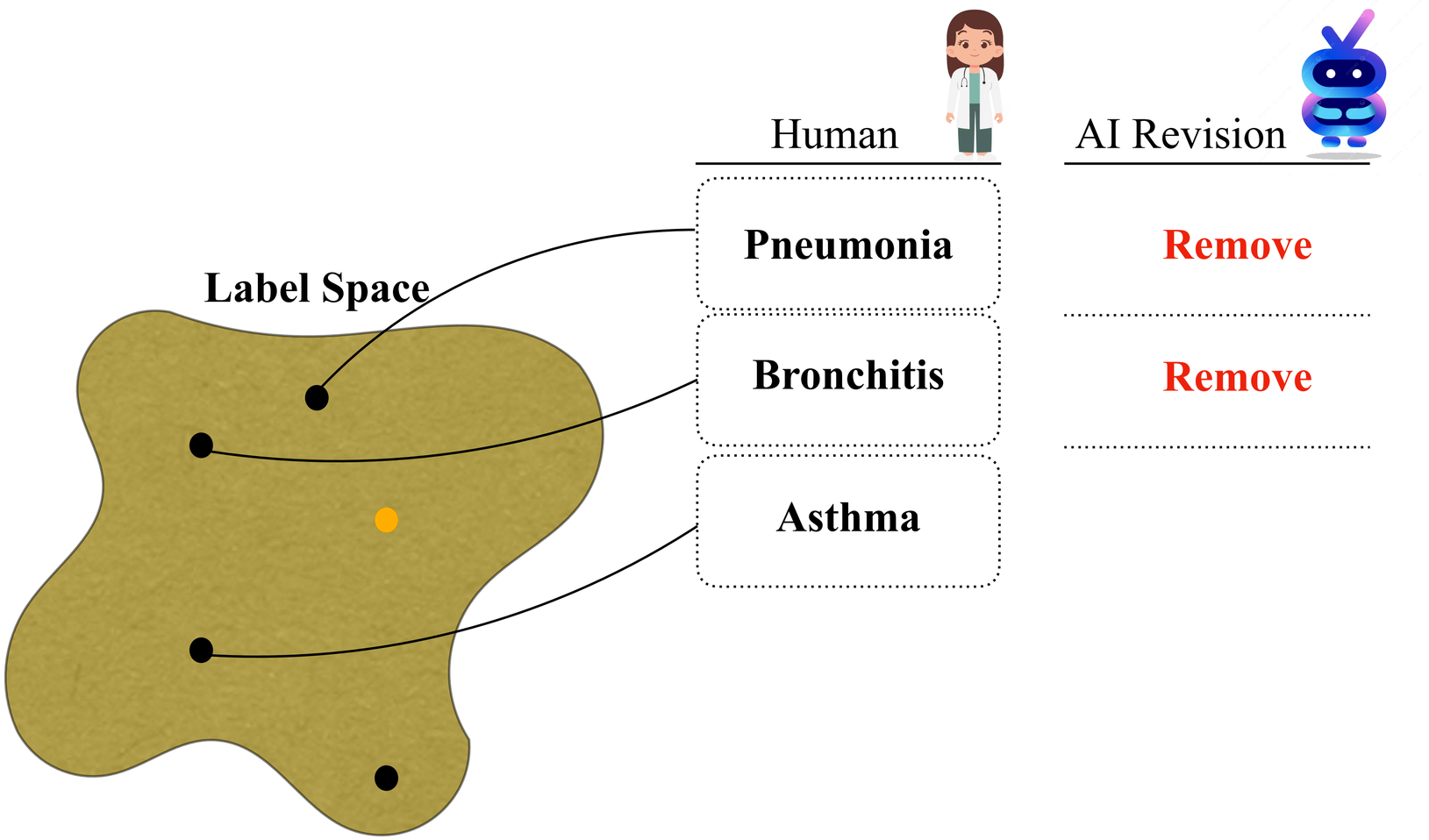
Question: what constitutes a good collaboration?



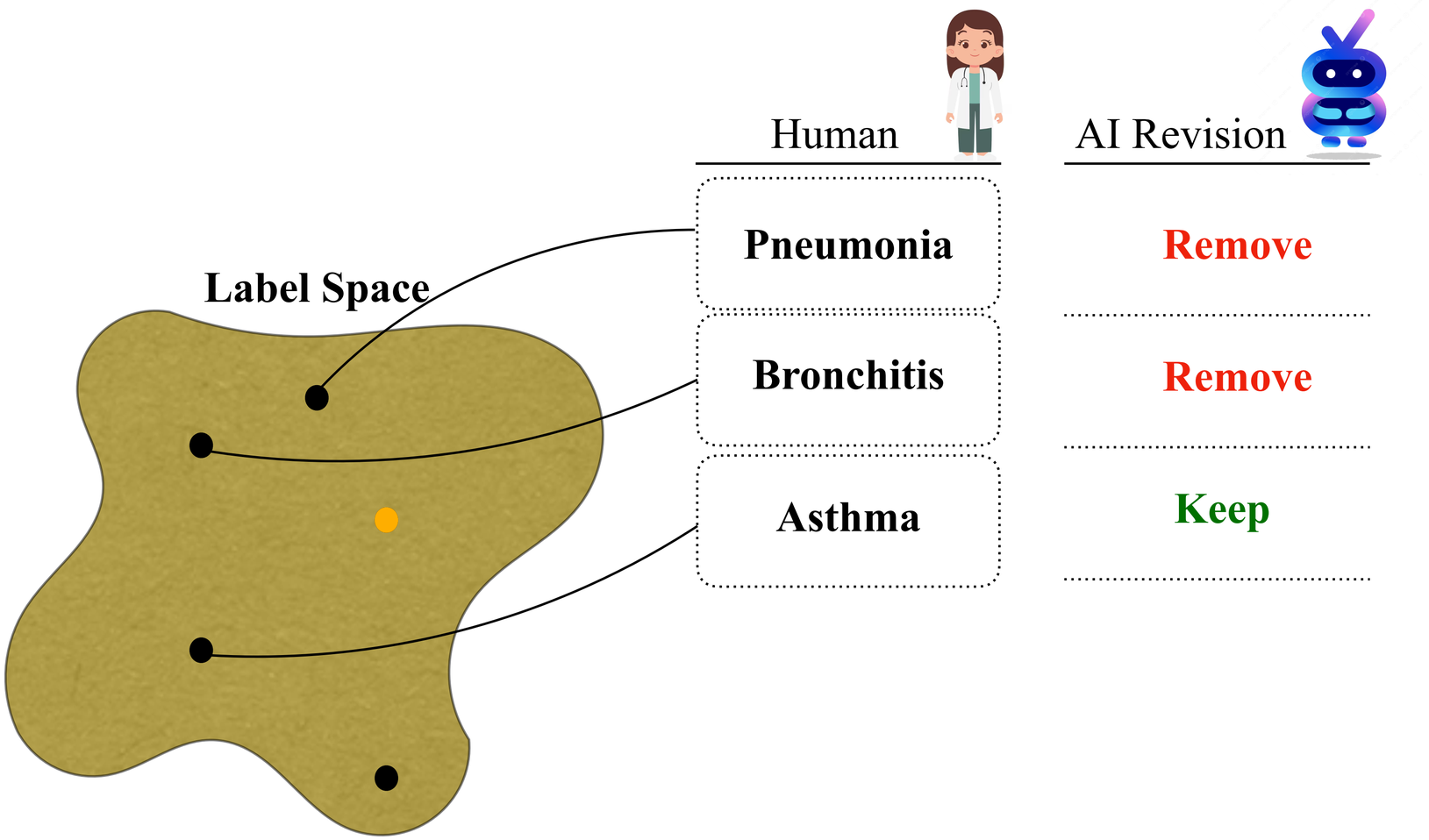
Question: what constitutes a good collaboration?



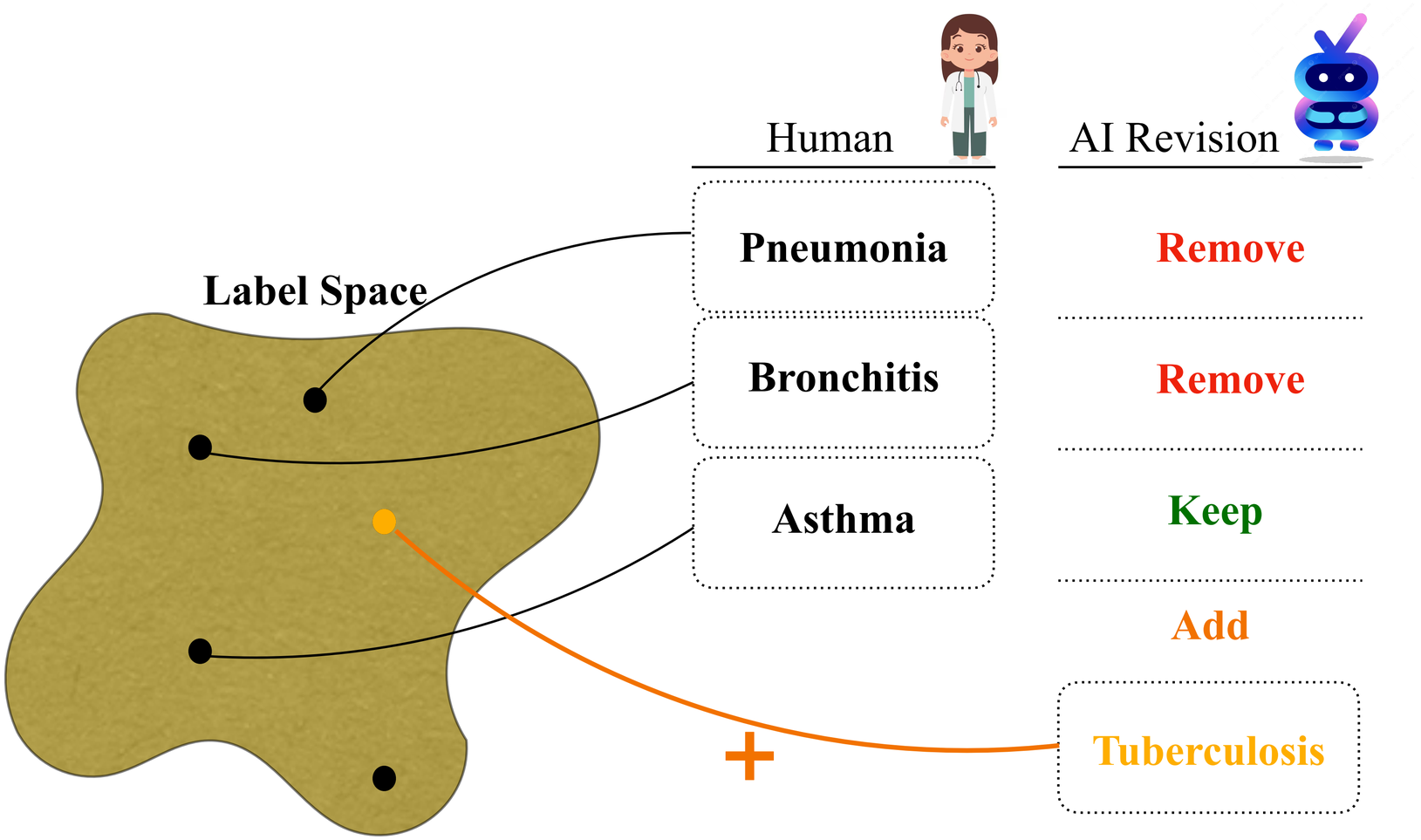
Question: what constitutes a good collaboration?



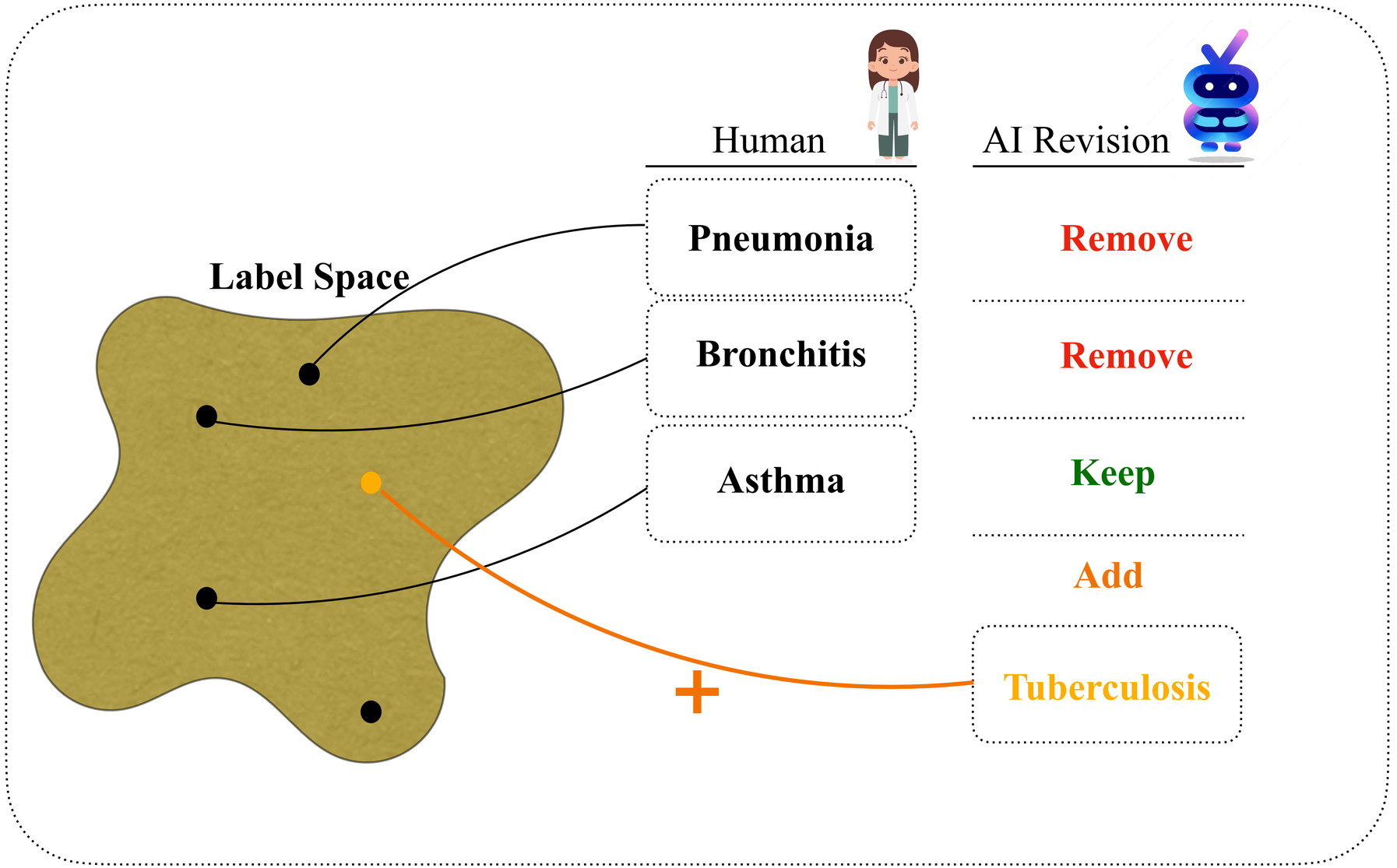
Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?

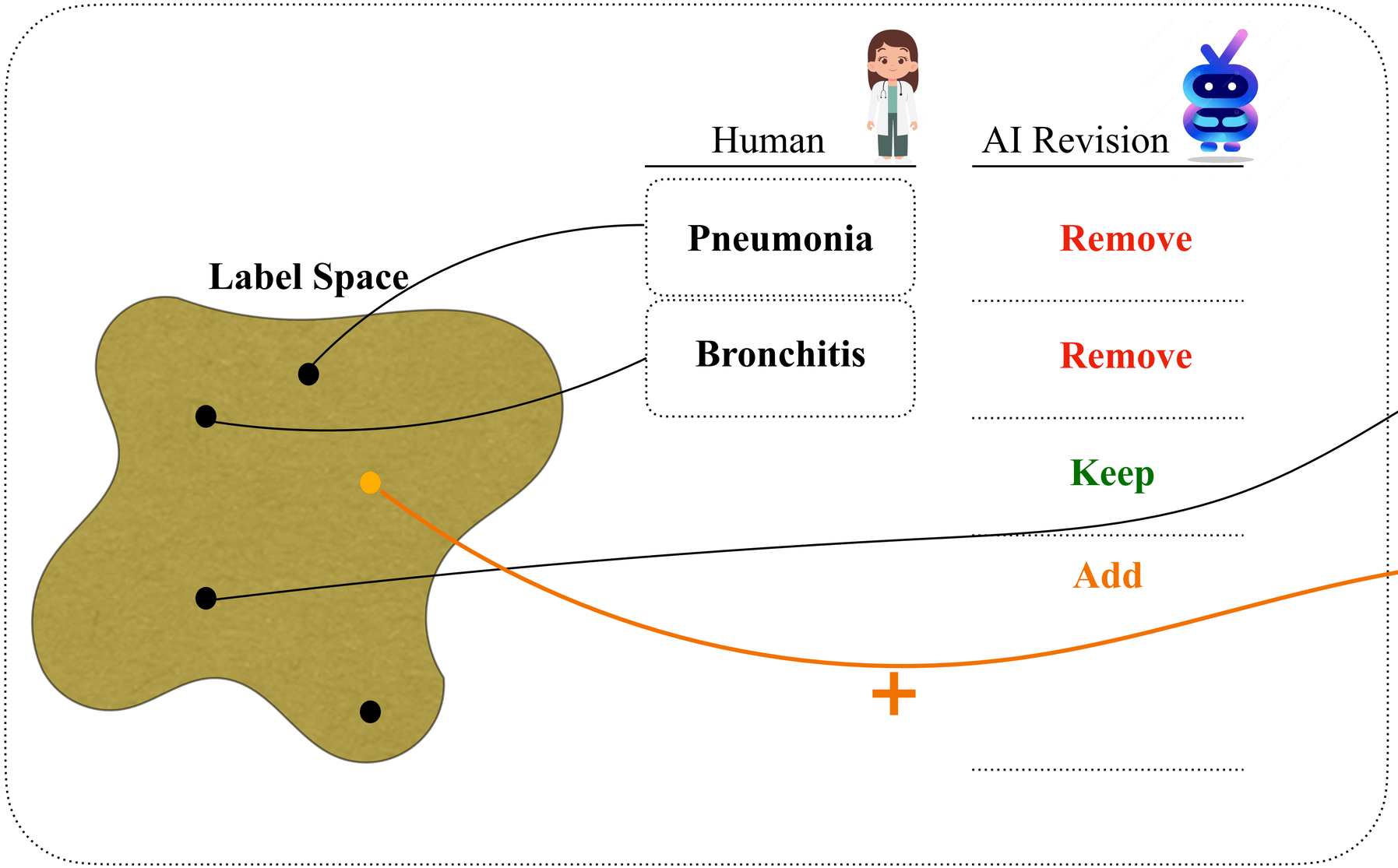


Question: what constitutes a good collaboration?

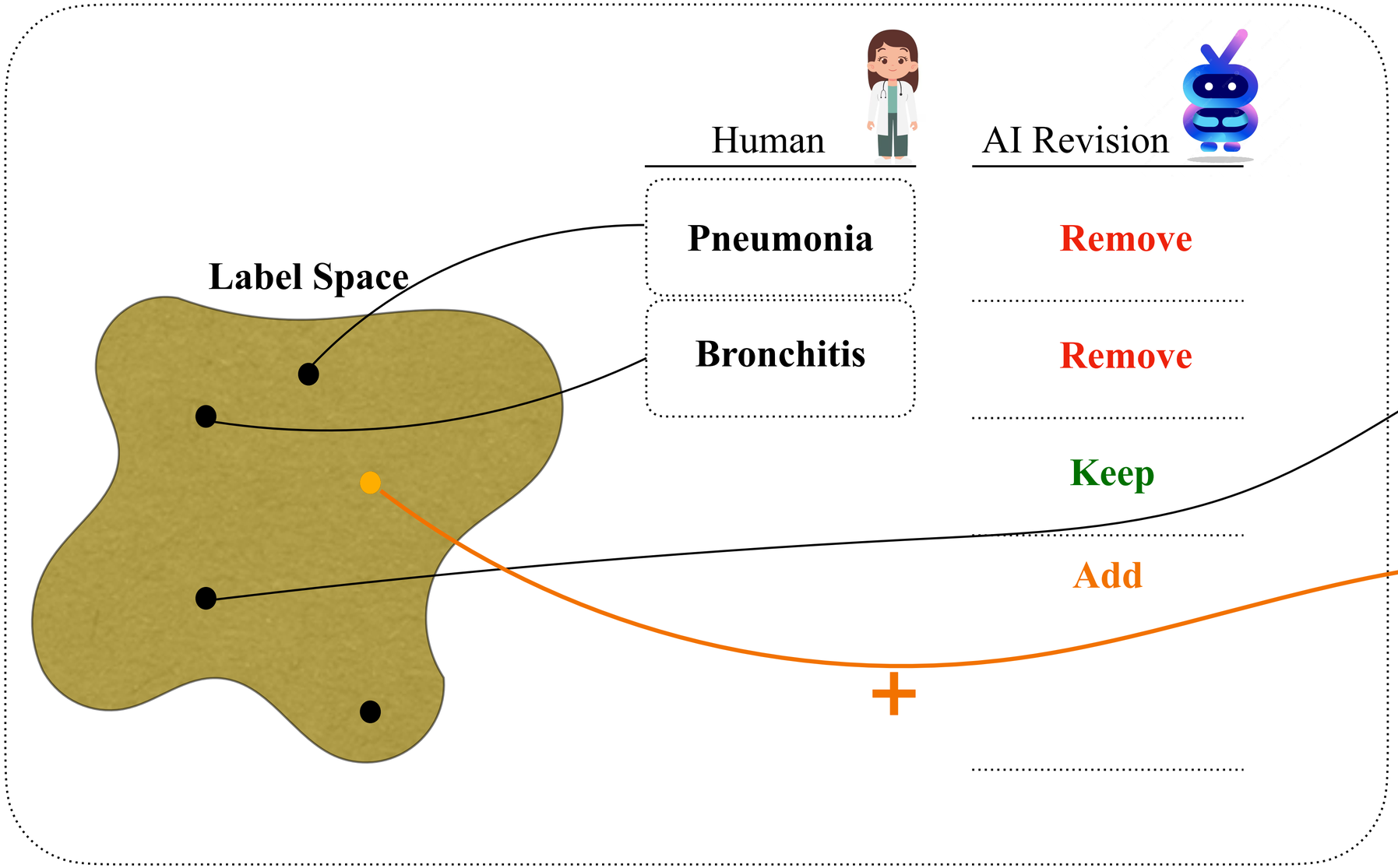


Collaborative Proposal

Question: what constitutes a good collaboration?



Question: what constitutes a good collaboration?



Collaborative Proposal

Asthma

Tuberculosis

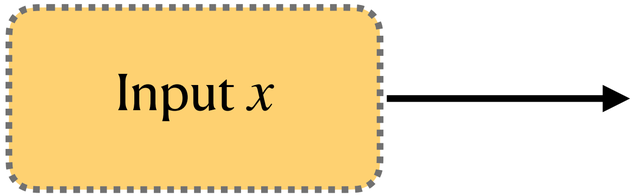
Second: Add value!

Two fundamentals of collaboration

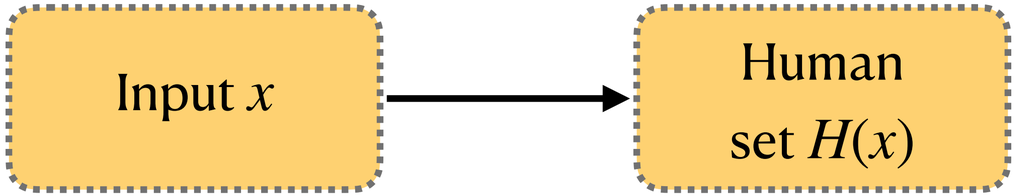
Two fundamentals of collaboration

Input x

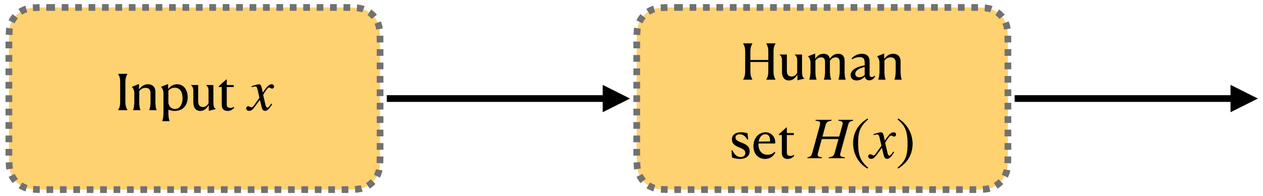
Two fundamentals of collaboration



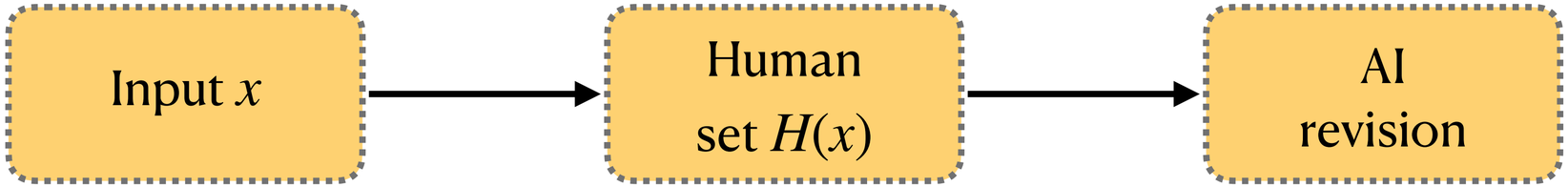
Two fundamentals of collaboration



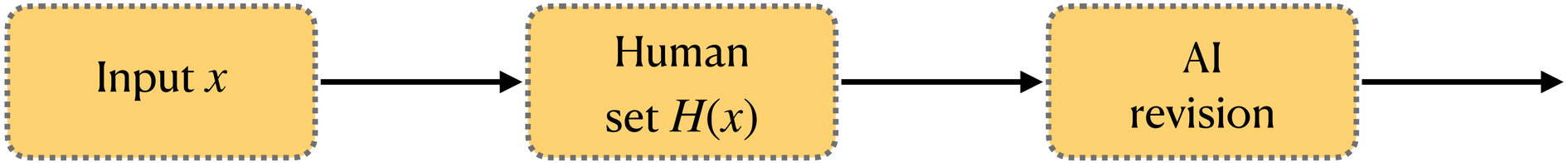
Two fundamentals of collaboration



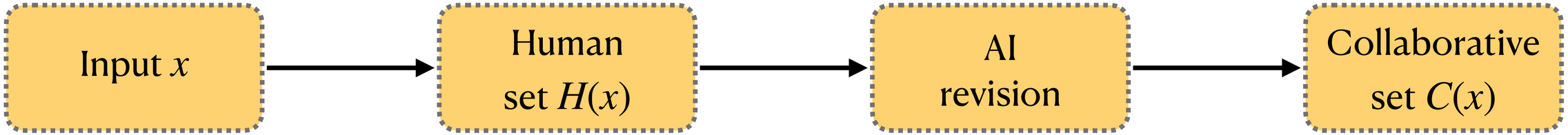
Two fundamentals of collaboration



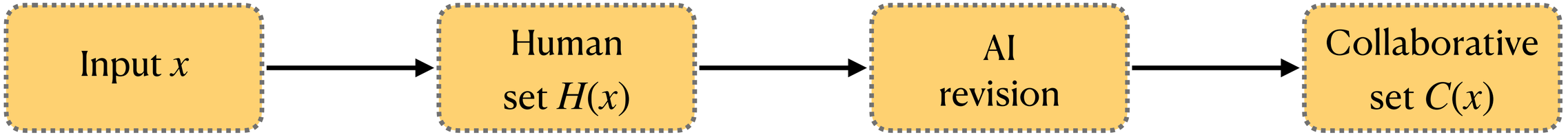
Two fundamentals of collaboration



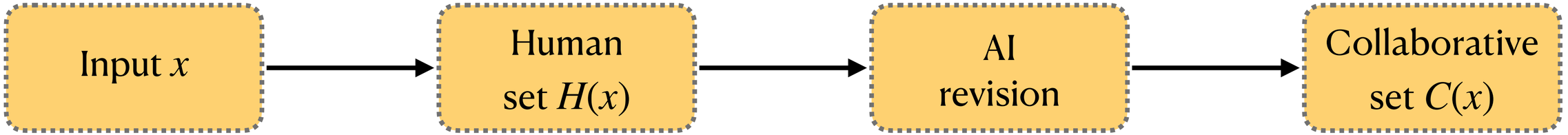
Two fundamentals of collaboration

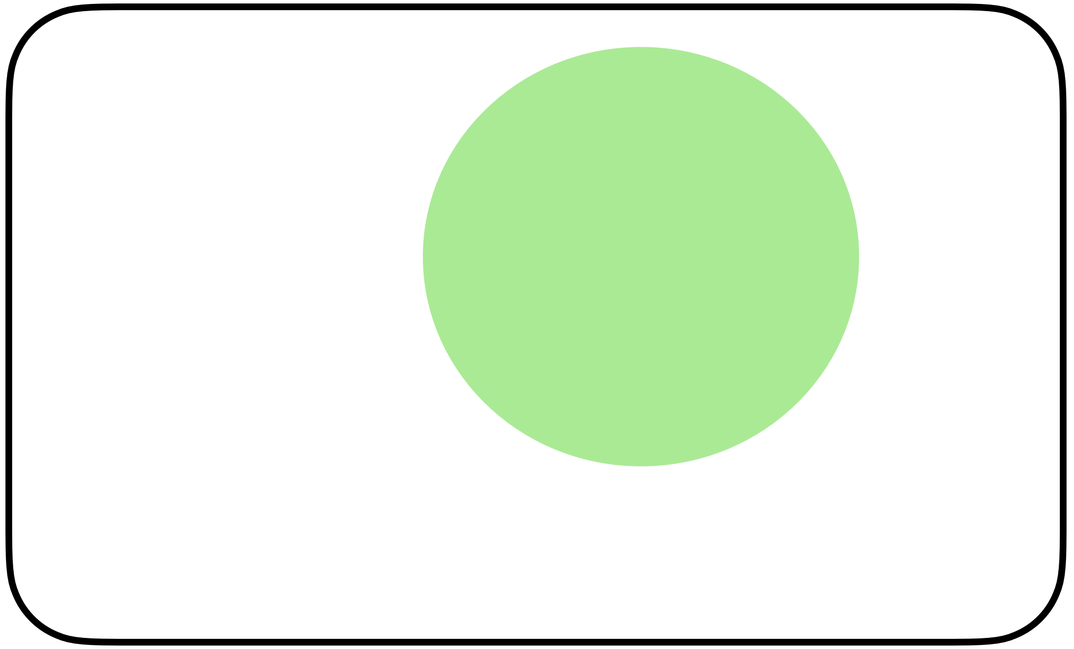


Two fundamentals of collaboration

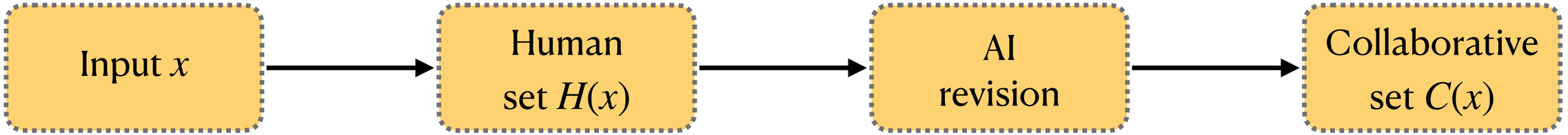


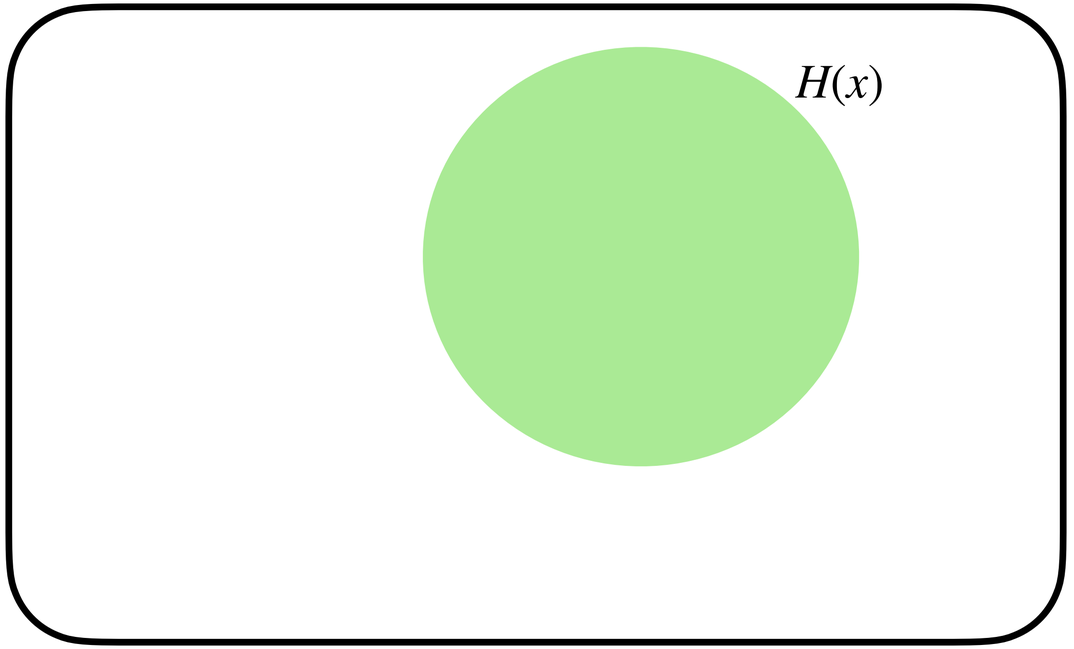
Two fundamentals of collaboration



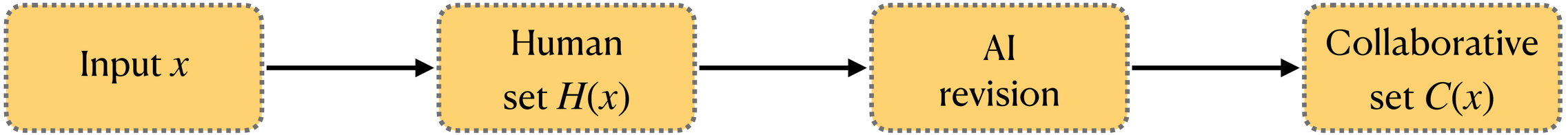


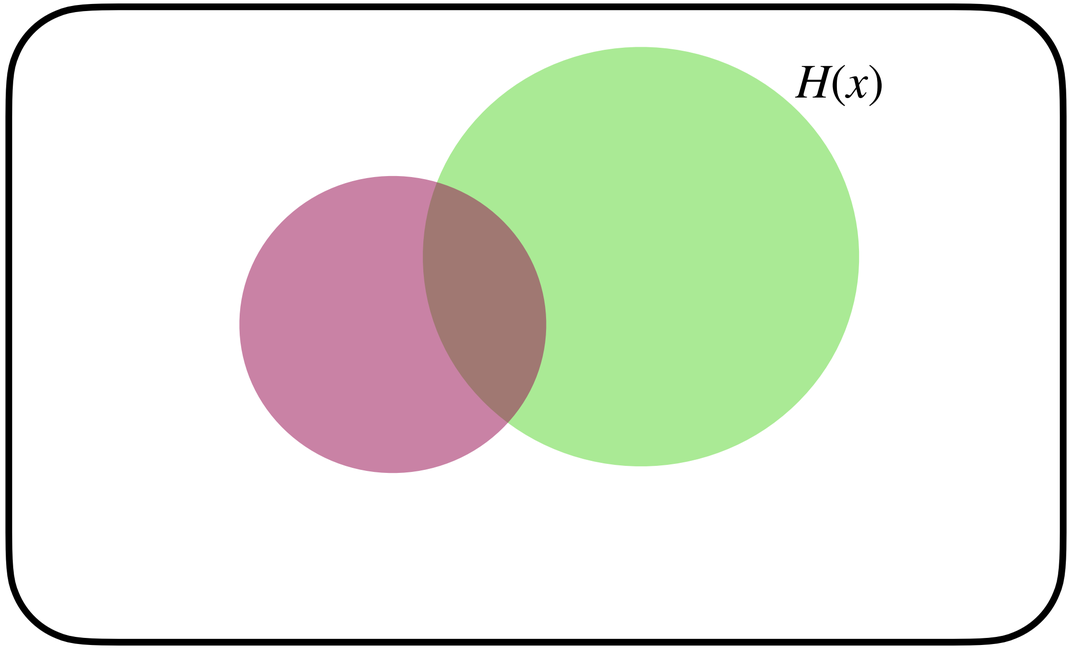
Two fundamentals of collaboration



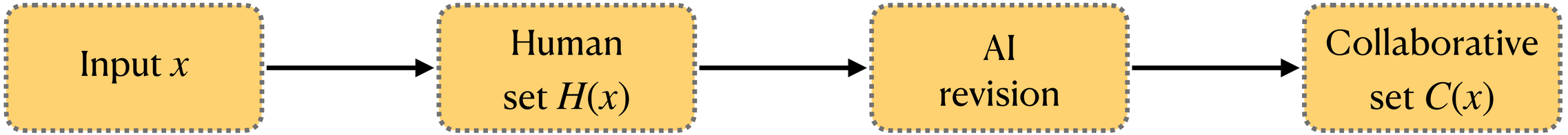


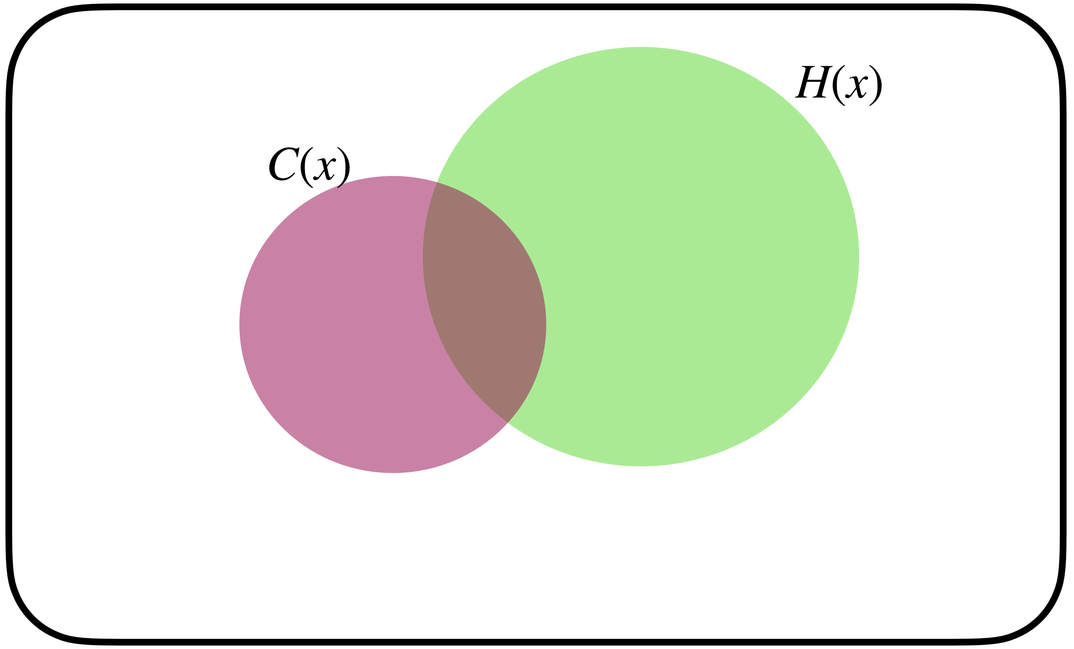
Two fundamentals of collaboration



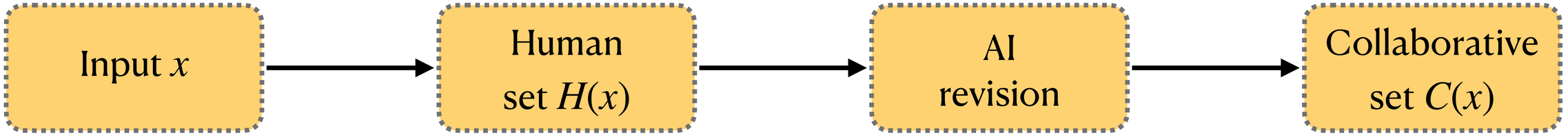


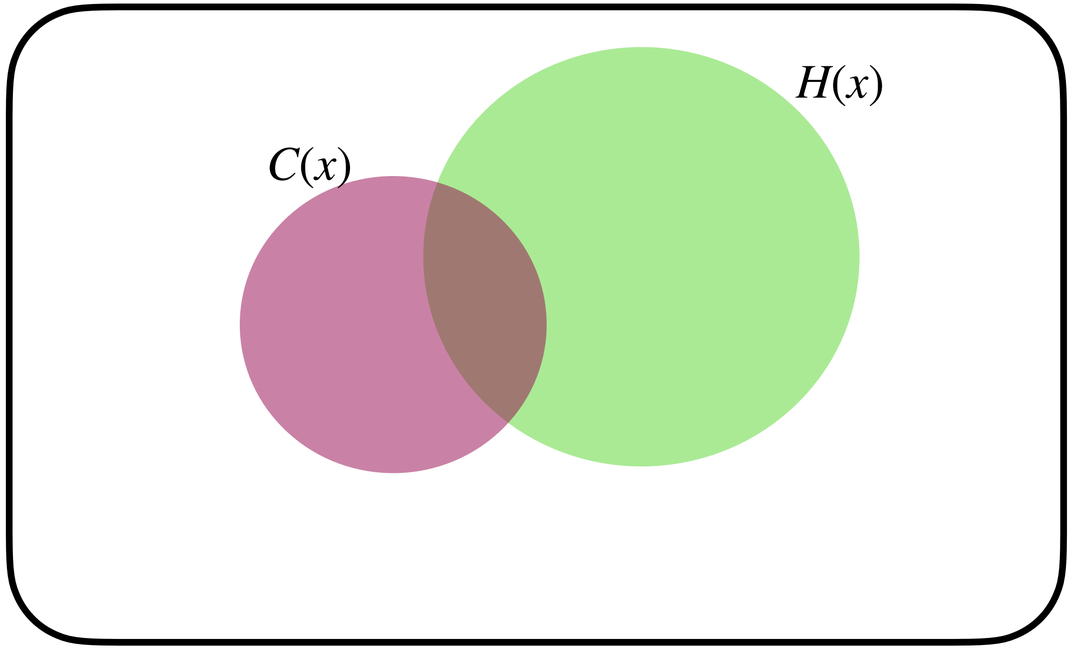
Two fundamentals of collaboration





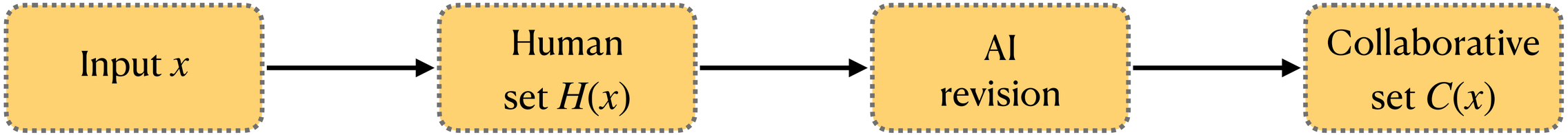
Two fundamentals of collaboration

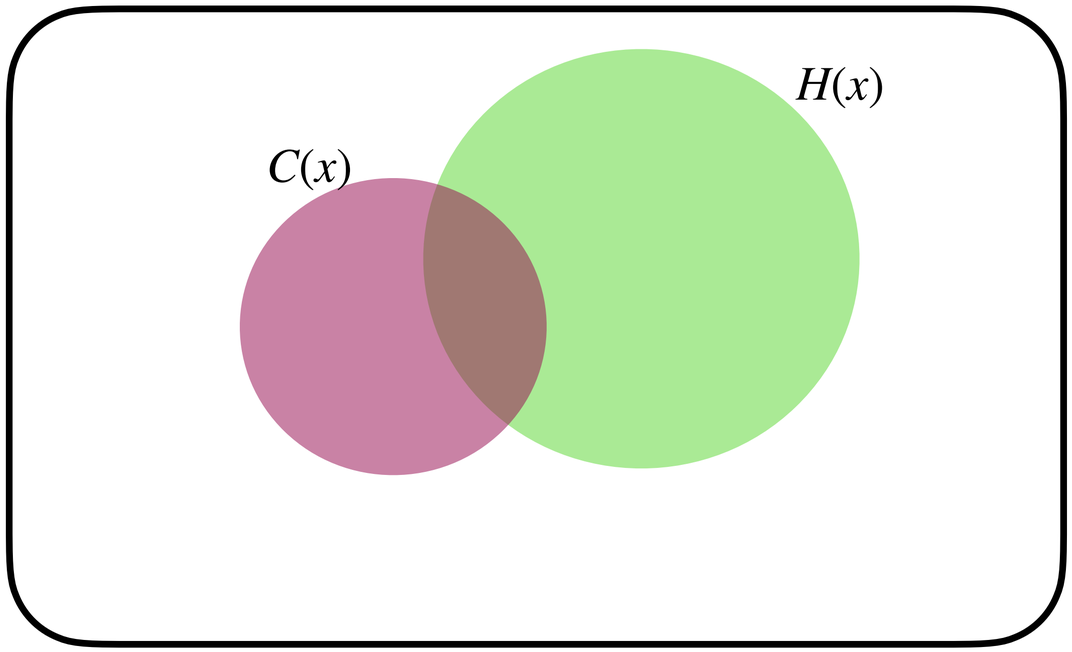




Counterfactual Harm

Two fundamentals of collaboration

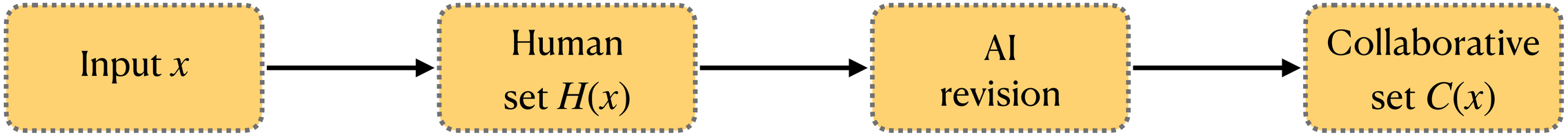


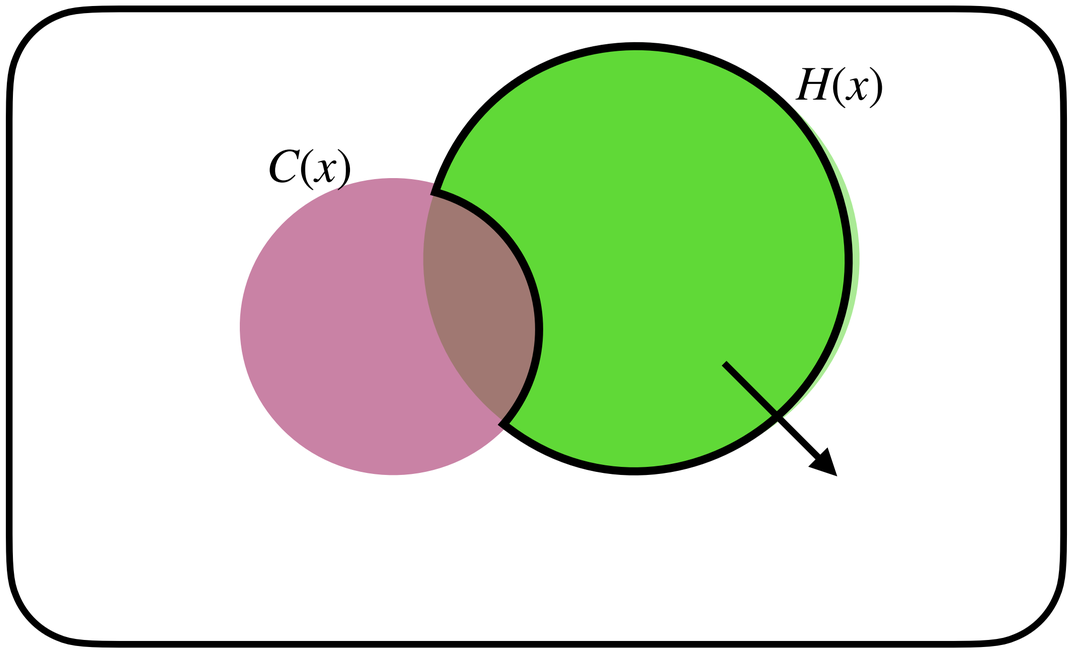


Counterfactual Harm

$$\mathbb{P}(Y \notin C(X) \mid Y \in H(X)) < \varepsilon$$

Two fundamentals of collaboration

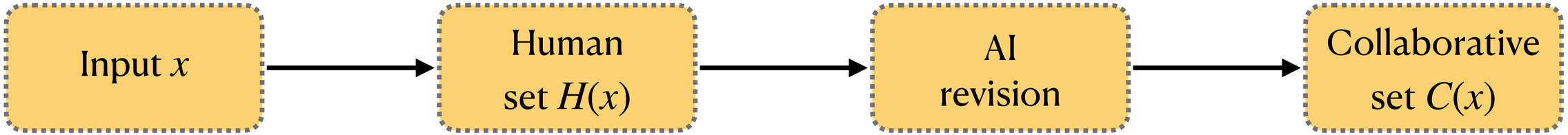


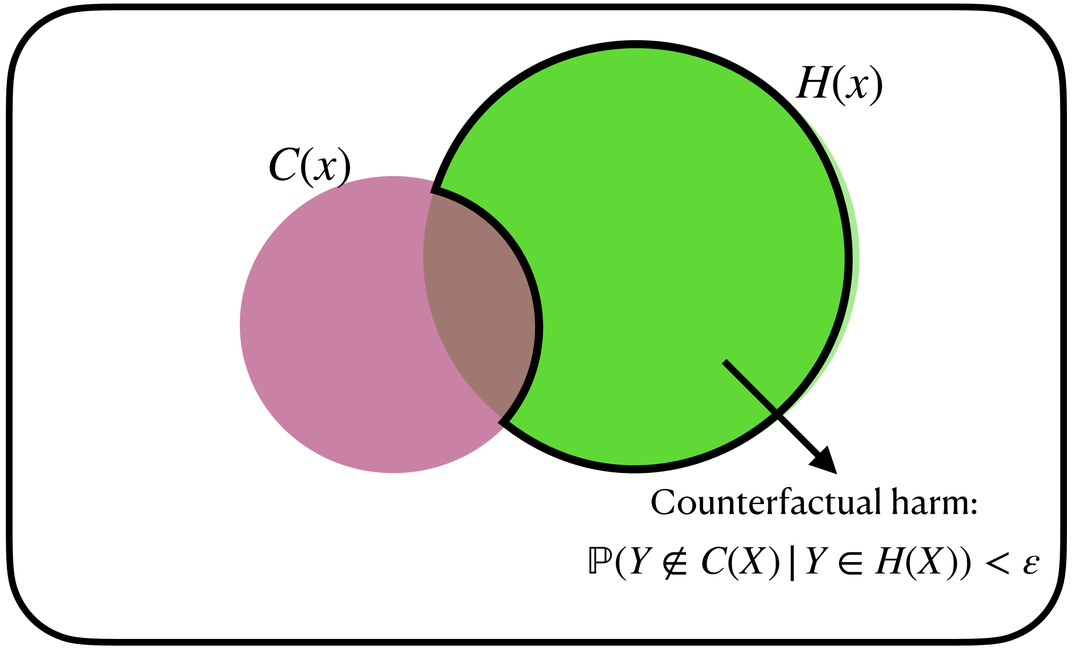


Counterfactual Harm

$$\mathbb{P}(Y \notin C(X) \mid Y \in H(X)) < \varepsilon$$

Two fundamentals of collaboration

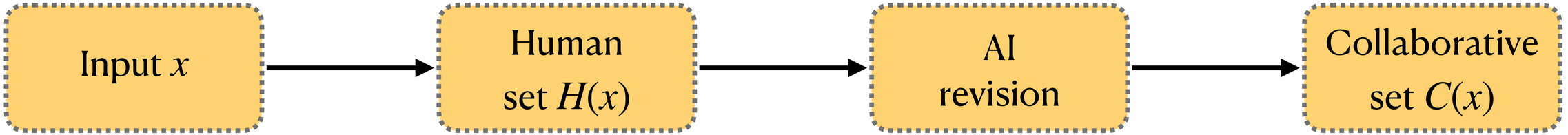


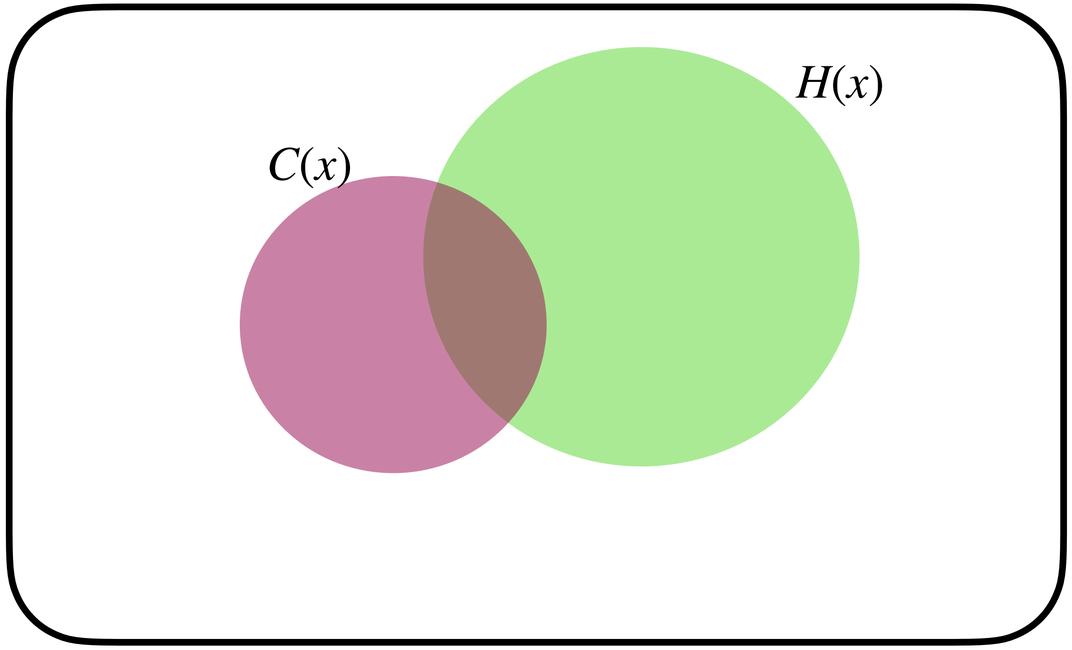


Counterfactual Harm

$$\mathbb{P}(Y \notin C(X) \mid Y \in H(X)) < \varepsilon$$

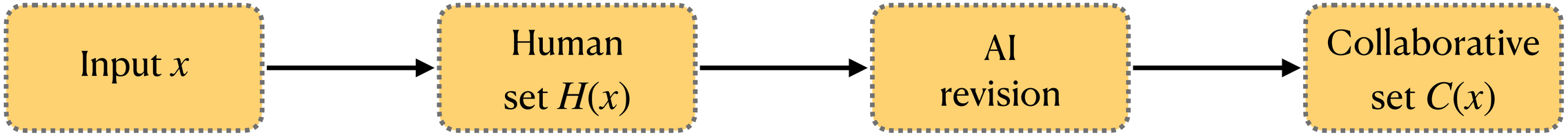
Two fundamentals of collaboration

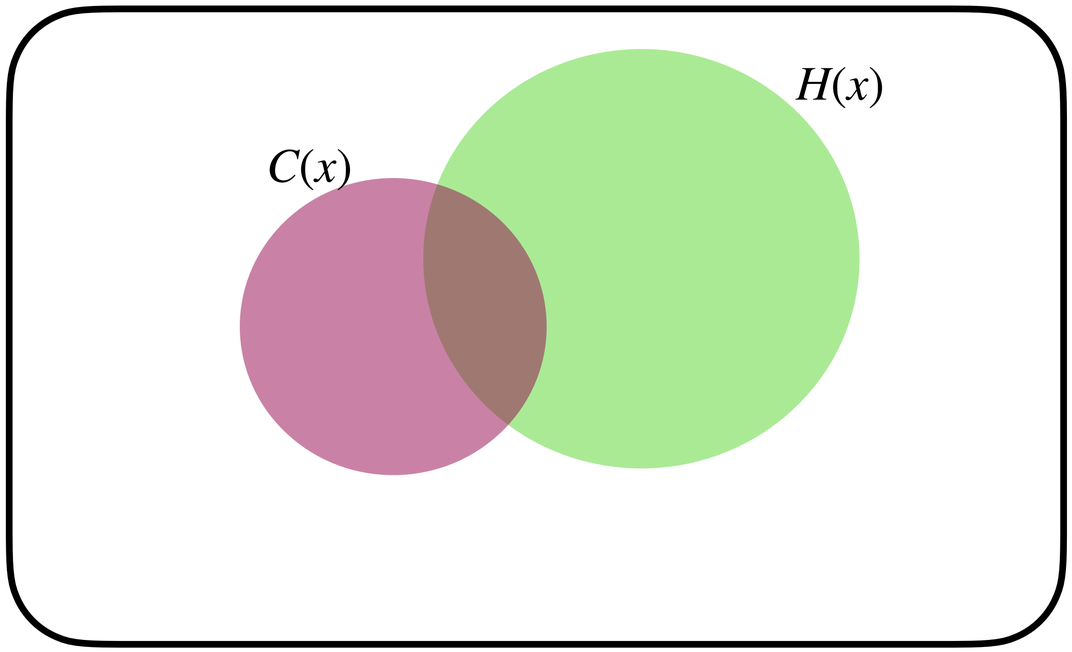




Complementarity

Two fundamentals of collaboration

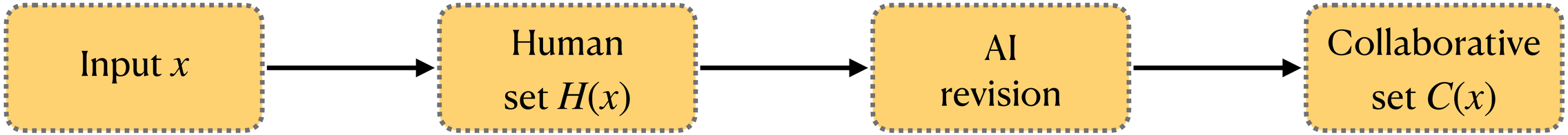


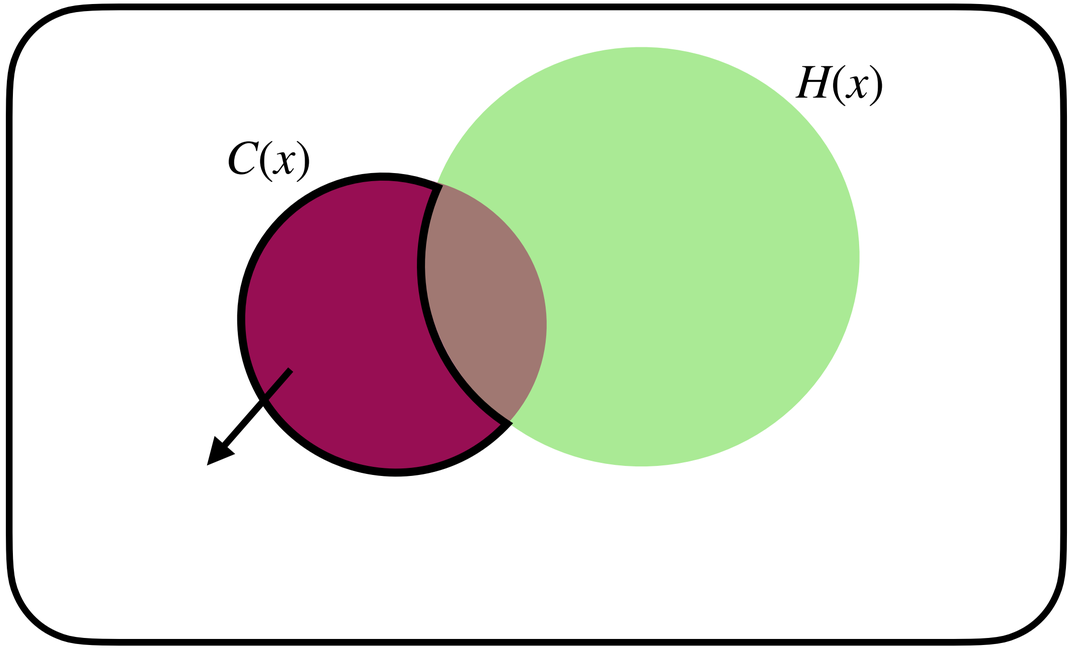


Complementarity

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \geq 1 - \delta$$

Two fundamentals of collaboration

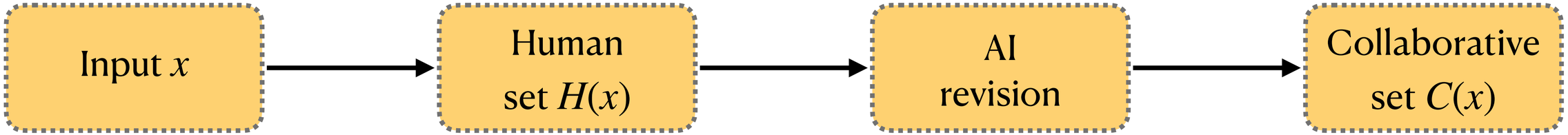


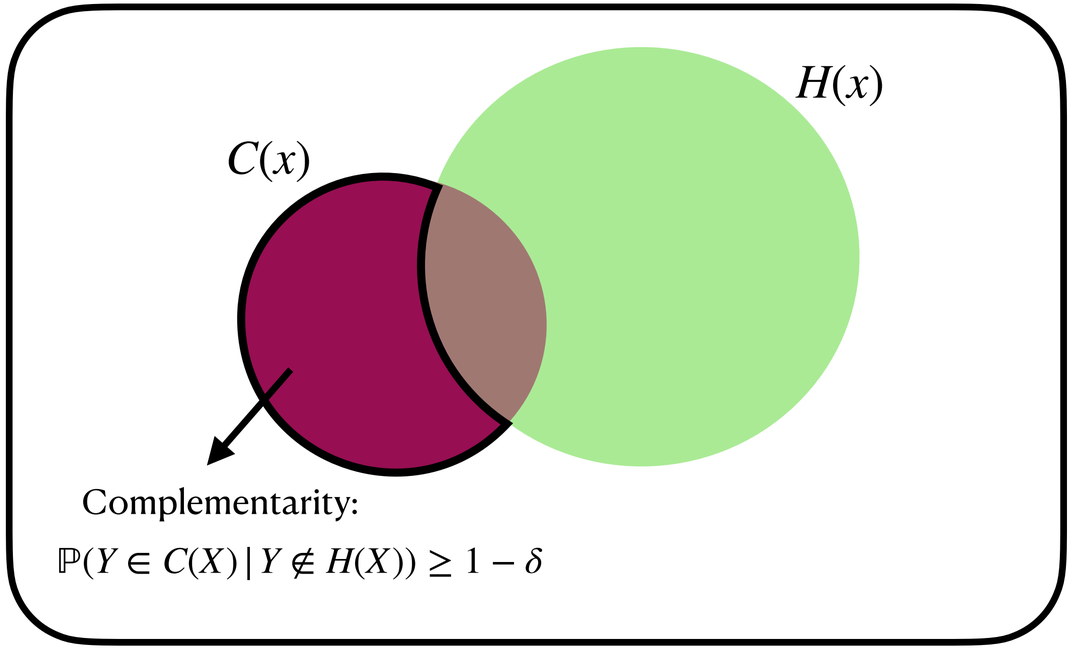


Complementarity

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \geq 1 - \delta$$

Two fundamentals of collaboration

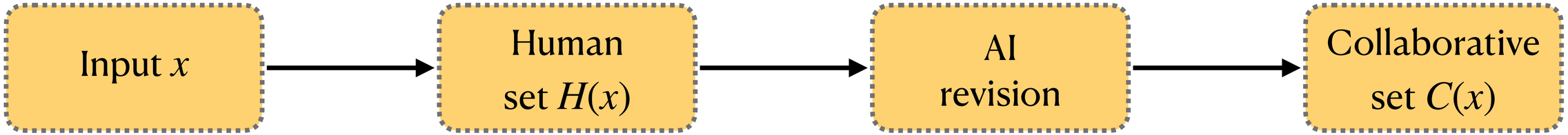


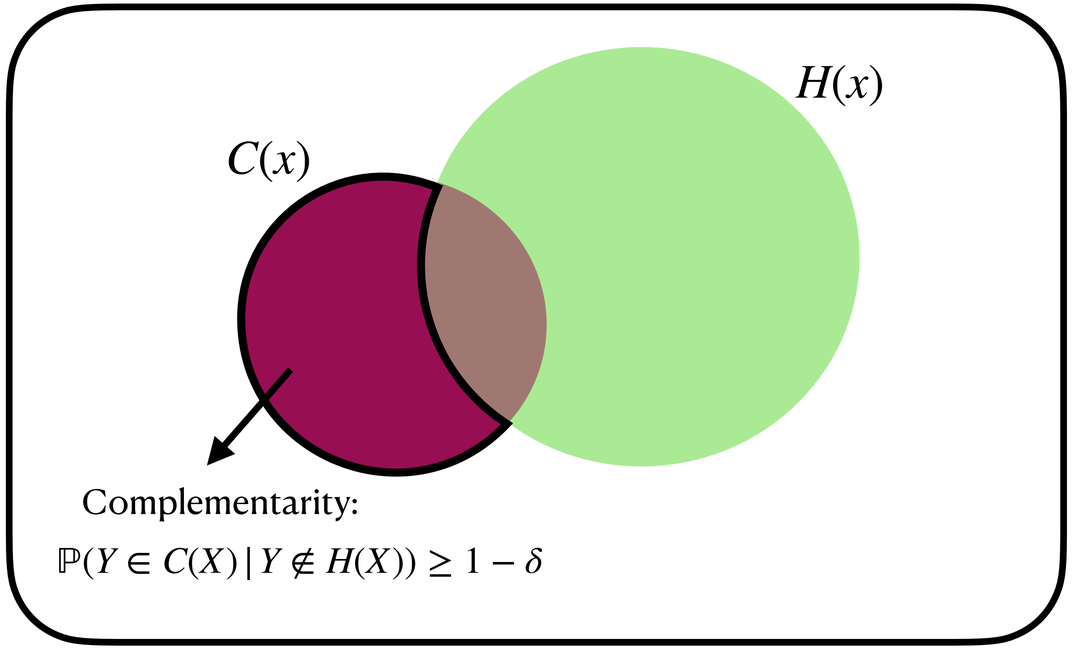


Complementarity

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \geq 1 - \delta$$

Two fundamentals of collaboration

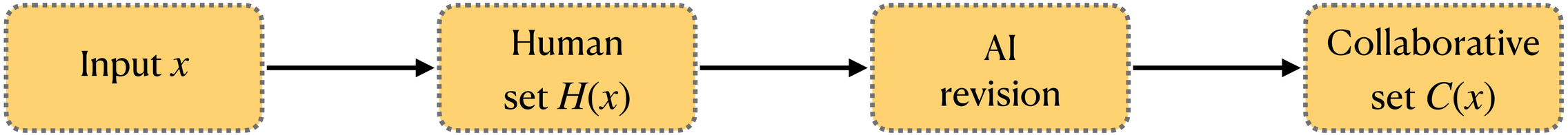


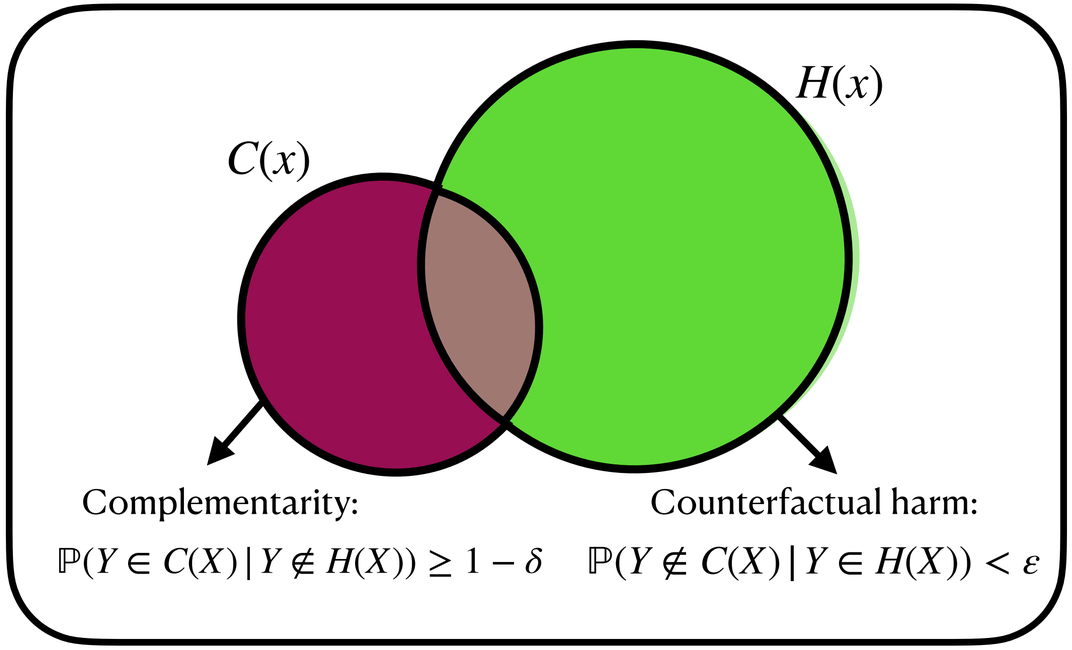


Complementarity

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \geq 1 - \delta$$

Two fundamentals of collaboration





Counterfactual Harm

$$\mathbb{P}(Y \notin C(X) \mid Y \in H(X)) < \varepsilon$$

Complementarity

$$\mathbb{P}(Y \in C(X) \mid Y \notin H(X)) \geq 1 - \delta$$

Question: what constitutes a good collaboration?

Question: **what constitutes a good collaboration?**

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \;\geq\; 1-\delta\,.$$

Question: what constitutes a good collaboration?

$$\min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|C(X)|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) < \varepsilon, \qquad \text{Counterfactual Harm}$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \geq 1-\delta\,.$$

Question: what constitutes a good collaboration?

$$\min_{C:\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|C(X)|$$

$$\text{s.t.} \quad \mathbb{P}\big(Y \notin C(X) \,\big|\, Y \in H(X)\big) < \varepsilon, \qquad \text{Counterfactual Harm}$$

$$\mathbb{P}\big(Y \in C(X) \,\big|\, Y \notin H(X)\big) \geq 1-\delta. \qquad \text{Complementarity}$$

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \;\geq\; 1-\delta\,.$$

HACO

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \;\geq\; 1-\delta\,.$$

HACO

Theorem: The optimal solution to HACO is of the form

Human-AI Collaborative Optimization

$$\min_{C:\,\mathcal{X}\to 2^{\mathcal{Y}}} \quad \mathbb{E}\,|\,C(X)\,|$$

$$\text{s.t.} \quad \mathbb{P}\Big(Y \notin C(X) \,\Big|\, Y \in H(X)\Big) \;<\; \varepsilon,$$

$$\mathbb{P}\Big(Y \in C(X) \,\Big|\, Y \notin H(X)\Big) \;\geq\; 1-\delta\,.$$

HACO

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

VS

Ordinary CP:

$$C(x) = \{ y \mid 1 - p(y \mid x) \leq q^* \}$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

$s(x, y)$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x, y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x, y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

1 **Question:** How to design the score function?

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; s(x,y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

1. **Question:** How to design the score function?

2. **Question:** How to debias the thresholds?

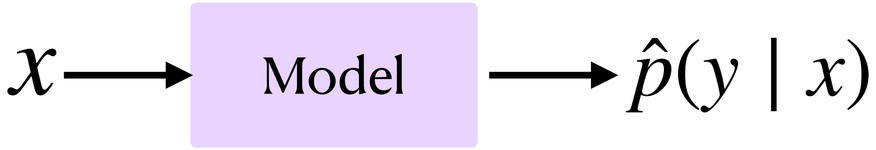
Question: How to design the score function?

Question: How to design the score function?

$$C^*(x) = \left\{ y \;\middle|\; \boxed{1 - p(y \mid x)} \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

$s(x, y)$

Classification



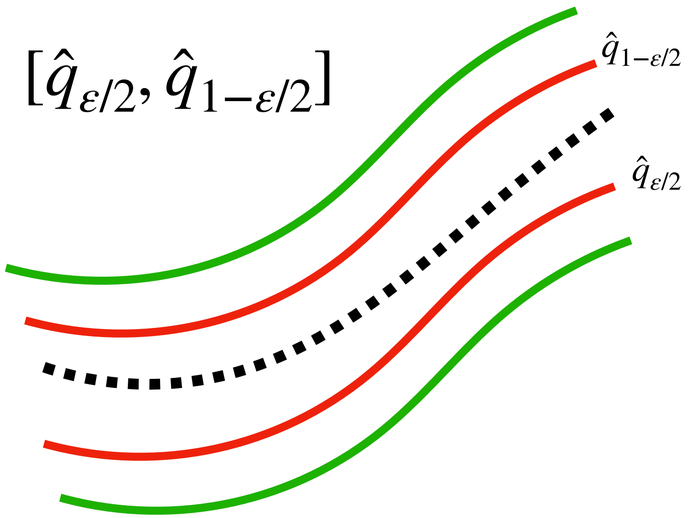
$$\hat{s}(x, y) = 1 - \hat{p}(y \mid x)$$

Question: How to design the score function?

$$C^*(x) = \left\{ y \;\middle|\; \boxed{1 - p(y \mid x)} \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

$s(x, y)$

Conformalized Quantile Regression



$$\hat{s}(x, y) = \max\left\{\hat{q}_{\varepsilon/2}(x) - y,\, y - \hat{q}_{1-\varepsilon/2}(x)\right\}$$

Our two threshold structure

$$\hat{s}(x, y) = \begin{cases} \max\left\{\hat{q}_{\varepsilon/2}(x) - y,\, y - \hat{q}_{1-\varepsilon/2}(x)\right\}, & y \in H(x), \\ \max\left\{\hat{q}_{\delta/2}(x) - y,\, y - \hat{q}_{1-\delta/2}(x)\right\}, & y \notin H(x)\,. \end{cases}$$

Human-AI Collaborative Optimization

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; \hat{s}(x,y) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

1. **Question:** How to design the score function?

2. **Question:** How to debias the thresholds?

Question: How to debias the thresholds?

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

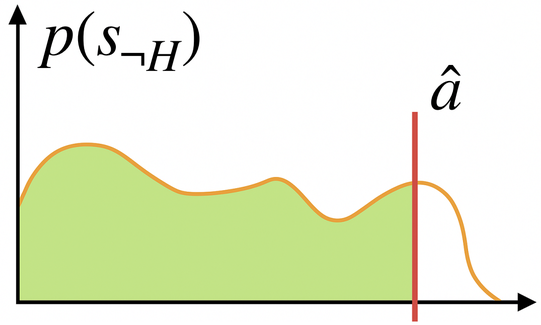
Assume Exchangeability **What should a and b be?**

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability **What should a and b be?**

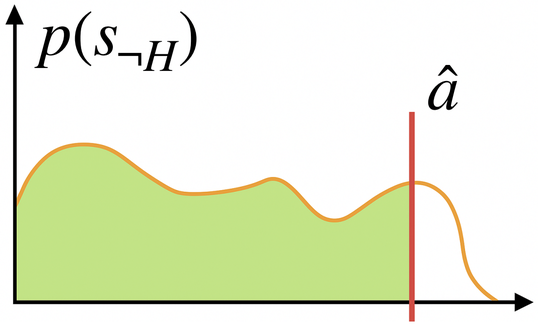


Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability **What should a and b be?**



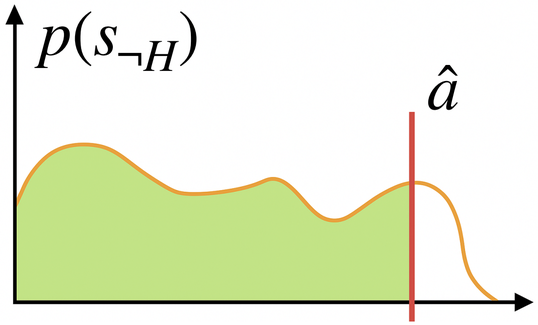
$$\hat{a} = Q_{1-\delta}\big(\{ s_i : \; Y_i \notin H(X_i) \} \cup \{\infty\}\big)$$

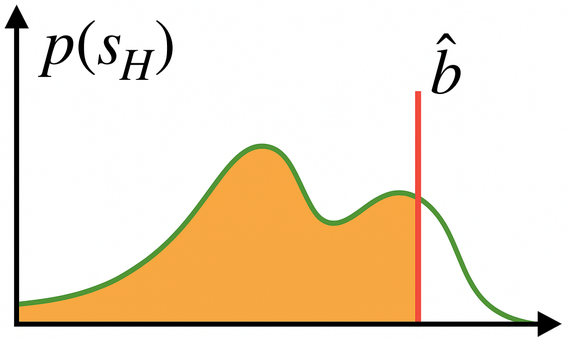
Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability **What should a and b be?**





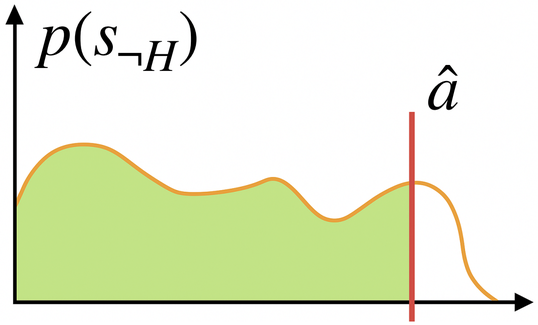
$$\hat{a} = Q_{1-\delta}\big(\{\, s_i : \;\; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)$$

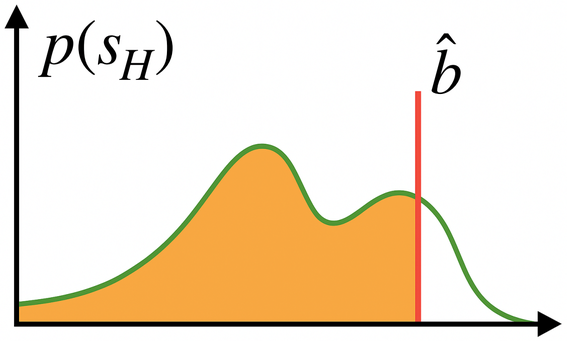
Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability **What should a and b be?**





$$\hat{a} = Q_{1-\delta}\big(\{\, s_i : \;\; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)$$

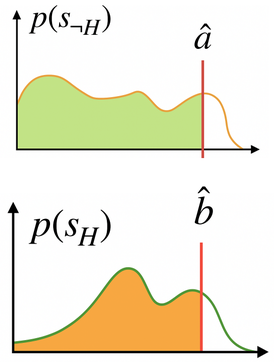
$$\hat{b} = Q_{1-\varepsilon}\big(\{\, s_i : \;\; Y_i \in H(X_i) \,\} \cup \{\infty\}\big)$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability

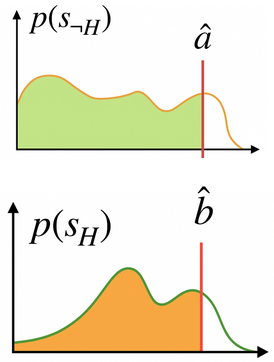


Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \leq a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



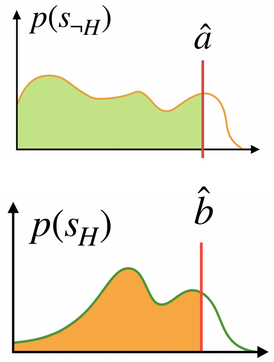
For a new test example (x_{test}, y_{test}) :

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

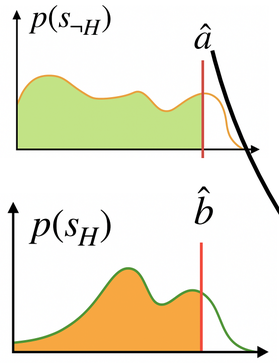
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

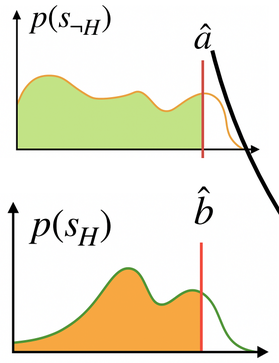
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

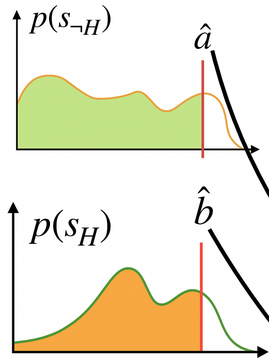
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

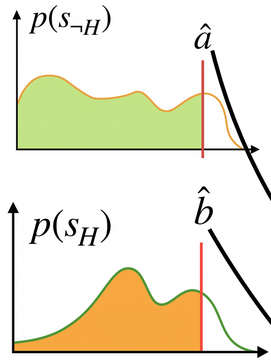
$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability



For a new test example (x_{test}, y_{test}) :

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b} \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall\, x \in \mathcal{X}\,.$$

Assume Exchangeability $\hat{a} = Q_{1-\delta}\big(\{\, s_i : \; Y_i \notin H(X_i) \,\} \cup \{\infty\}\big)\,; \quad \hat{b} = Q_{1-\varepsilon}\big(\{\, s_i : \; Y_i \in H(X_i) \,\} \cup \{\infty\}\big)$

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a}\, \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b}\, \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

$$1 - \varepsilon \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \in H(X_{test})\right] \;<\; 1 - \varepsilon + \frac{1}{n_1 + 1}$$

$$1 - \delta \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \notin H(X_{test})\right] \;<\; 1 - \delta + \frac{1}{n_2 + 1}$$

Question: How to debias the thresholds?

Theorem: The optimal solution to HACO is of the form

$$C^*(x) = \left\{ y \;\middle|\; 1 - p(y \mid x) \le a^* \mathbf{1}_{\{y \notin H(x)\}} + b^* \mathbf{1}_{\{y \in H(x)\}} \right\}, \quad \forall x \in \mathcal{X}.$$

Assume Exchangeability $\hat{a} = Q_{1-\delta}\big(\{ s_i : \; Y_i \notin H(X_i) \} \cup \{\infty\}\big)$; $\hat{b} = Q_{1-\varepsilon}\big(\{ s_i : \; Y_i \in H(X_i) \} \cup \{\infty\}\big)$

$$C(x_{\text{test}}) = \left\{ y \;\middle|\; s(x_{\text{test}}, y) \le \hat{a} \mathbf{1}_{\{y \notin H(x_{\text{test}})\}} + \hat{b} \mathbf{1}_{\{y \in H(x_{\text{test}})\}} \right\}.$$

Offline Guarantees

$$1 - \varepsilon \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \in H(X_{test})\right] \;<\; 1 - \varepsilon + \frac{1}{n_1 + 1}$$

$$1 - \delta \;\le\; \Pr\left[Y_{test} \in C(X_{test}) \mid Y_{test} \notin H(X_{test})\right] \;<\; 1 - \delta + \frac{1}{n_2 + 1}$$

~~Assume Exchangeability~~

~~Assume Exchangeability~~

Exchangeability is fragile

~~Assume Exchangeability~~

Exchangeability is fragile

Question: How to debias the thresholds in the online setting

Question: How to debias the thresholds in the online setting

Question: How to debias the thresholds in the online setting

At round t

Question: How to debias the thresholds in the online setting

At round t

input x_t

Question: How to debias the thresholds in the online setting

At round t

input x_t

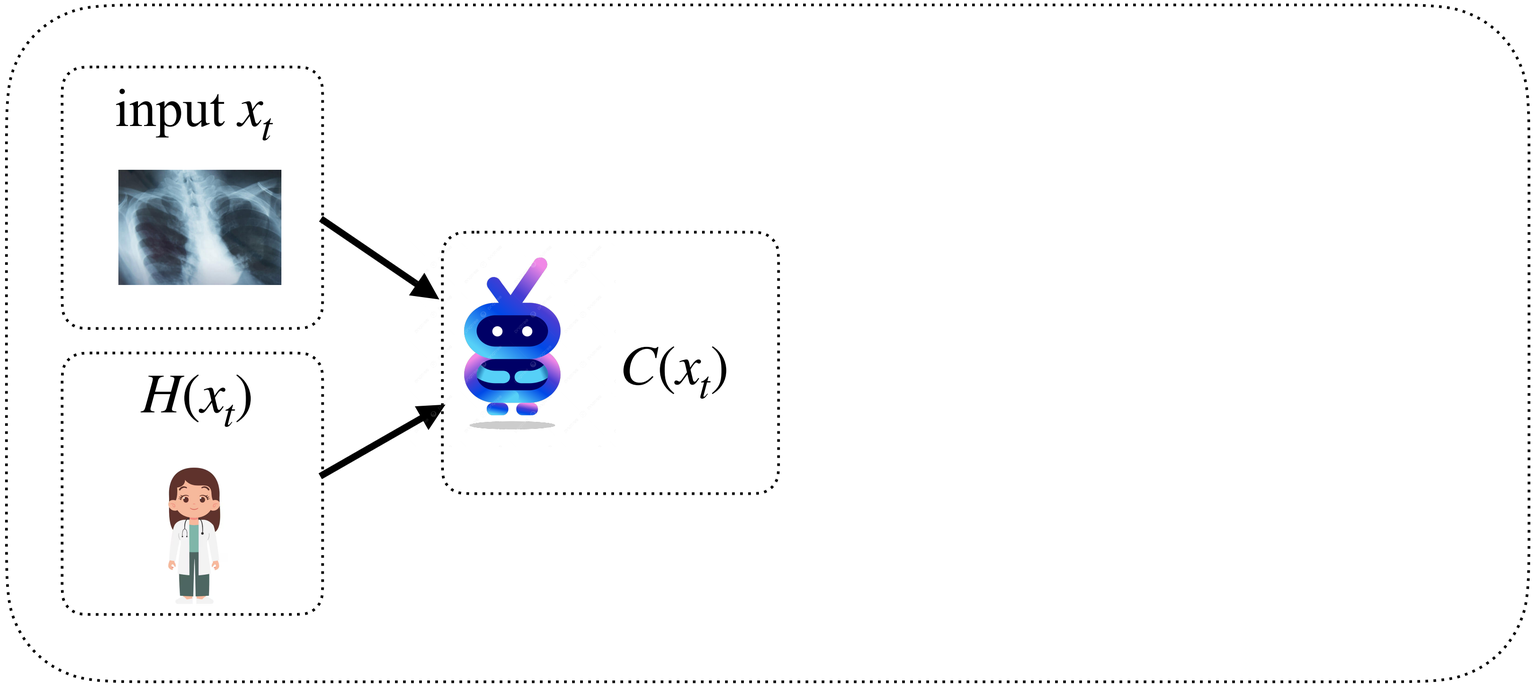
$H(x_t)$

Question: How to debias the thresholds in the online setting

At round t

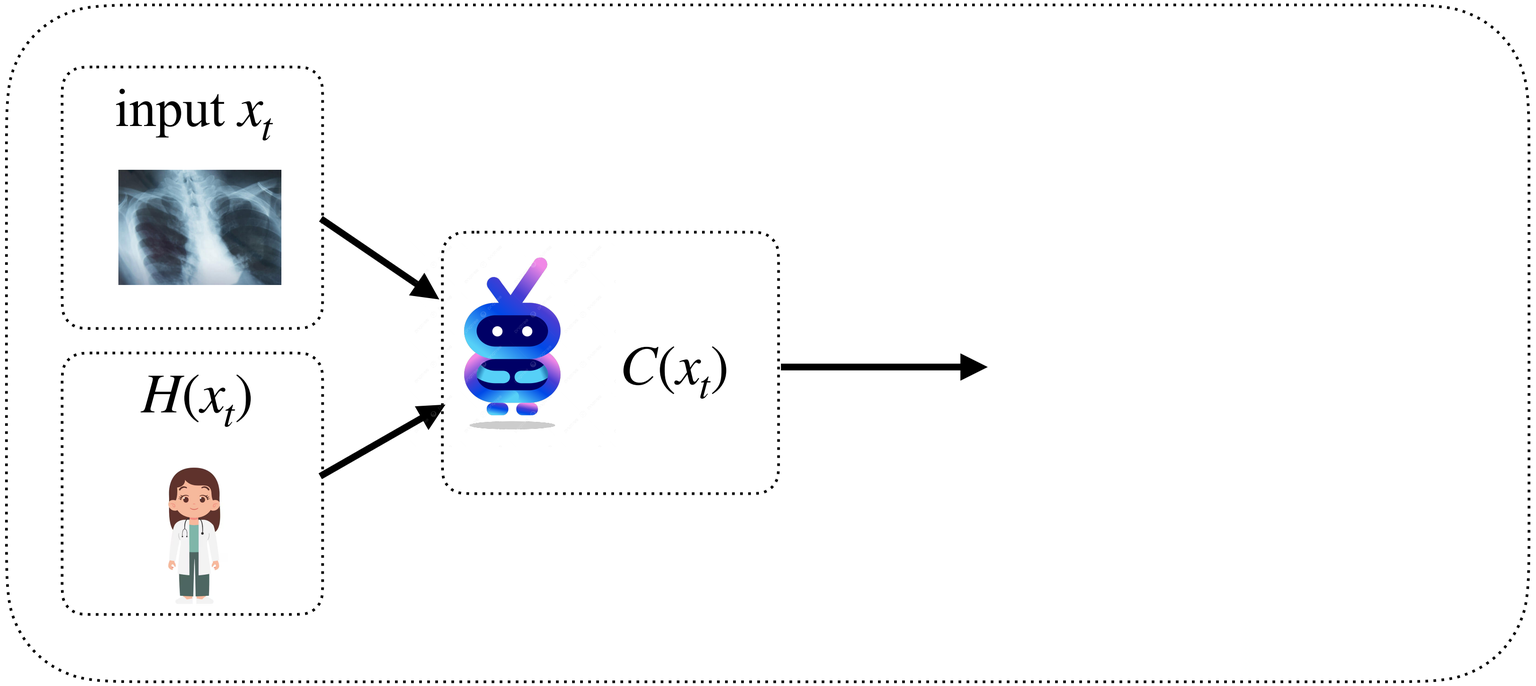
Question: How to debias the thresholds in the online setting

At round t



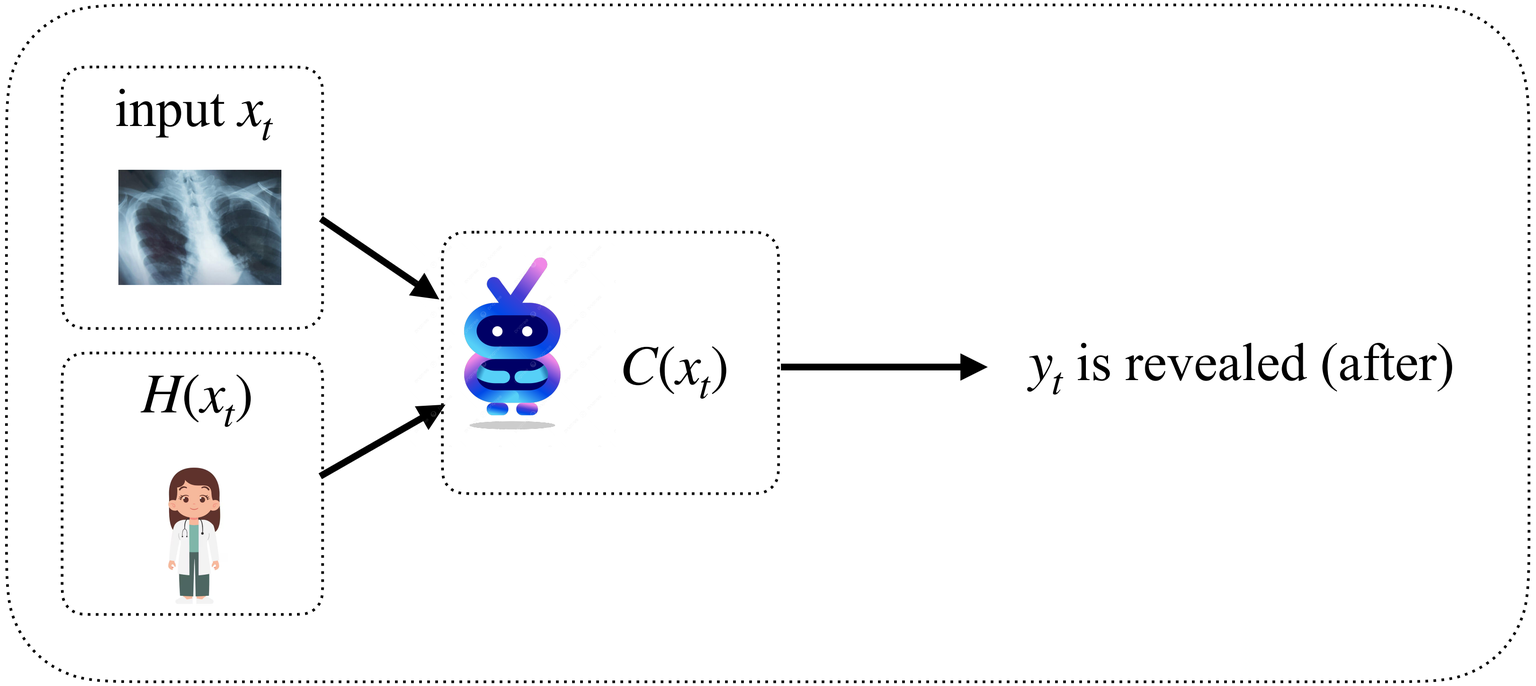
Question: How to debias the thresholds in the online setting

At round t



Question: How to debias the thresholds in the online setting

At round t



Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \middle| \; s(x_t, y) \leq a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \middle| \; s(x_t, y) \leq a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Question: How to debias the thresholds in the online setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Counterfactual Harm Error

Question: How to debias the thresholds in the **online** setting

At round t input x_t ; $H(x_t)$; y_t is revealed (after)

$$C_t(x_t) = \left\{ y \,\middle|\, s(x_t, y) \le a_t \mathbf{1}_{\{y \notin H(x_t)\}} + b_t \mathbf{1}_{\{y \in H(x_t)\}} \right\}$$

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Complementarity Error

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Complementarity Error

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t) \,) \longrightarrow$$

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}} \qquad \mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t)\,) \longrightarrow \quad b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}} \qquad \mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

$$\textit{If} \quad (\, y_t \in H(x_t)\,) \longrightarrow \quad b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}} \qquad \mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Counterfactual Harm Error | **Complementarity Error**

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Complementarity Error

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

Question: How to debias the thresholds in the online setting

We track two errors

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

Complementarity Error

We update the thresholds as follows

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| \leq \frac{1 + \eta \ \max(\varepsilon, 1-\varepsilon)}{\eta\, N_1(T)}$$

Question: How to debias the thresholds in the online setting

Counterfactual Harm Error

$$\mathrm{err}_t^{\mathrm{in}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \in H(x_t) \,\}}$$

If $(\, y_t \in H(x_t) \,)$

$$b_{t+1} = b_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > b_t \,\}} - \varepsilon\right),$$

$$a_{t+1} = a_t$$

Complementarity Error

$$\mathrm{err}_t^{\mathrm{out}} := \mathbf{1}_{\{\, y_t \notin C_t(x_t),\ y_t \notin H(x_t) \,\}}$$

If $(\, y_t \notin H(x_t) \,)$

$$a_{t+1} = a_t + \eta\left(\mathbf{1}_{\{\, s(x_t,y_t) > a_t \,\}} - \delta\right),$$

$$b_{t+1} = b_t$$

Theorem: Assume the conformity score is bounded, i.e $s(x, y) \in [0,1]$:

$$\left| \frac{1}{N_1(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{in}} - \varepsilon \right| \leq \frac{1 + \eta \ \max(\varepsilon, 1-\varepsilon)}{\eta\, N_1(T)}$$

$$\left| \frac{1}{N_2(T)} \sum_{t=1}^{T} \mathrm{err}_t^{\mathrm{out}} - \delta \right| \leq \frac{1 + \eta \ \max(\delta, 1-\delta)}{\eta\, N_2(T)}$$

Experiments

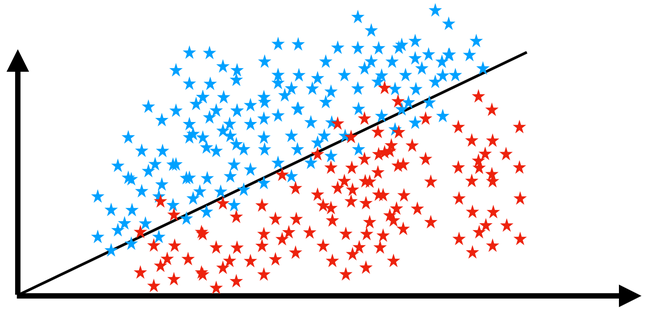
We consider three modalities of data

Experiments

We consider three modalities of data

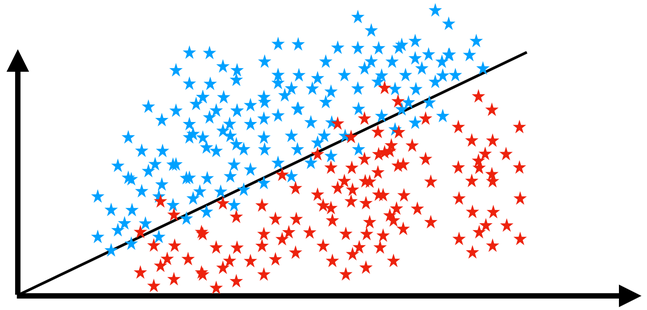
Experiments

We consider three modalities of data



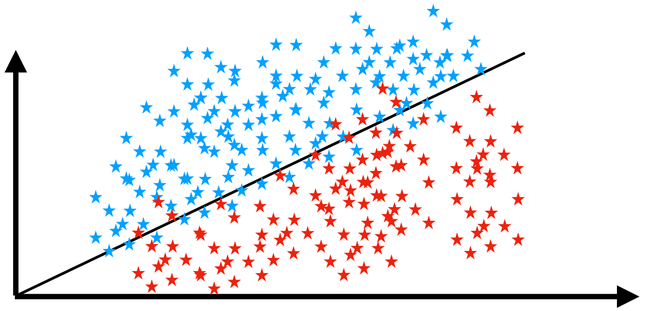
Experiments

We consider three modalities of data



Experiments

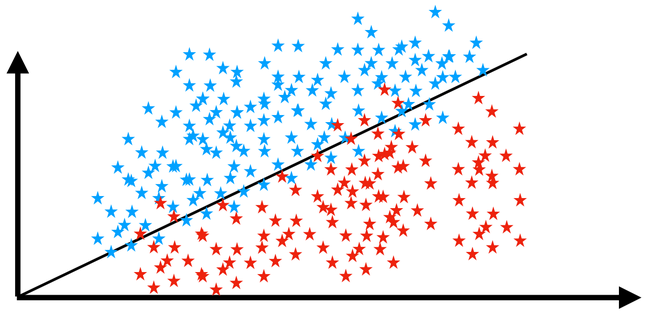
We consider three modalities of data



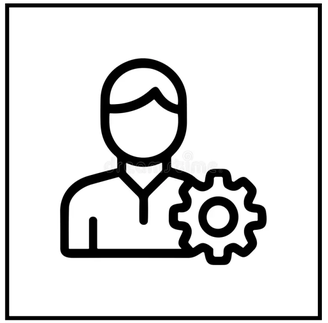
Baselines

Experiments

We consider three modalities of data



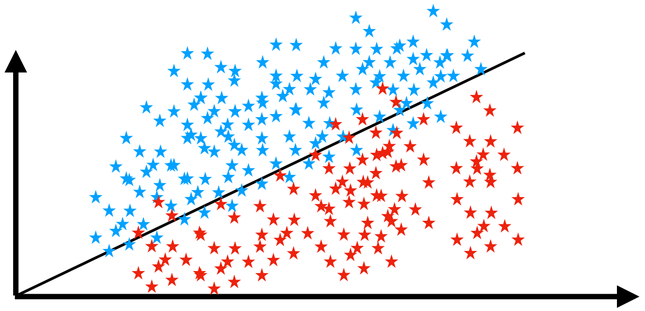
Baselines



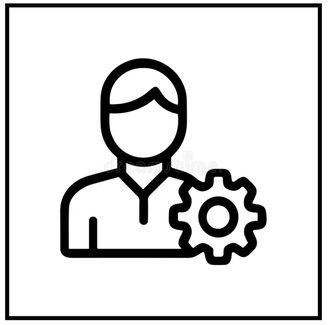
Human Alone

Experiments

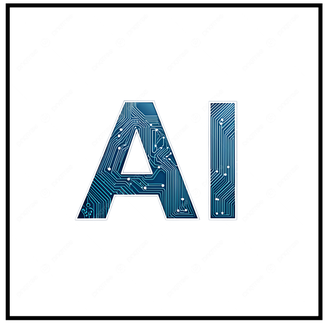
We consider three modalities of data



Baselines



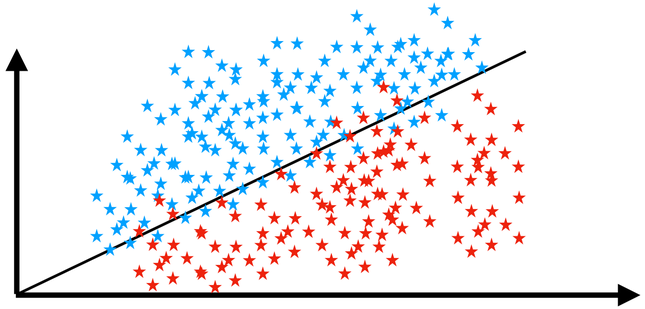
Human Alone



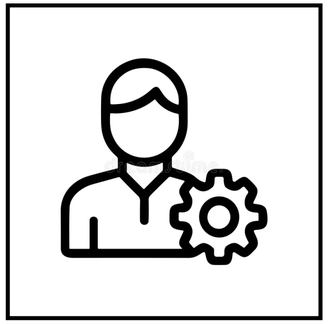
AI Alone

Experiments

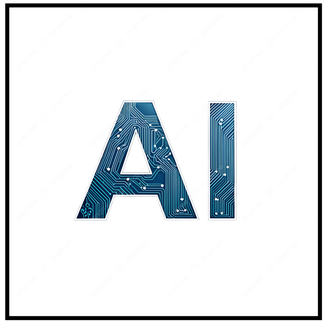
We consider three modalities of data



Baselines



Human Alone

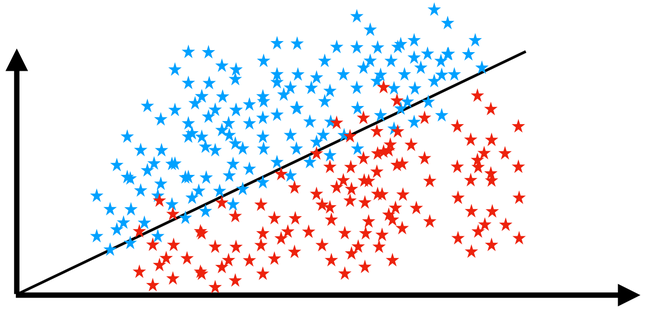


AI Alone

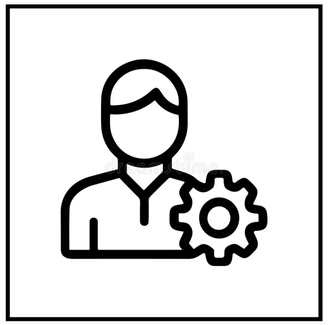
Evaluation Metics

Experiments

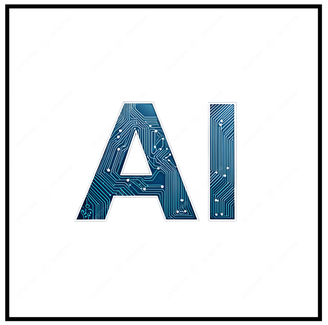
We consider three modalities of data



Baselines

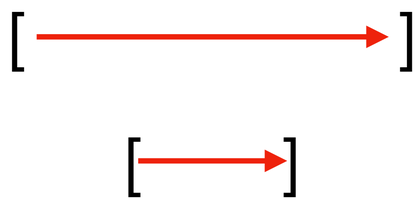


Human Alone



AI Alone

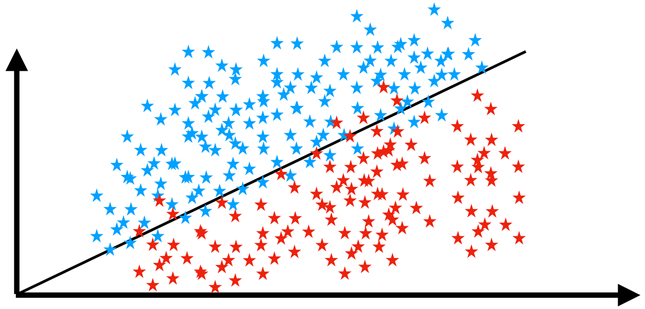
Evaluation Metics



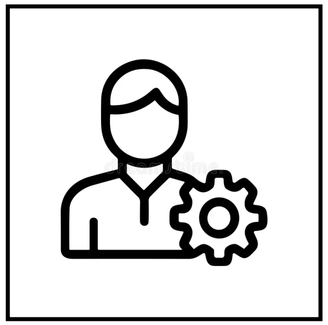
Set Size

Experiments

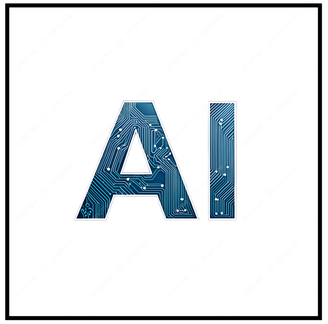
We consider three modalities of data



Baselines

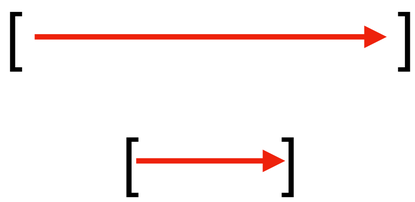


Human Alone

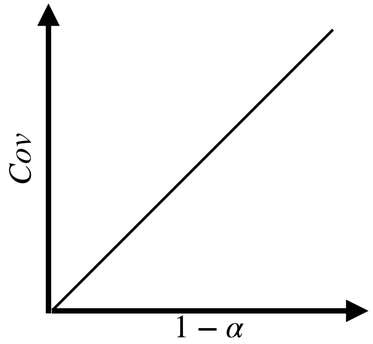


AI Alone

Evaluation Metics

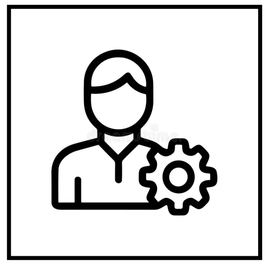


Set Size

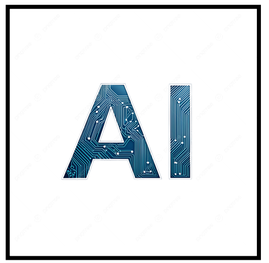


Marginal Coverage

Classification

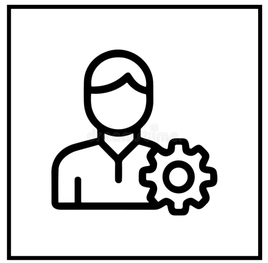


Crowdsourcing

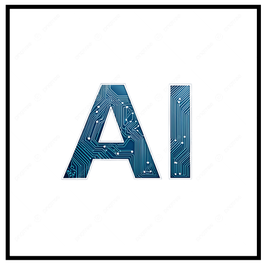


AlexNet

Classification



Crowdsourcing

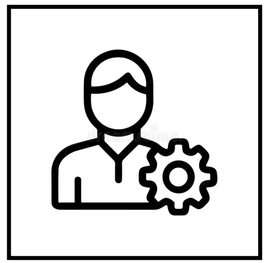


AlexNet

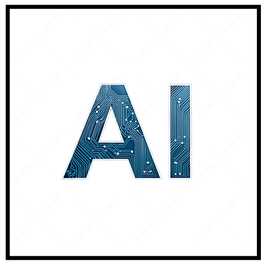
Offline Setting

$\omega = 125$								
	Human Alone		**CUP**				**AI Alone**	
Strategy	Coverage	Size	Coverage	Size	ε	δ	Coverage	Size
Top-2	0.8008 ± 0.0090	2.00 ± 0.00	$\mathbf{0.9022 \pm 0.0083}$	$\mathbf{1.49 \pm 0.04}$	0.05	0.70	0.9072 ± 0.0138	1.65 ± 0.07
Top-1	0.7245 ± 0.0103	1.00 ± 0.00	$\mathbf{0.8823 \pm 0.0134}$	1.36 ± 0.07	0.05	0.70	0.8828 ± 0.0140	1.48 ± 0.05

Classification



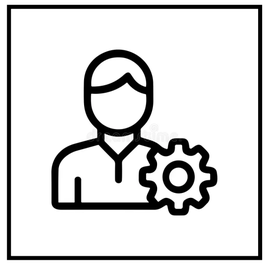
Crowdsourcing



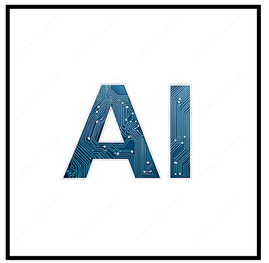
AlexNet

Online Setting - “ Human Adaptation to AI”

Classification

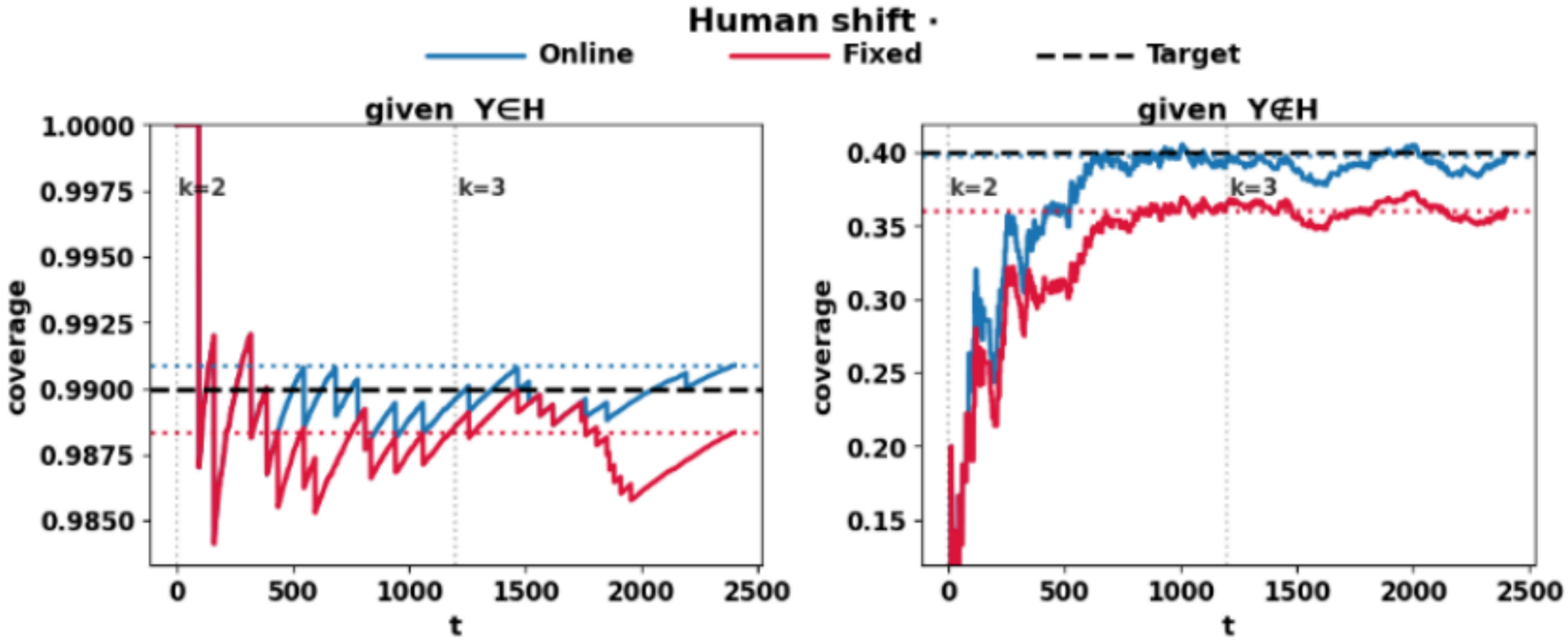


Crowdsourcing

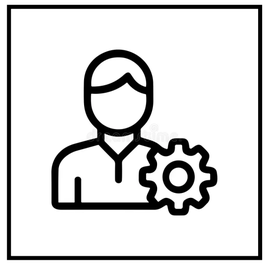


AlexNet

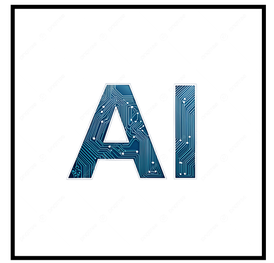
Online Setting - " Human Adaptation to AI"



Classification

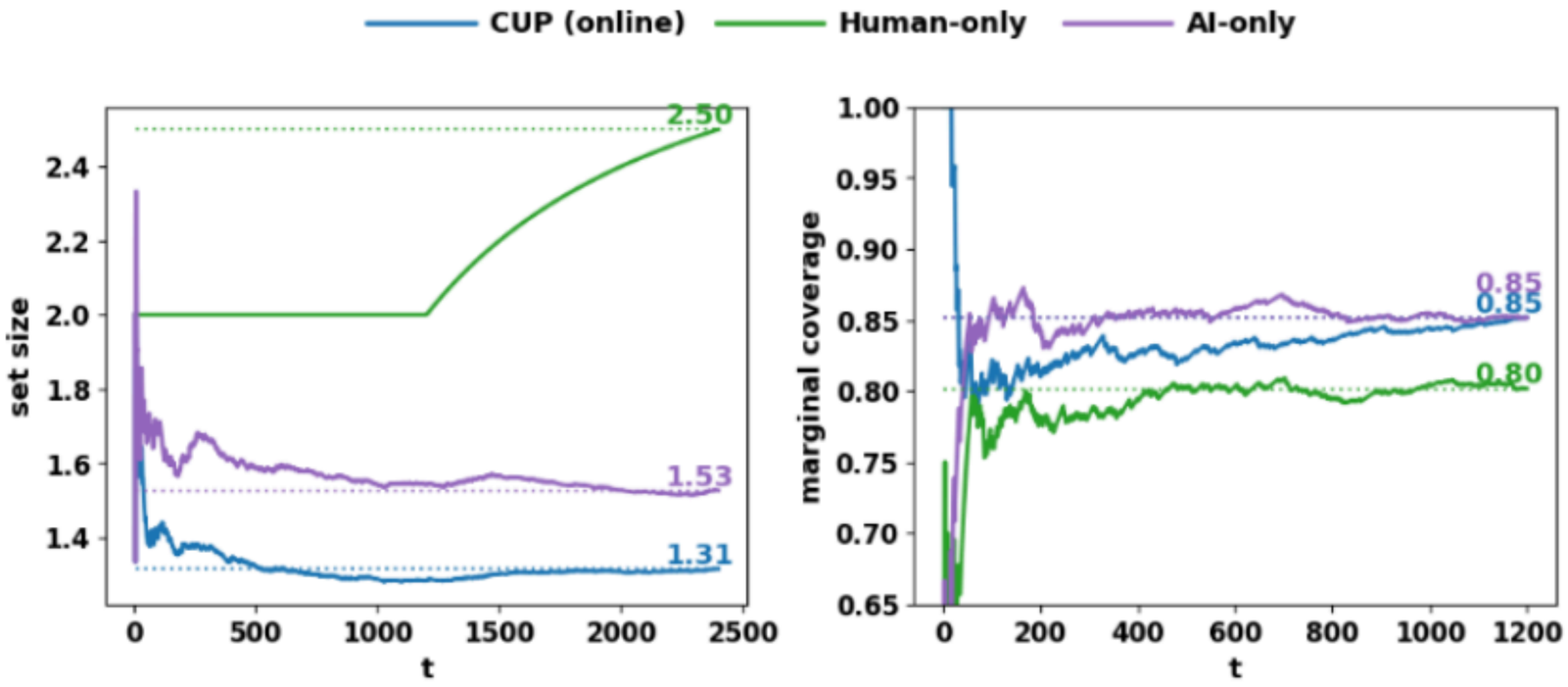


Crowdsourcing

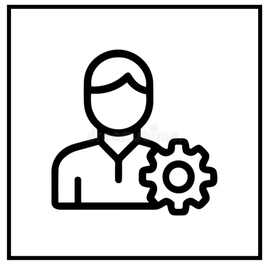


AlexNet

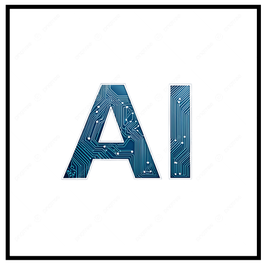
Online Setting - " Human Adaptation to AI"



Text Based Medical Diagnosis

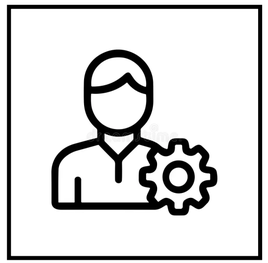


Rule-based diagnostic system

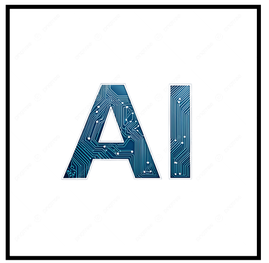


LLMs

Text Based Medical Diagnosis



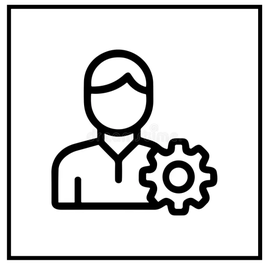
Rule-based diagnostic system



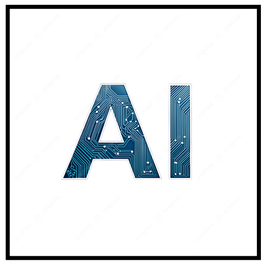
LLMs

Offline Setting

Text Based Medical Diagnosis



Rule-based diagnostic system

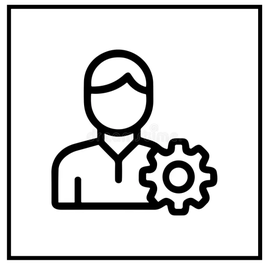


LLMs

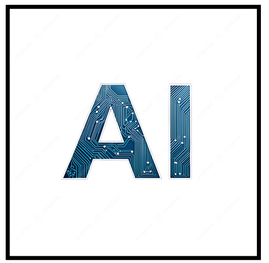
Offline Setting

	Human	GPT-4o			GPT-5		
Strategy	**C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**
Top-1	0.71 / 1.00	0.90 / 2.84	(0.02, 0.70)	0.88 / 4.64	0.91 / 1.59	(0.02, 0.70)	0.91 / 1.76
Top-2	0.87 / 1.95	0.93 / 3.14	(0.01, 0.45)	0.90 / 9.12	0.93 / 1.65	(0.02, 0.45)	0.93 / 1.95

Text Based Medical Diagnosis



Rule-based diagnostic system



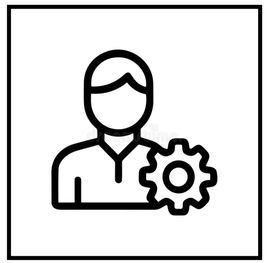
LLMs

Offline Setting

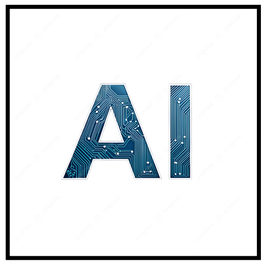
	Human	GPT-4o			GPT-5		
Strategy	**C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**	**CUP C/S**	**(ϵ, δ)**	**AI C/S**
Top-1	0.71 / 1.00	0.90 / 2.84	(0.02, 0.70)	0.88 / 4.64	0.91 / 1.59	(0.02, 0.70)	0.91 / 1.76
Top-2	0.87 / 1.95	0.93 / 3.14	(0.01, 0.45)	0.90 / 9.12	0.93 / 1.65	(0.02, 0.45)	0.93 / 1.95

AI's quality affects the overall collaboration quality!

Text Based Medical Diagnosis



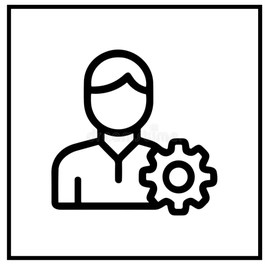
Rule-based diagnostic system



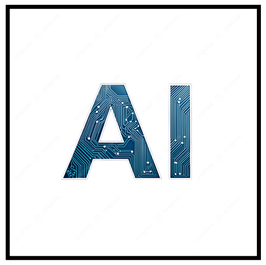
LLMs

Online Setting - "age shift"

Text Based Medical Diagnosis

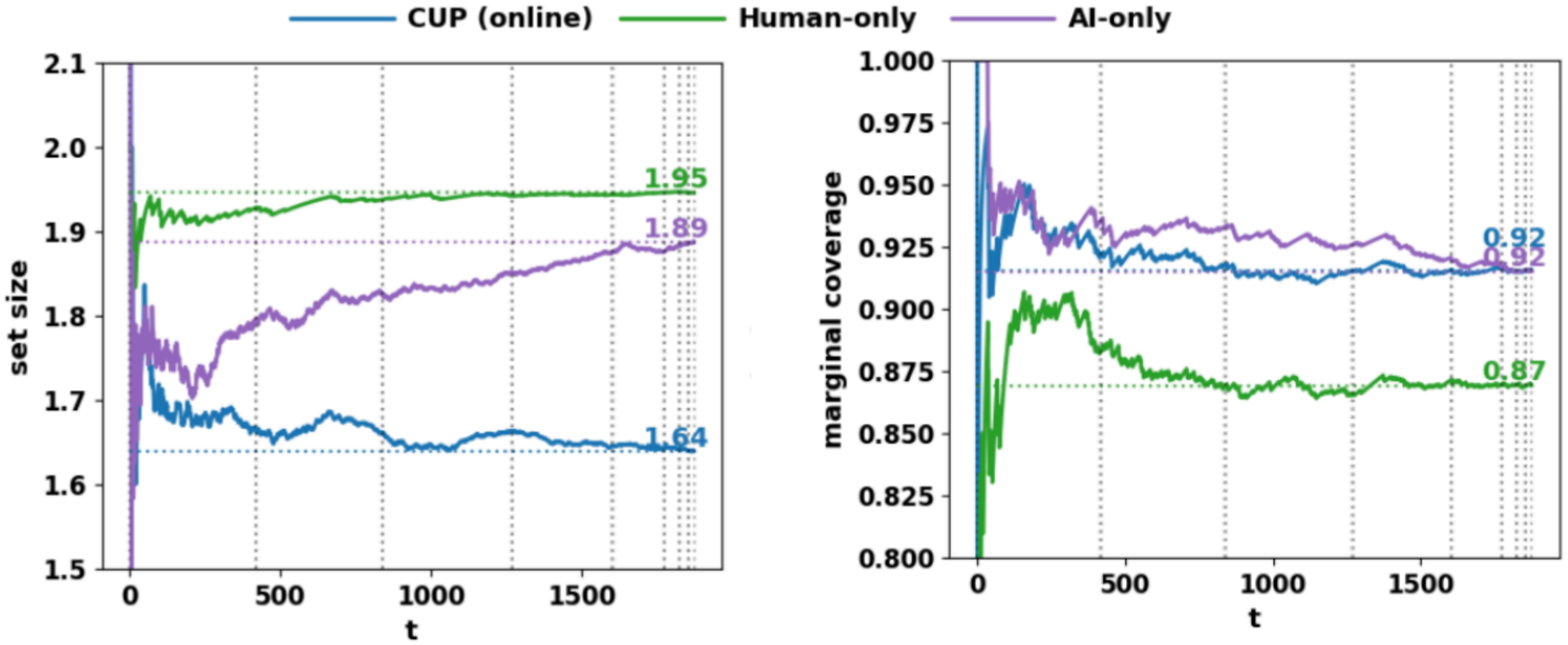


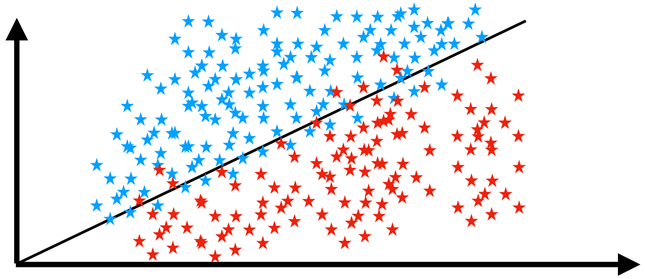
Rule-based diagnostic system



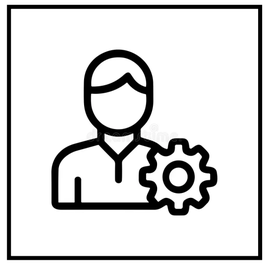
LLMs

Online Setting - "age shift"

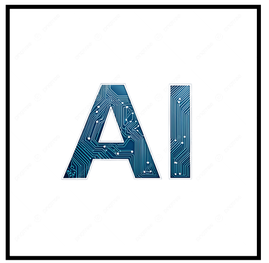




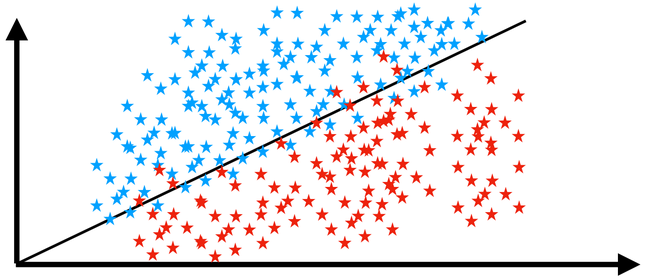
Regression



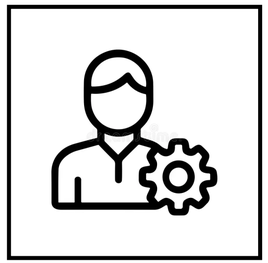
Synthetic



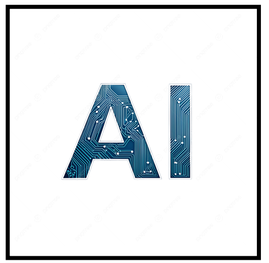
MLP



Regression



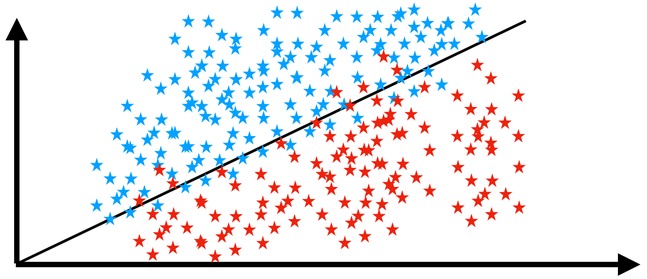
Synthetic



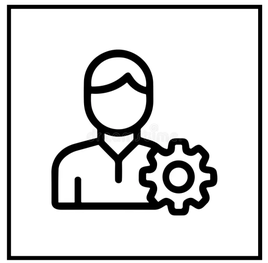
MLP

Offline Setting

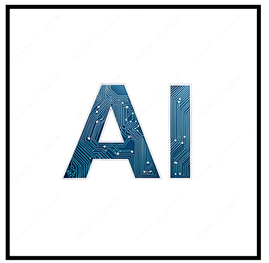
Human A				Human B			
Human A C/S	**CUP C/S**	$(\varepsilon, 1-\delta)$	**AI C/S**	**Human B C/S**	**CUP C/S**	$(\varepsilon, 1-\delta)$	**AI C/S**
0.760 / 0.581	0.862 / 0.380	(0.10, 0.70)	0.862 / 0.394	0.872 / 0.618	0.948 / 0.528	(0.05, 0.90)	0.948 / 0.608
0.760 / 0.581	0.825 / 0.326	(0.15, 0.70)	0.825 / 0.337	0.872 / 0.618	0.953 / 0.558	(0.05, 0.95)	0.953 / 0.588



Regression



Synthetic

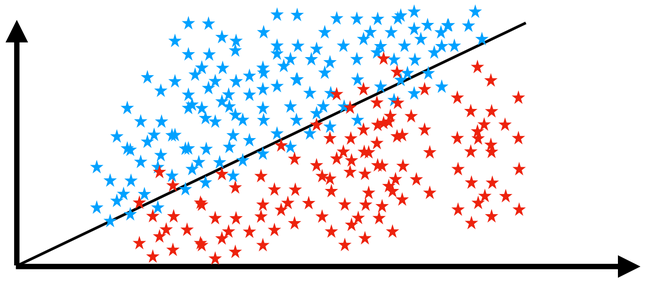


MLP

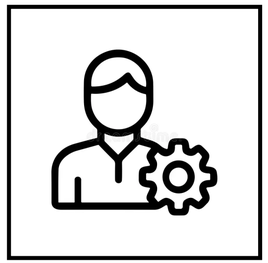
Offline Setting

Human A				Human B			
Human A C/S	**CUP C/S**	$(\varepsilon, 1-\delta)$	**AI C/S**	**Human B C/S**	**CUP C/S**	$(\varepsilon, 1-\delta)$	**AI C/S**
0.760 / 0.581	0.862 / 0.380	(0.10, 0.70)	0.862 / 0.394	0.872 / 0.618	0.948 / 0.528	(0.05, 0.90)	0.948 / 0.608
0.760 / 0.581	0.825 / 0.326	(0.15, 0.70)	0.825 / 0.337	0.872 / 0.618	0.953 / 0.558	(0.05, 0.95)	0.953 / 0.588

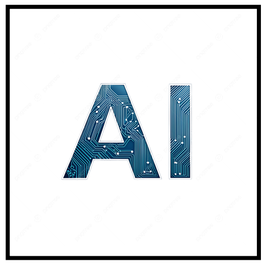
Human's quality affects the overall collaboration quality!



Regression

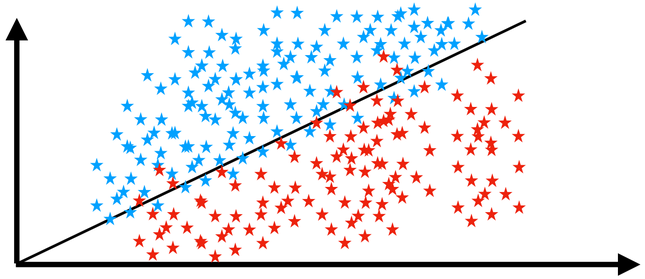


Synthetic

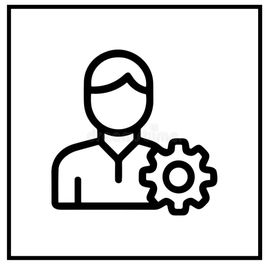


MLP

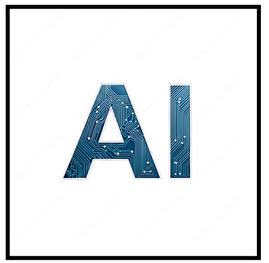
Online Setting - “demographic shift”



Regression

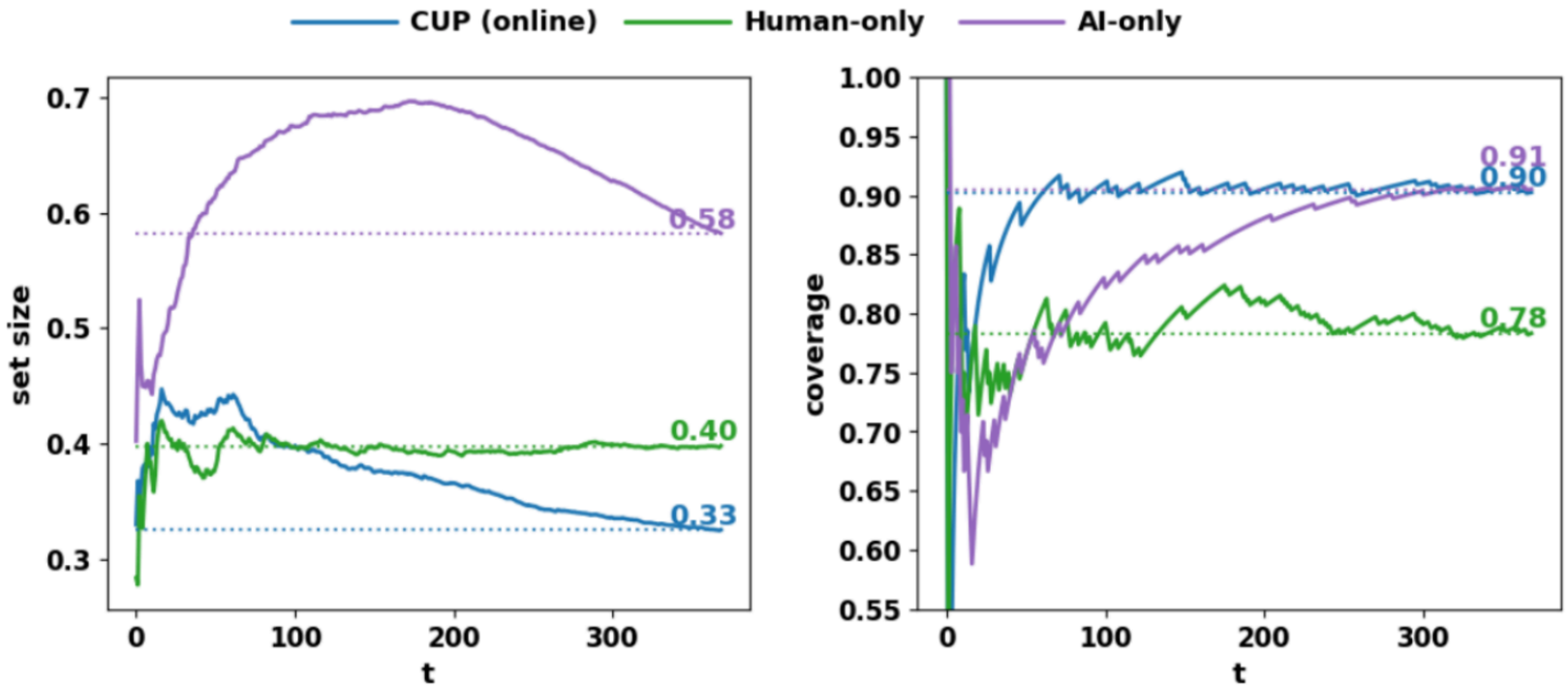


Synthetic



MLP

Online Setting - "demographic shift"



Thank You!