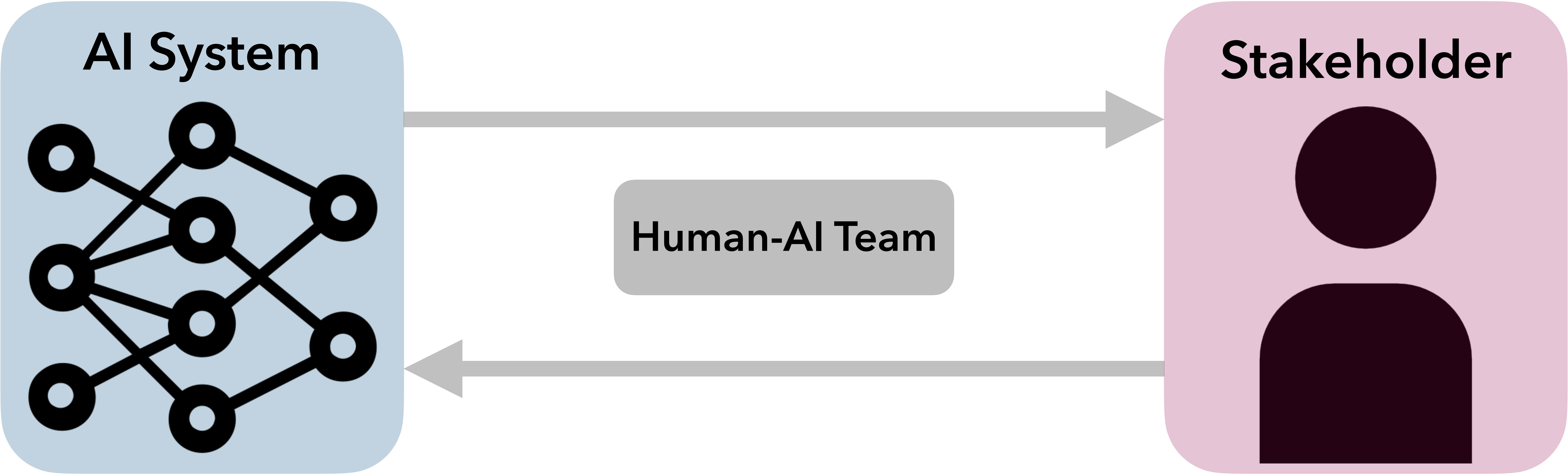
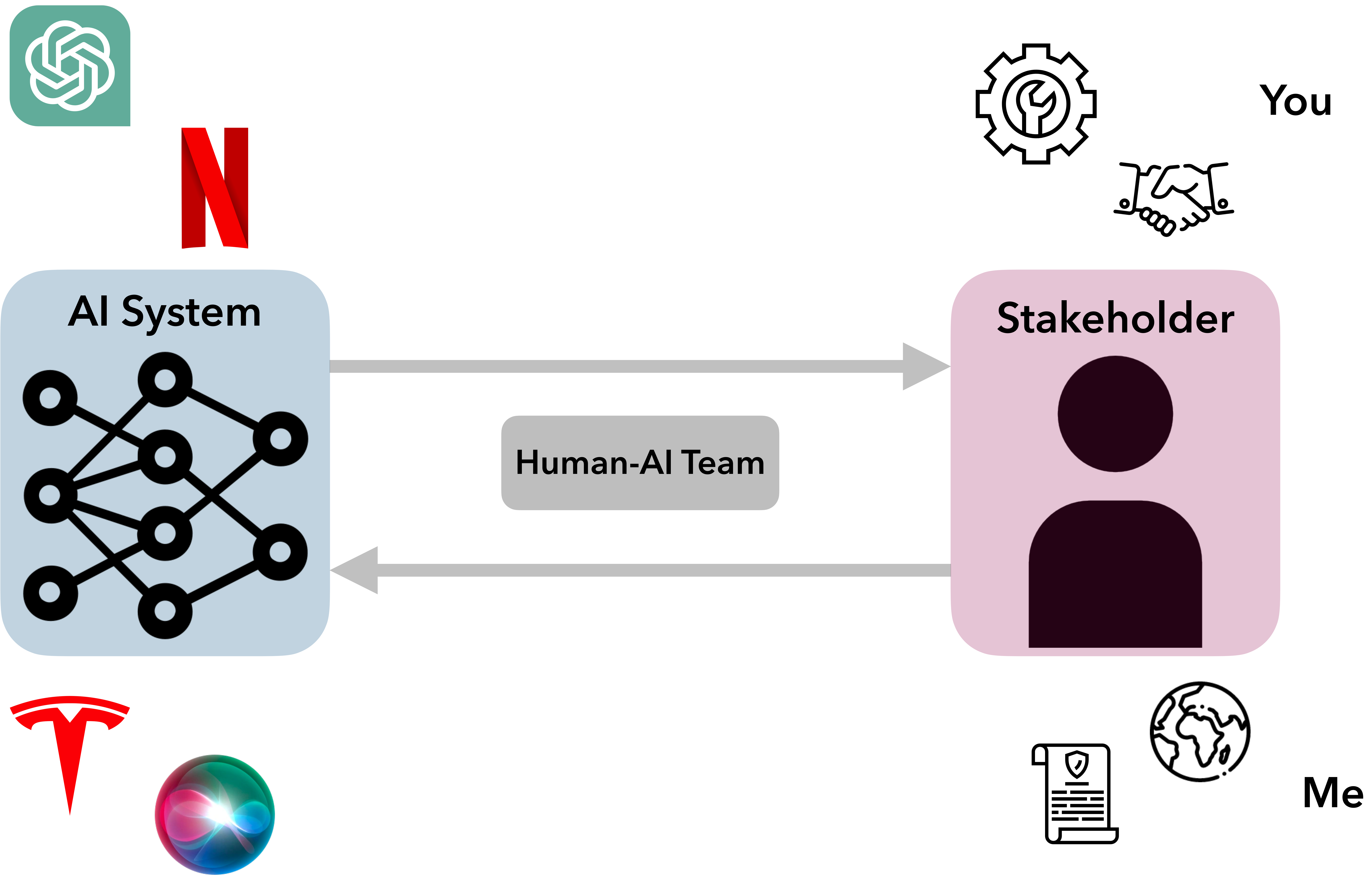
Trustworthy Machine Learning

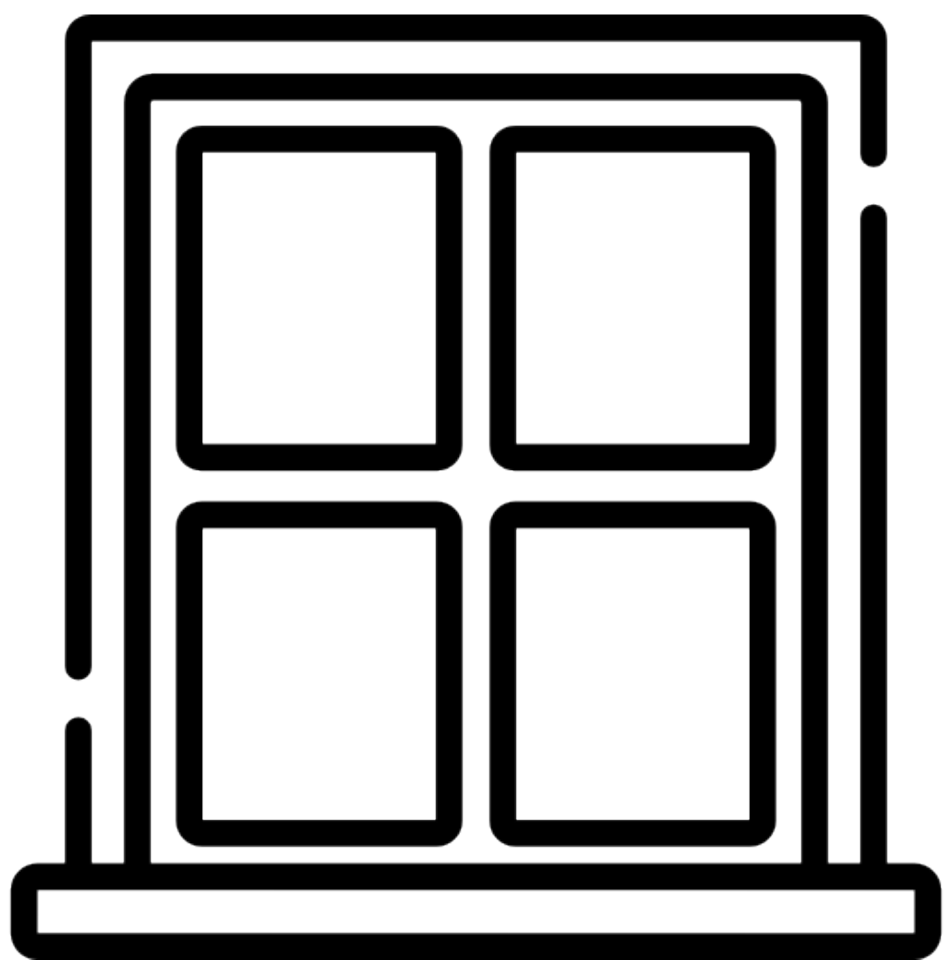
Umang Bhatt
Assistant Professor/Faculty Fellow, New York University
Senior Research Associate, The Alan Turing Institute
Associate Fellow, Leverhulme Center for the Future of Intelligence

@umangsbhatt
umangbhatt@nyu.edu


AI System
Human-AI Team
Stakeholder


AI System
Stakeholder
Human-AI Team
You
Me

Research Agenda



Transparency

Collaboration

Evaluation

Research Style

Convenings

Methods

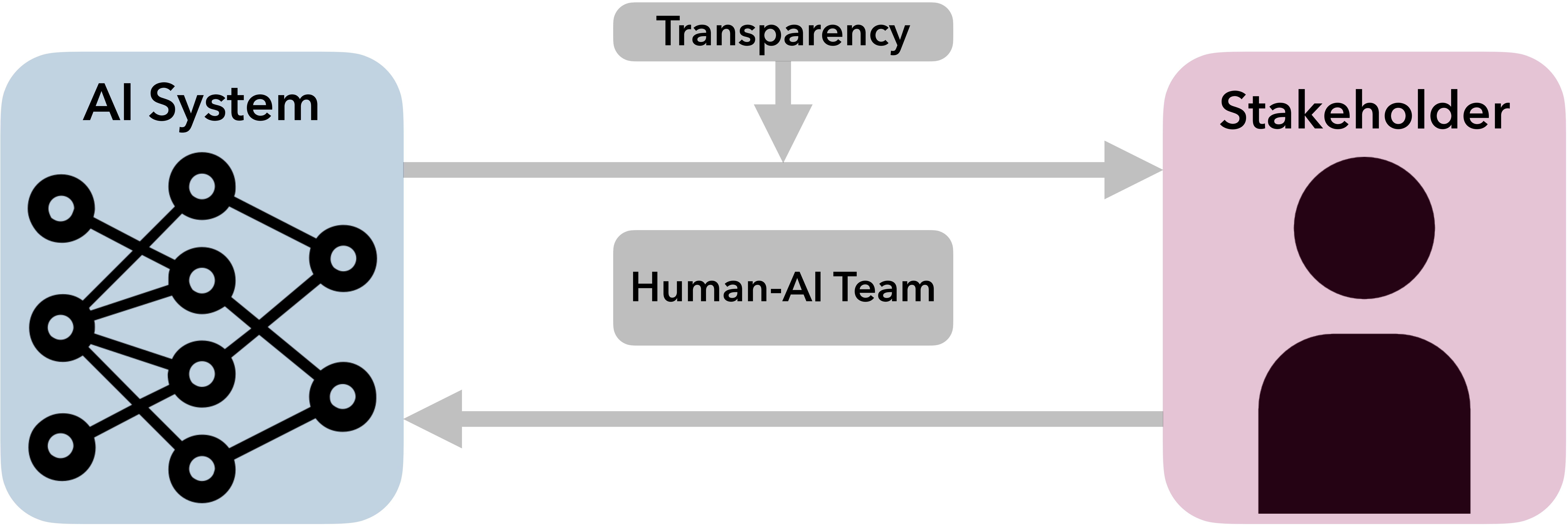
User Studies

Transparency Mechanisms

*How can we use transparency mechanisms to demonstrate the **trustworthiness** of AI systems?*

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

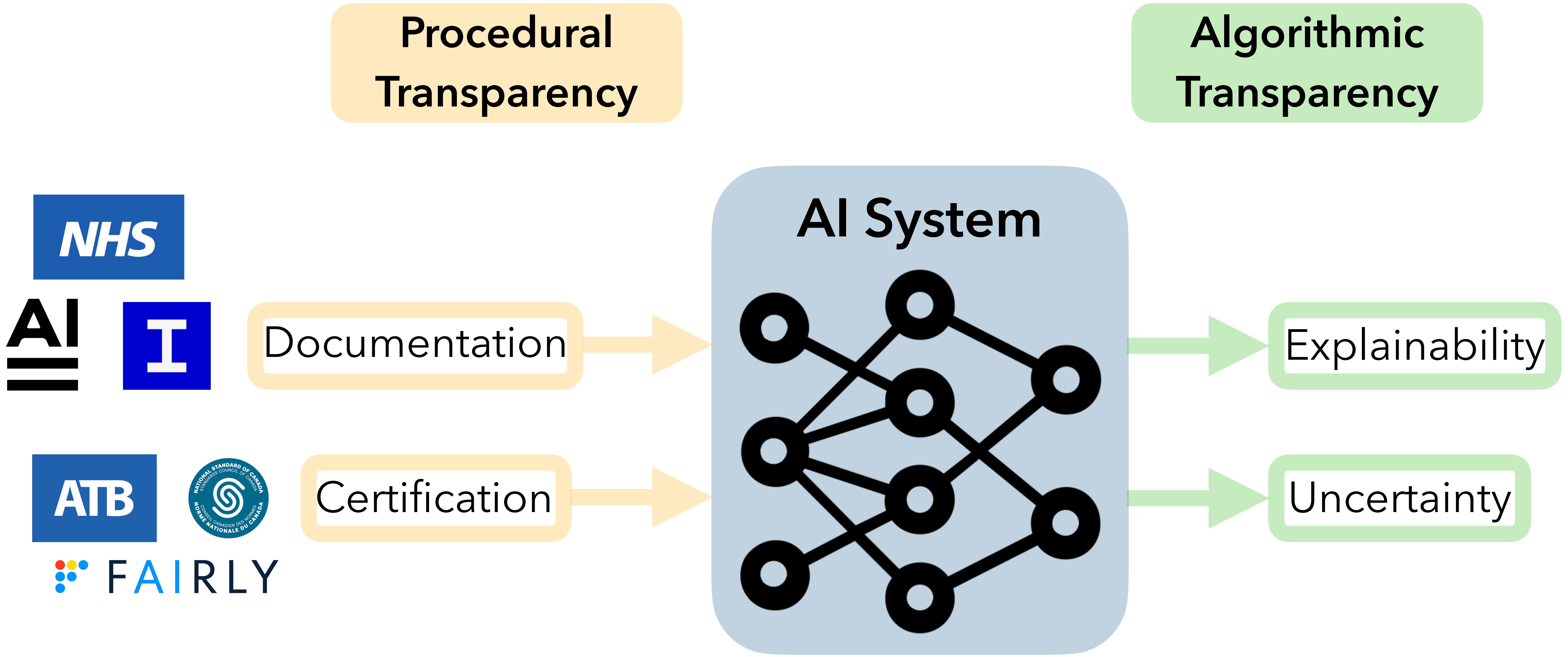
Transparency Mechanisms



Transparency means providing stakeholders with *relevant* information about how a system works

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Transparency Mechanisms

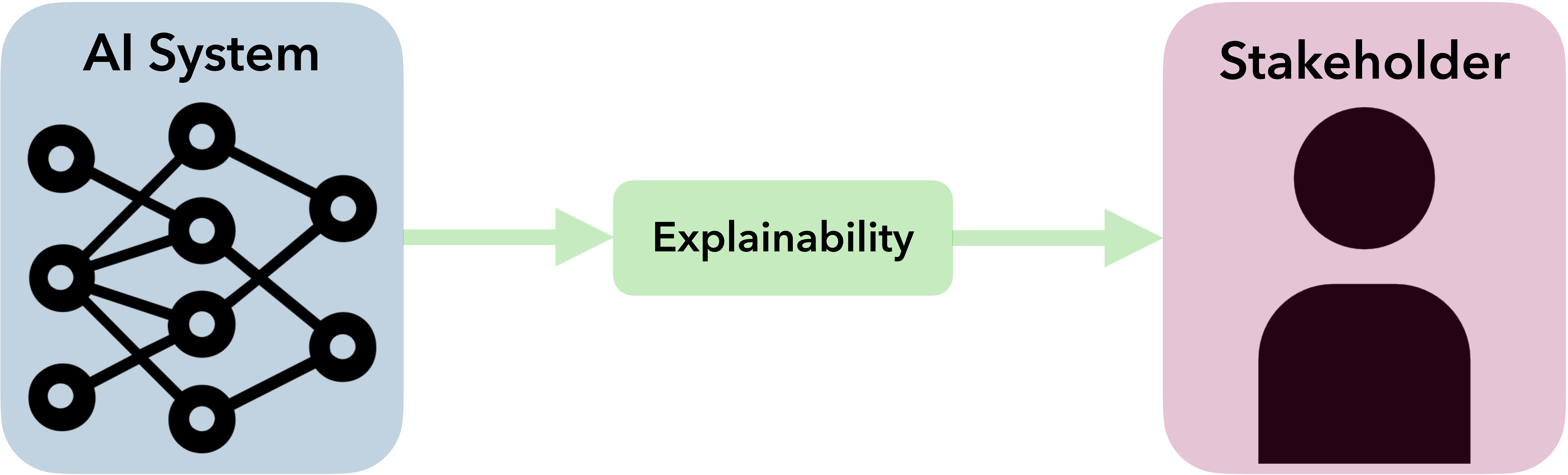


Brogle, Kallina, Sargeant, Shankar, Casovan, Weller, **B**. *Context-Specific Certification of AI Systems: A Pilot in the Financial Industry*. Under Review. 2024.

Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback*. ACM EAAMO. 2023.

B, Shams. *Trust in Artificial Intelligence: Clinicians Are Essential*. Chapter 10 in Healthcare Information Technology for Cardiovascular Medicine. 2021.

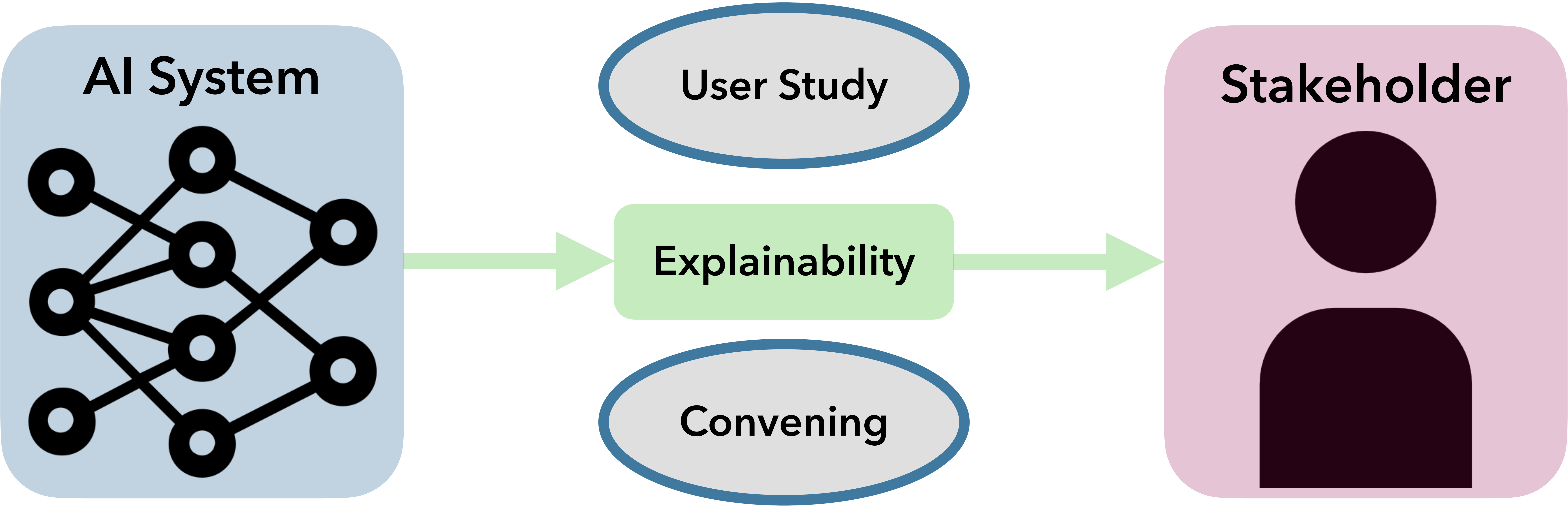
Transparency Mechanisms



Explainability means providing insight into a model's behavior for specific datapoint(s)

B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders.* ICML WHI. 2020.
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Transparency Mechanisms



PARTNERSHIP ON AI

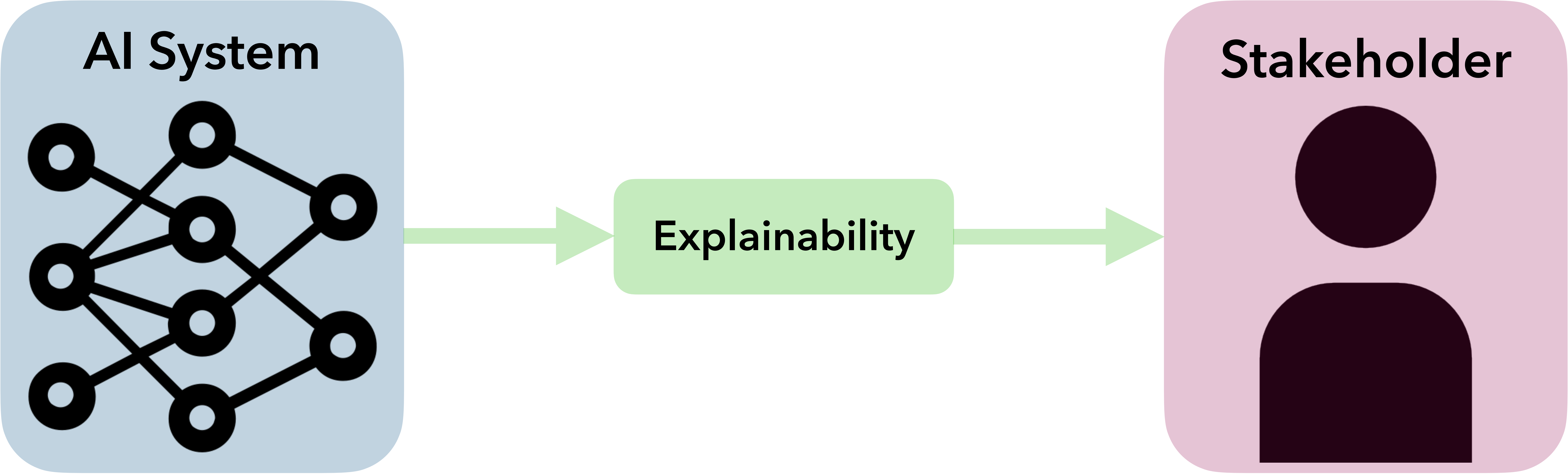
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

IBM

B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders.* ICML WHI. 2020.

B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

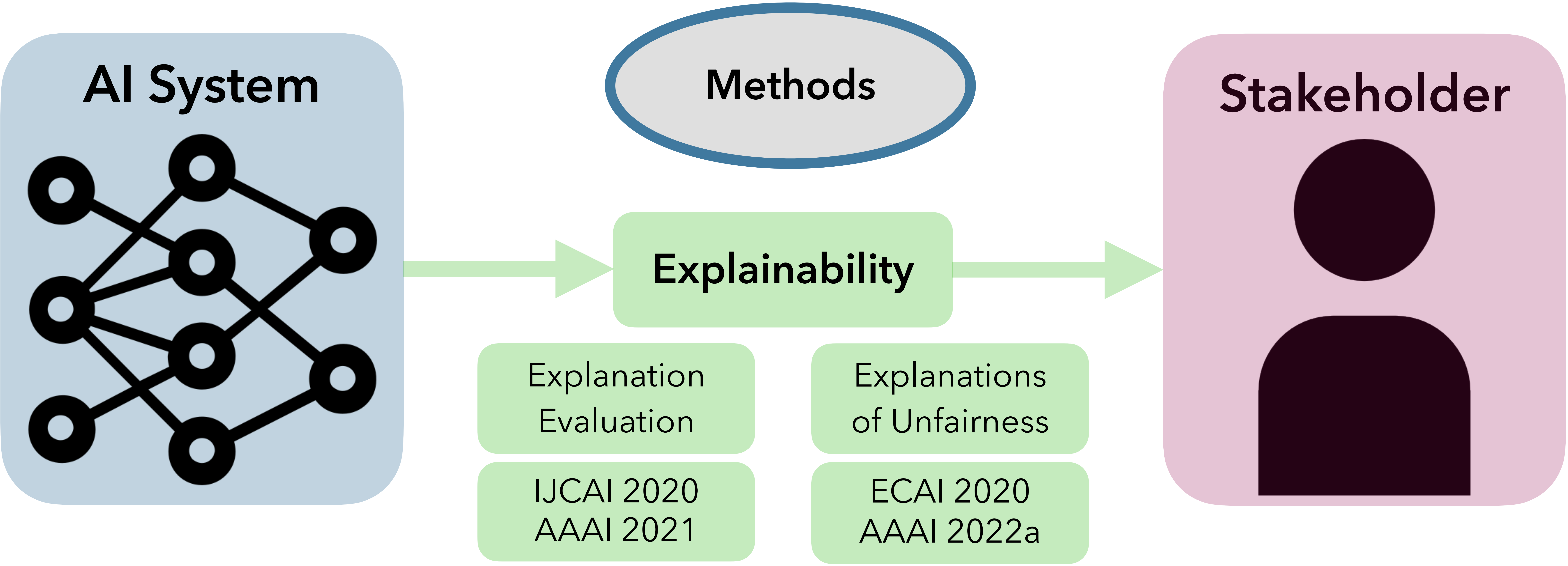
Transparency Mechanisms



Explainability methods are **not** in service of transparency goals within organizations

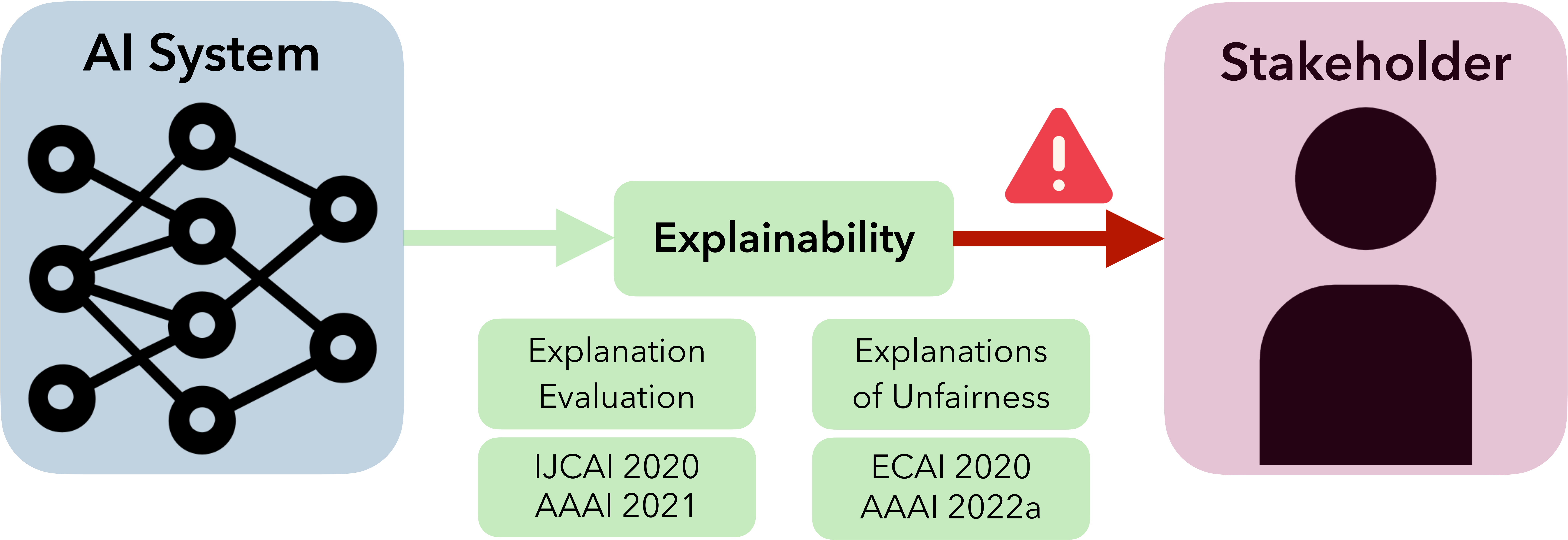
B, Andrus, Xiang, Weller. *Machine Learning Explainability for External Stakeholders*. ICML WHI. 2020.
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment*. ACM FAccT. 2020.

Transparency Mechanisms



B, Moura, Weller. *Evaluating and Aggregating Feature-based Model Explanations*. IJCAI. 2020.
Chapman, **B**, Pazos, Schulz, Georgatzis. *FIMAP: Feature Importance by Minimal Adversarial Perturbation*. AAAI. 2021.
Dimanov, **B**, Jamnik, Weller. *You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods*. ECAI. 2020.
von Kügelgen, Karimi, **B**, Valera, Weller, Schölkopf. *On the fairness of causal algorithmic recourse*. AAAI. 2022.

Transparency Mechanisms



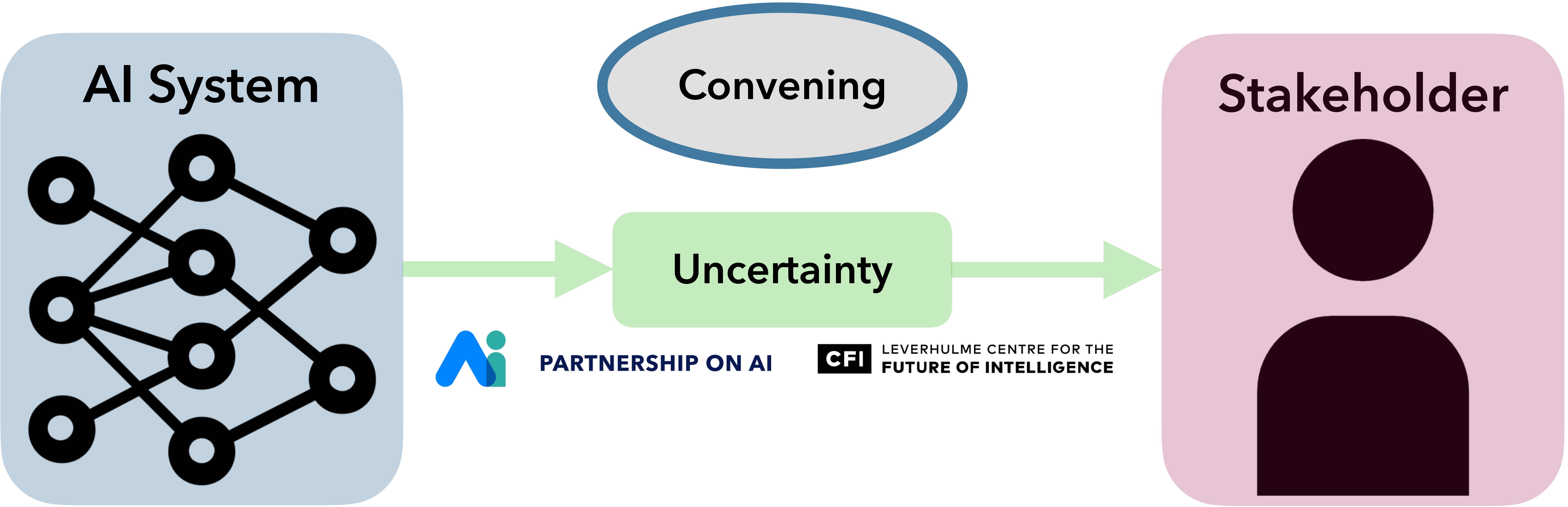
B, Moura, Weller. *Evaluating and Aggregating Feature-based Model Explanations*. IJCAI. 2020.

Chapman, **B**, Pazos, Schulz, Georgatzis. *FIMAP: Feature Importance by Minimal Adversarial Perturbation*. AAAI. 2021.

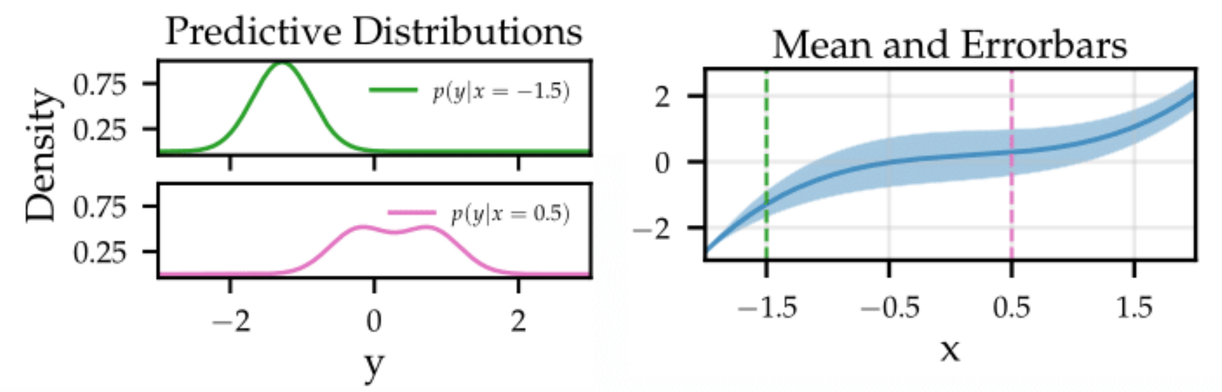
Dimanov, **B**, Jamnik, Weller. *You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods*. ECAI. 2020.

von Kügelgen, Karimi, **B**, Valera, Weller, Schölkopf. *On the fairness of causal algorithmic recourse*. AAAI. 2022.

Transparency Mechanisms



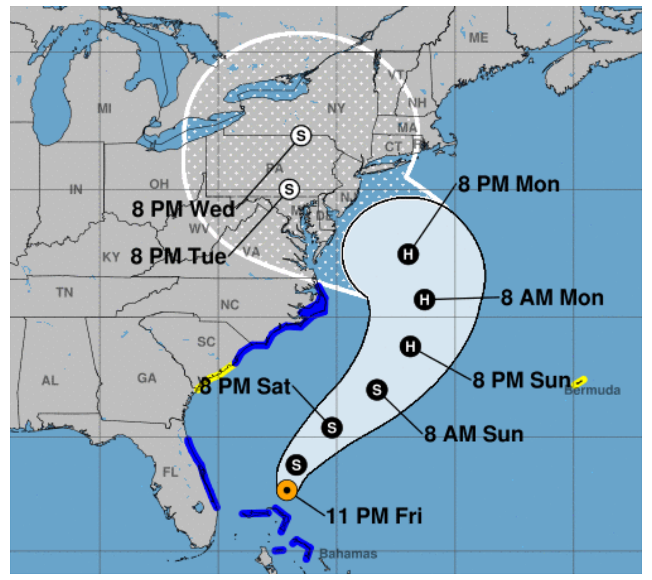
Step 1: Measuring



Step 2: Using

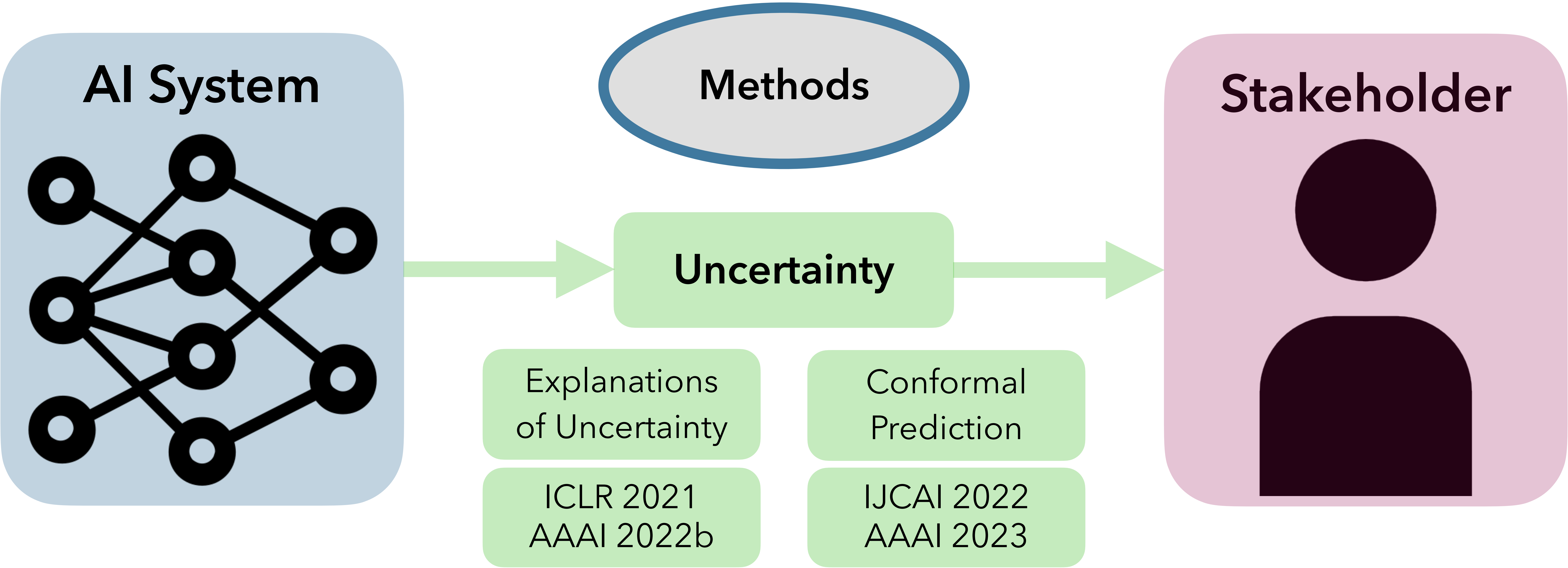
- **Fairness:** Measurement and Sampling Bias
- **Decision-Making:** Building Reject Option Classifiers
- **Trust Formation:** ABI

Step 3: Communicating



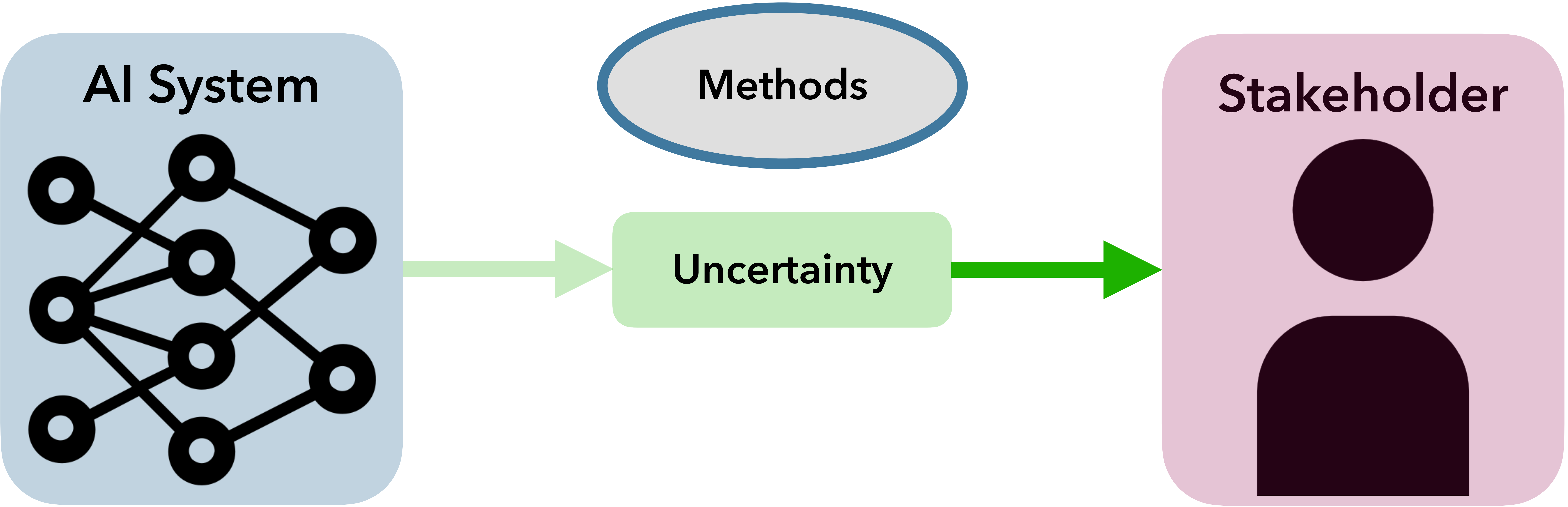
B, Antoran, Zhang, Liao, Sattigeri, Fogliato, et al. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*. ACM AIES. 2021.

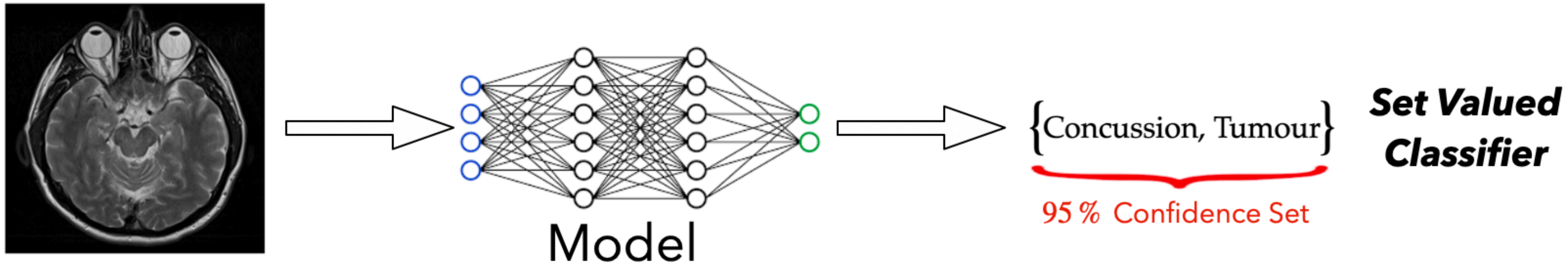
Transparency Mechanisms



Antoran, **B**, Adel, Weller, Hernandez-Lobato. *Getting a CLUE: A Method for Explaining Uncertainty Estimates.* ICLR. 2021.
Ley, **B**, Weller. *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates.* AAAI. 2022.
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams.* IJCAI. 2022.
Martinez, **B**, Weller, Cherubin. *Approximating full conformal prediction at scale via influence functions.* AAAI. 2023.

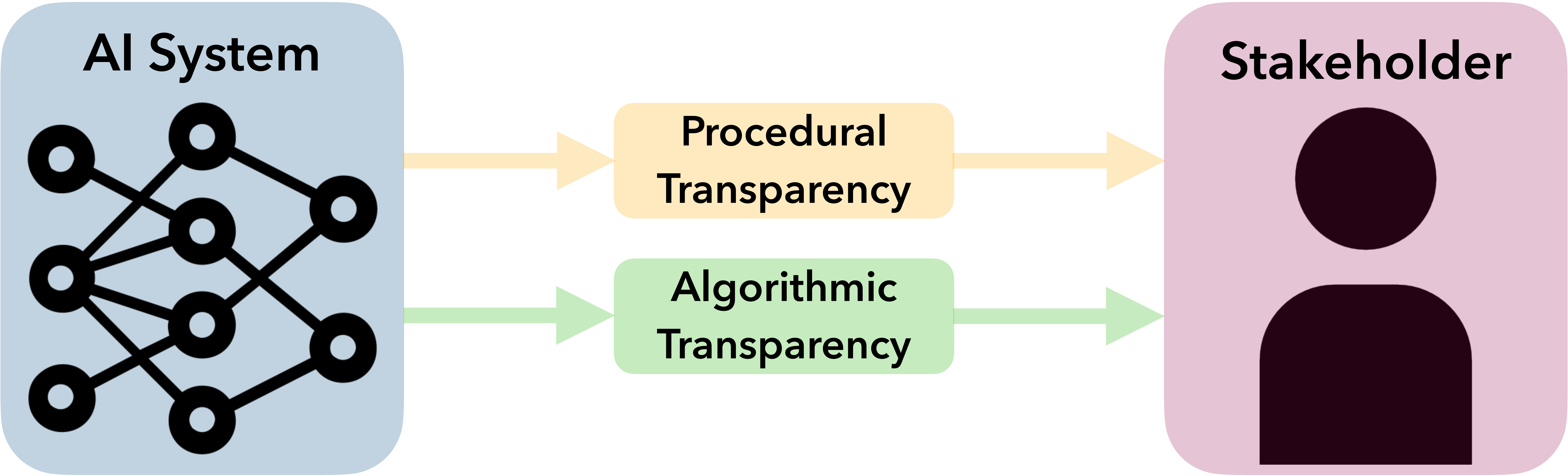
Transparency Mechanisms





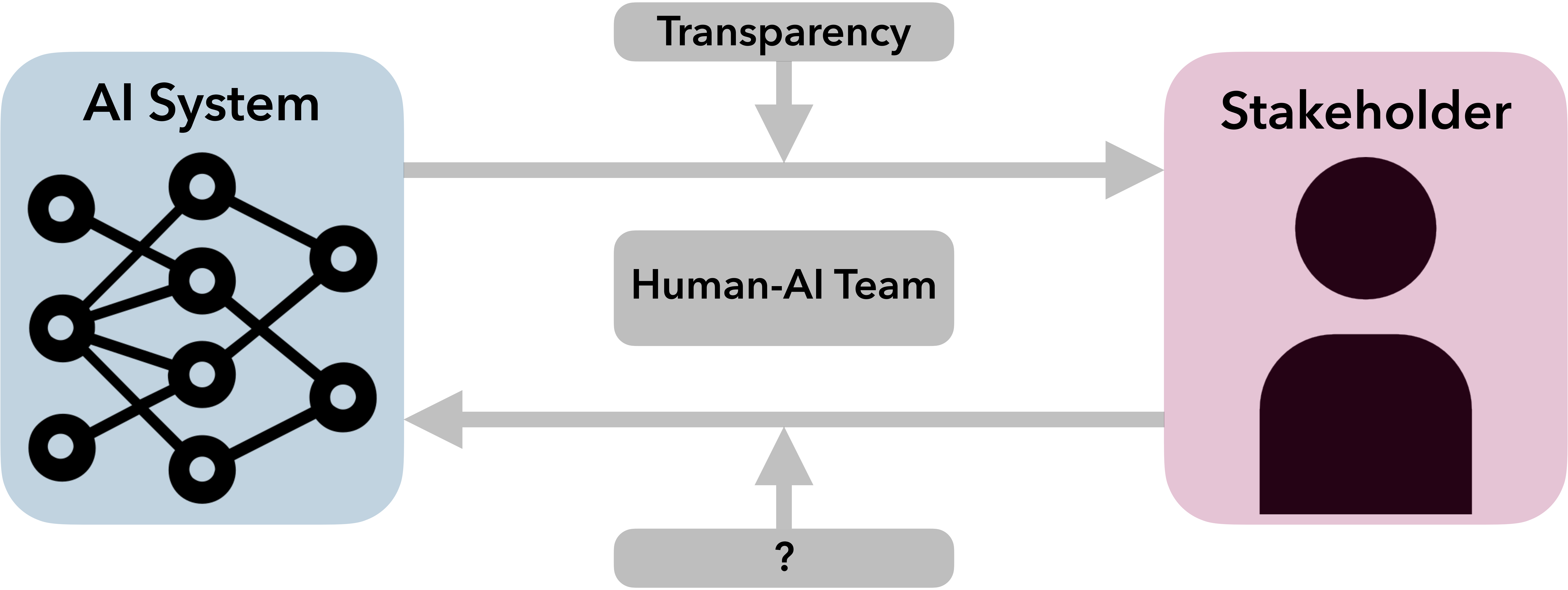
Babbar, B, Weller. *On the Utility of Prediction Sets in Human-AI Teams.* IJCAI. 2022.

Transparency Mechanisms



*How can we align transparency mechanisms with **regulatory** requirements?*

*How can we use **natural language** uncertainty explanations to improve trustworthiness?*



Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams.* IJCAI. 2022.
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* Patterns. 2023.

Effective Human-AI Collaboration

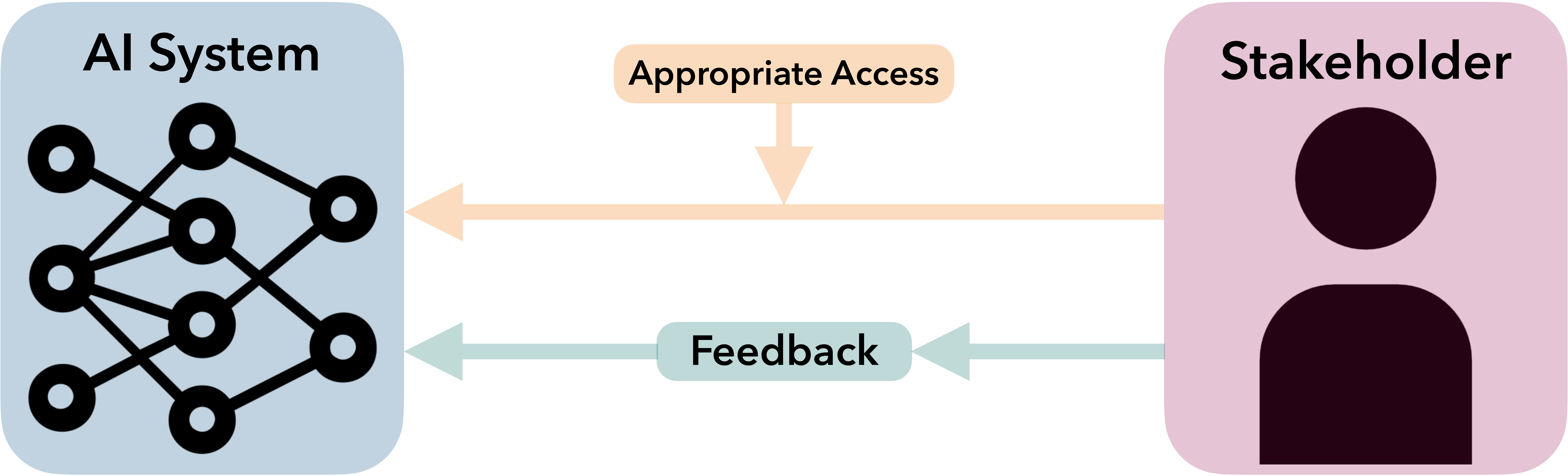
*How can AI systems work **alongside** human decision-makers?*

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies.* AAAI. 2025.

Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* Patterns. 2023.

Effective Human-AI Collaboration



B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies.* AAAI. 2025.

Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* Patterns. 2023.

Effective Human-AI Collaboration

Collins*, Sucholutsky*, **B***, Chandra, Wong, Lee, Zhang, Zhi-Xuan, Ho, Mansinghka, Weller, Tenenbaum, Griffiths. *Building machines that learn and think with people.* Nature Human Behavior. 2024.

Effective Human-AI Collaboration

Loafing
Stakeholder aligns ***all*** decisions with AI

Appreciation
Stakeholder aligns ***most*** decisions with AI

Vigilance

Aversion
Stakeholder aligns ***few*** decisions with AI

Opposition
Stakeholder aligns ***no*** decisions with AI

Overtrust

Distrust

$\ell(\hat{y}, \tilde{y}) = 0$ → *increases*

Dietvorst, Simmons, Massey. *Algorithm aversion: People Erroneously Avoid Algorithms after Seeing Them Err.* Journal of Experimental Psychology. 2015.
Logg, Minson, Moore. *Algorithm appreciation: People prefer algorithmic to human judgment.* Organizational Behavior and Human Decision Processes. 2019.
Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence.* Patterns. 2022.

Effective Human-AI Collaboration

POLITICS

Judge sanctions lawyers for brief written by A.I. with fake citations

PUBLISHED THU, JUN 22 2023·2:34 PM EDT | UPDATED THU, JUN 22 2023·3:53 PM EDT

Dan Mangan
@_DANMANGAN

SHARE

FROM AFP NEWS

Brazil Judge Investigated For AI Errors In Ruling

By AFP - Agence France Presse November 13, 2023

Tesla wins first US Autopilot trial involving fatal crash

By **Dan Levine** and **Hyunjoo Jin**

November 1, 2023 12:58 AM EDT · Updated a month ago

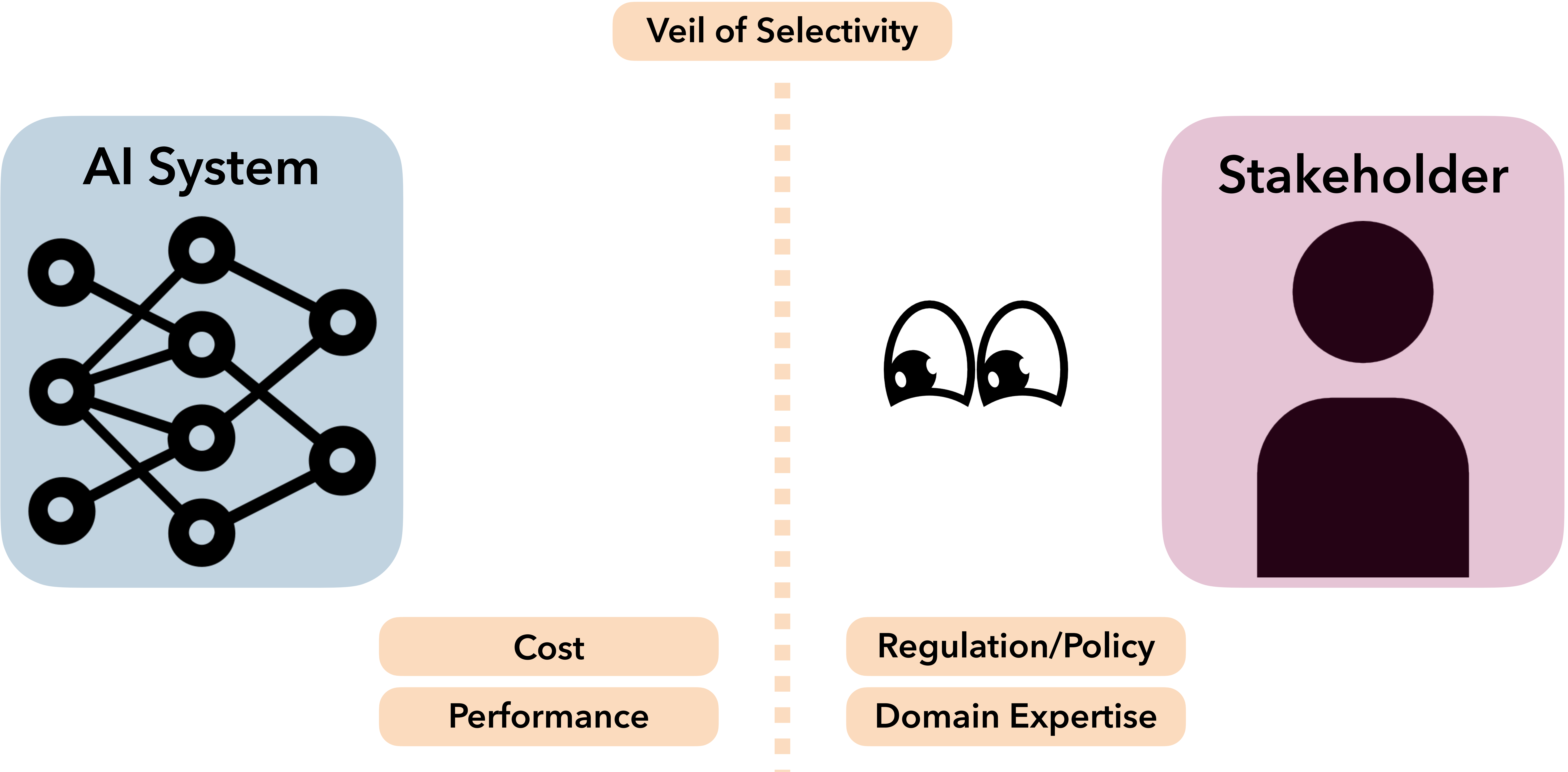
Is your health insurer using AI to deny you services? Lawsuit says errors harmed elders.

Ken Alltucker
USA TODAY

Published 5:18 a.m. ET Nov. 19, 2023 | **Updated 11:19 a.m. ET Nov. 20, 2023**

Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence*. Patterns. 2022.

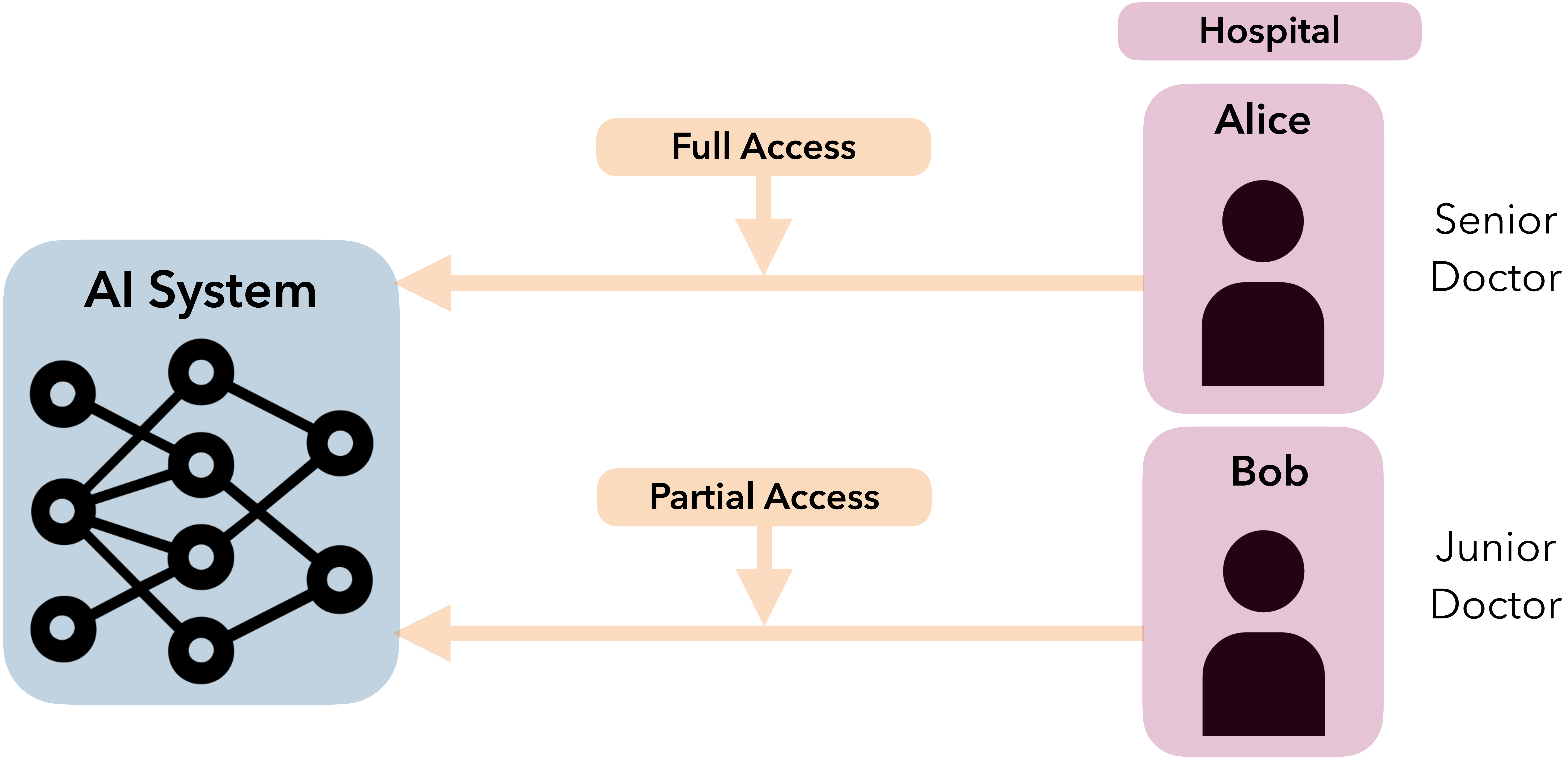
Effective Human-AI Collaboration



B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. AAAI. 2025.

Effective Human-AI Collaboration

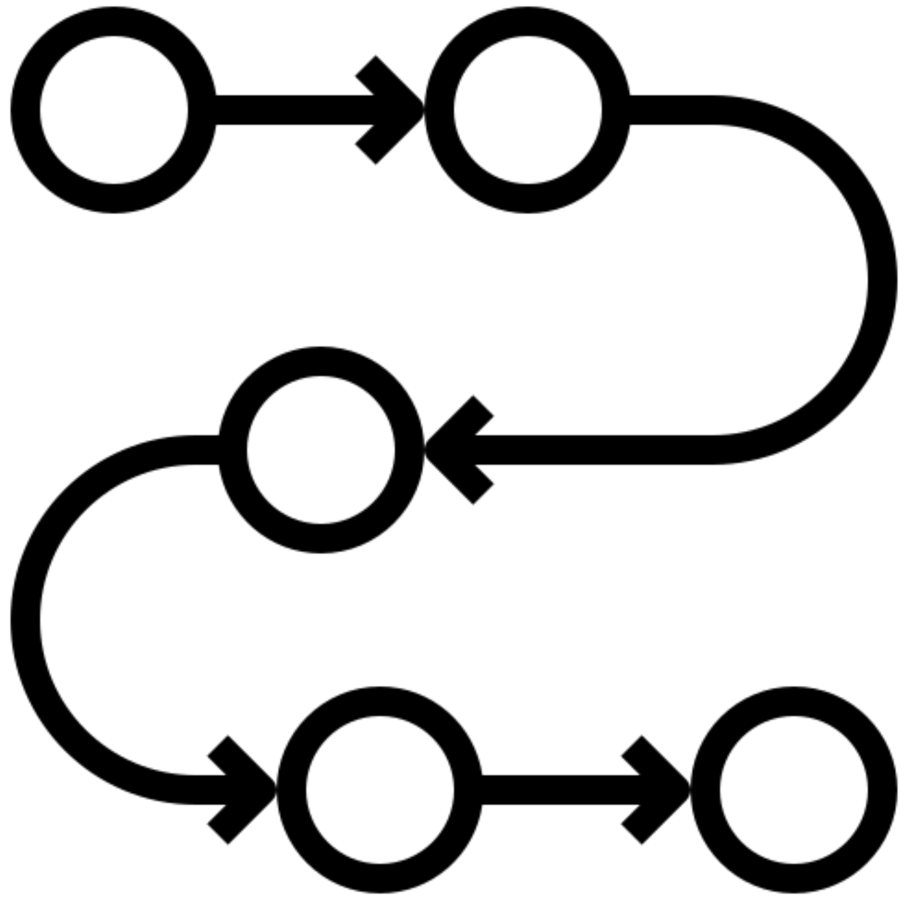


B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.

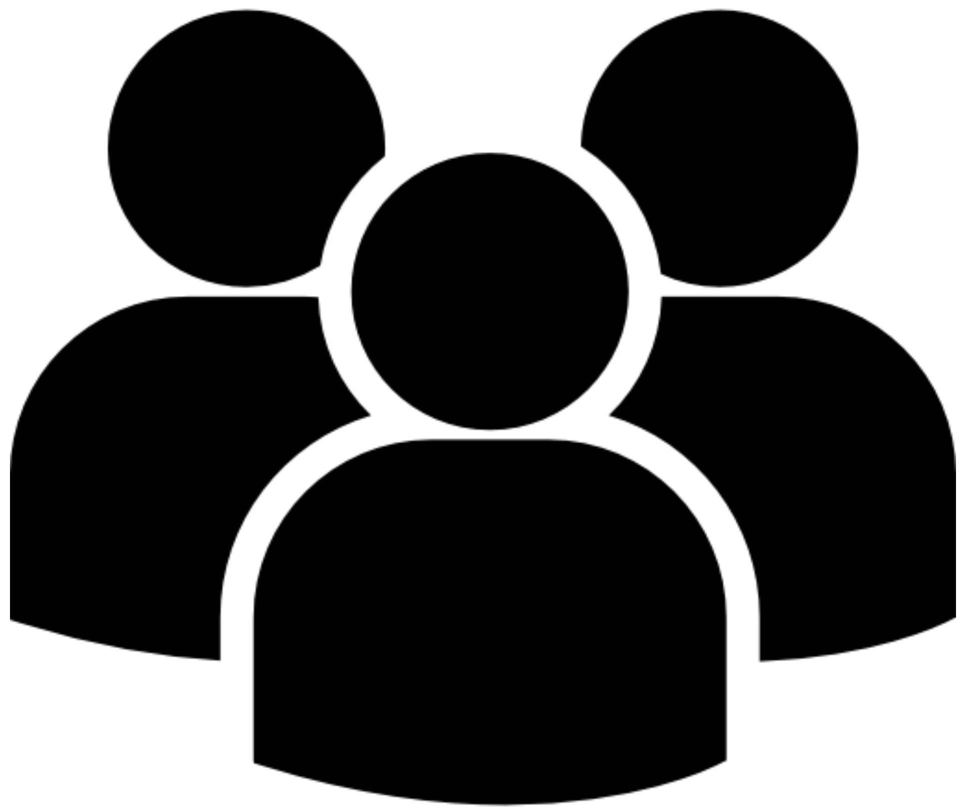
B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. AAAI. 2025.

Effective Human-AI Collaboration

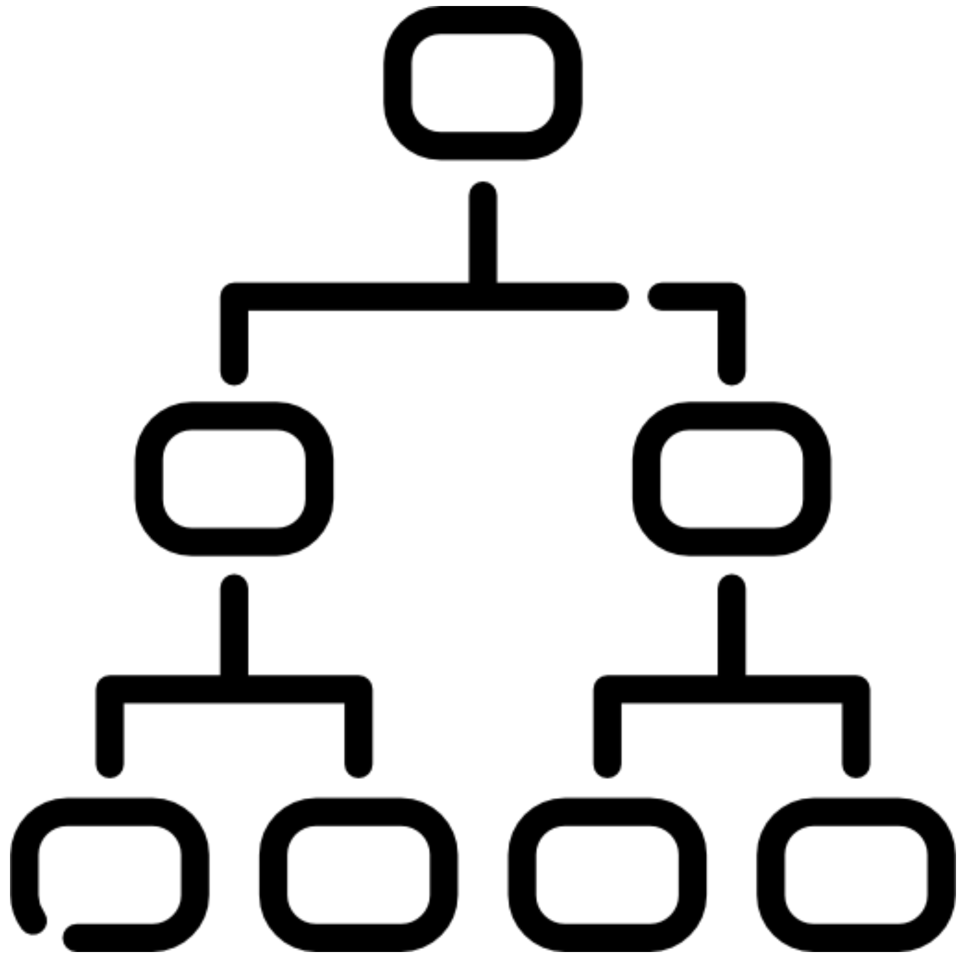
Methods



Online Learning



Learning from Prior Data



Rule-Based

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies.* AAAI. 2025.

Collins, Chen, Sucholutsky, Kirk, Sadek, Sargeant, Talwalkar, Weller, **B**. *Modulating Language Model Experiences through Frictions*. Under Review. 2024.

Effective Human-AI Collaboration

User Study

*Under what conditions is **selective** access to AI assistance helpful?*

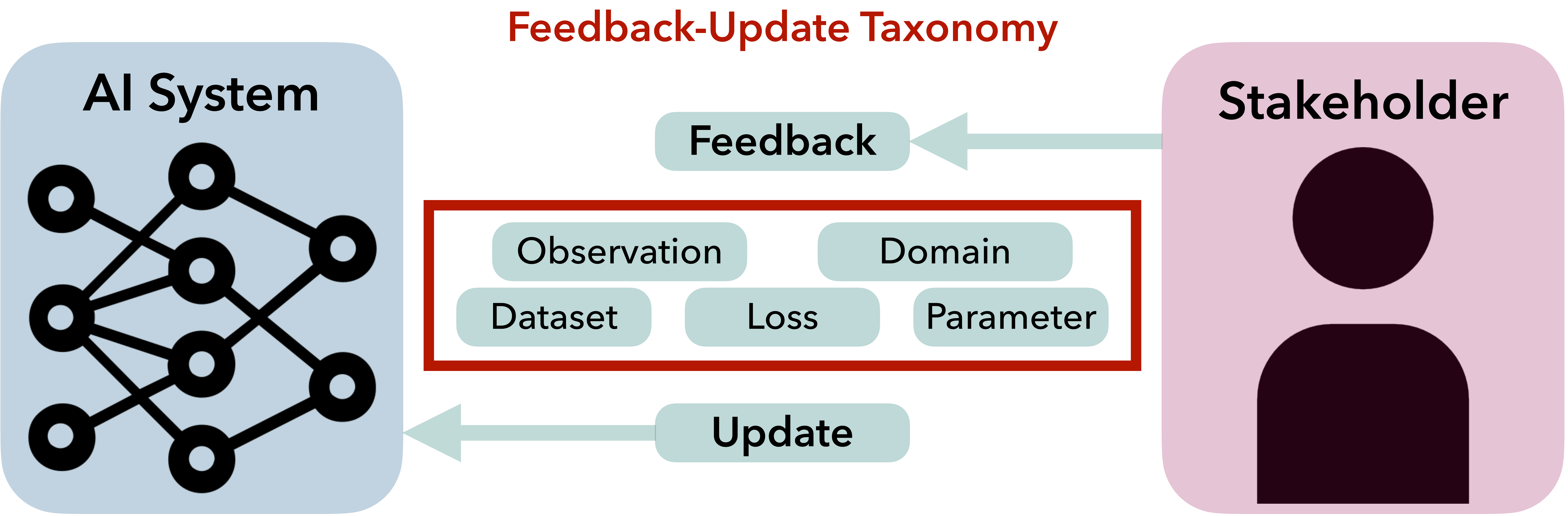
Foul detection with soccer referees

Visual pollution detection with city inspectors

Mortality prediction with cardiologists

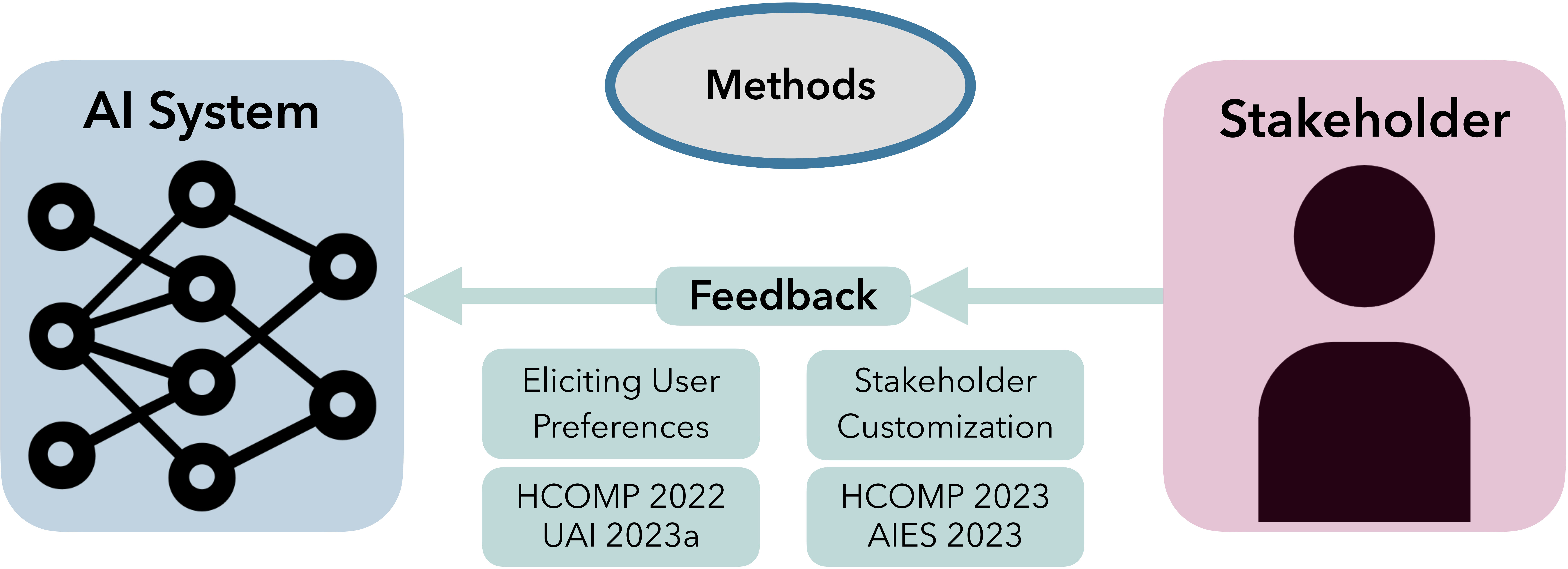
B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer. 2024.
B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies.* AAAI. 2025.
Collins, Chen, Sucholutsky, Kirk, Sadek, Sargeant, Talwalkar, Weller, **B**. *Modulating Language Model Experiences through Frictions*. Under Review. 2024.

Effective Human-AI Collaboration



Hertwig, Erev. *The description–experience gap in risky choice*. Trends in Cognitive Science. 2009.
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. Patterns. 2023.

Effective Human-AI Collaboration



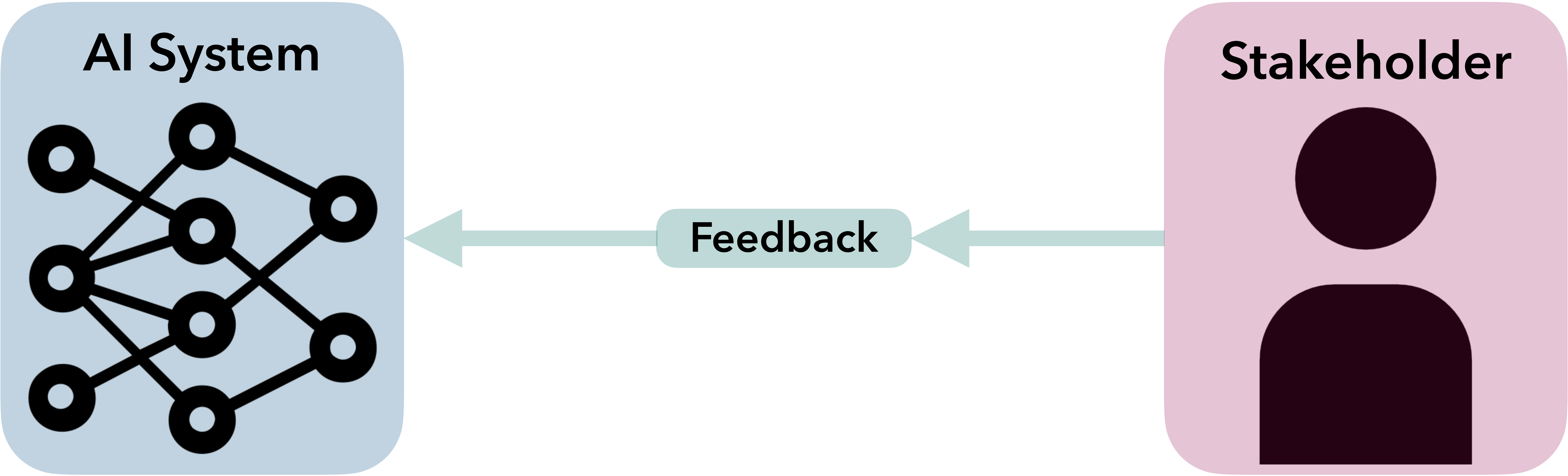
Collins*, **B***, Weller. *Eliciting and Learning with Soft Labels from Every Annotator*. AAAI HCOMP. 2022.

Collins, **B**, Liu, Piratla, Sucholutsky, Love, Weller. *Human-in-the-Loop mixUp*. UAI. 2023.

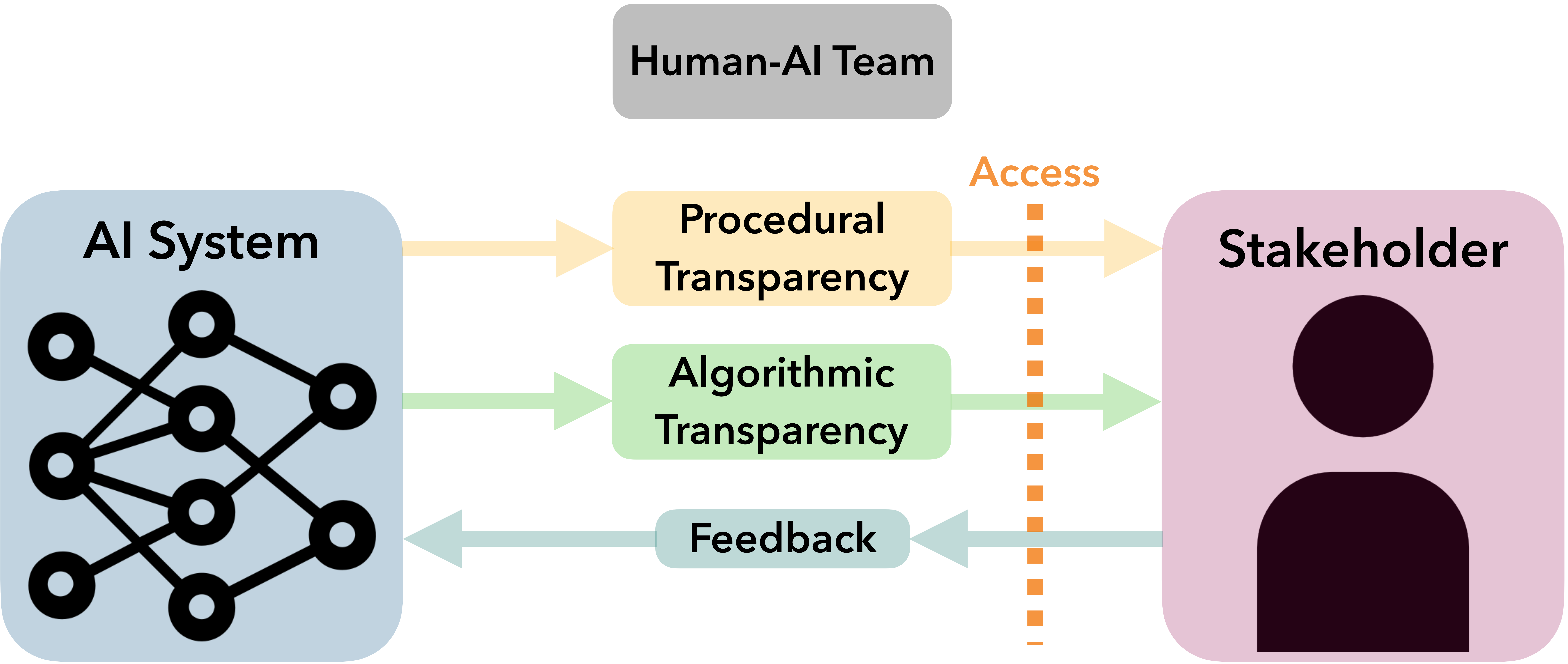
Collins, Barker, Espinosa, Raman, **B**, Jamnik, Sucholutsky, Weller, Dvijotham. *Human Uncertainty in Concept-Based AI Systems*. ACM AIES. 2023.

Barker, Collins, Dvijotham, Weller, **B**. *Selective Concept Models: Permitting Stakeholder Customization at Test-Time*. AAAI HCOMP. 2023.

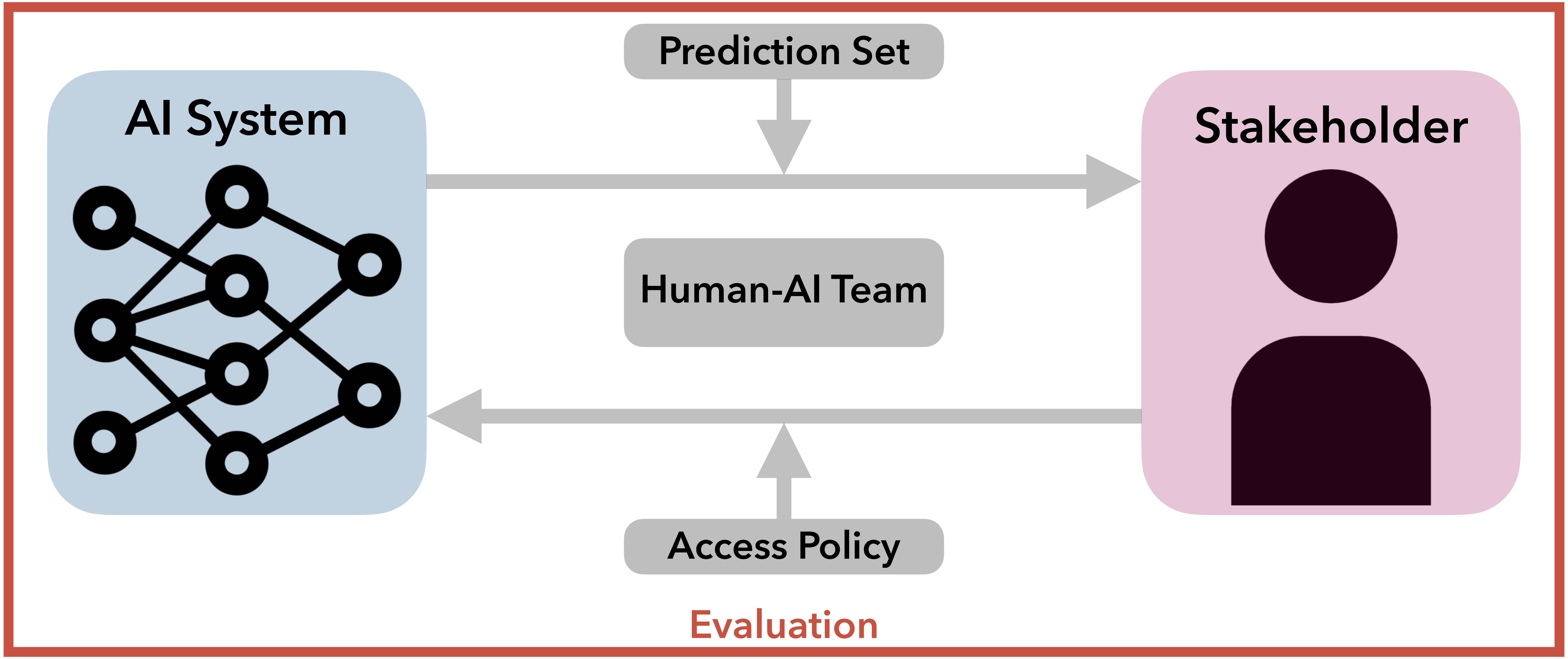
Effective Human-AI Collaboration



*How do feedback mechanisms **vary** across cultures and contexts?*

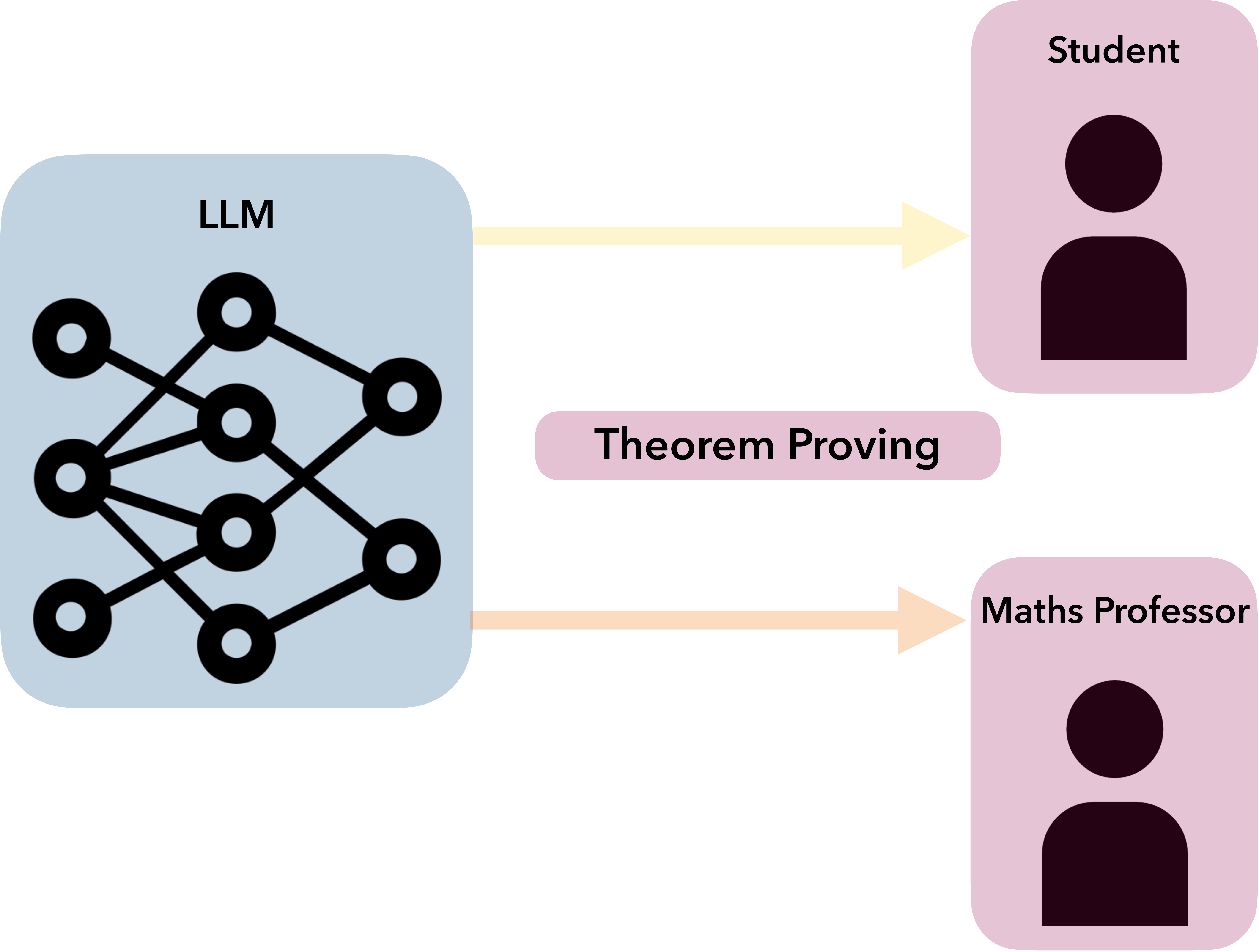

Human-AI Team
Access
AI System
Procedural Transparency
Algorithmic Transparency
Feedback
Stakeholder

Interactive Human-Centered Evaluation



Babbar, B, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.

Interactive Human-Centered Evaluation

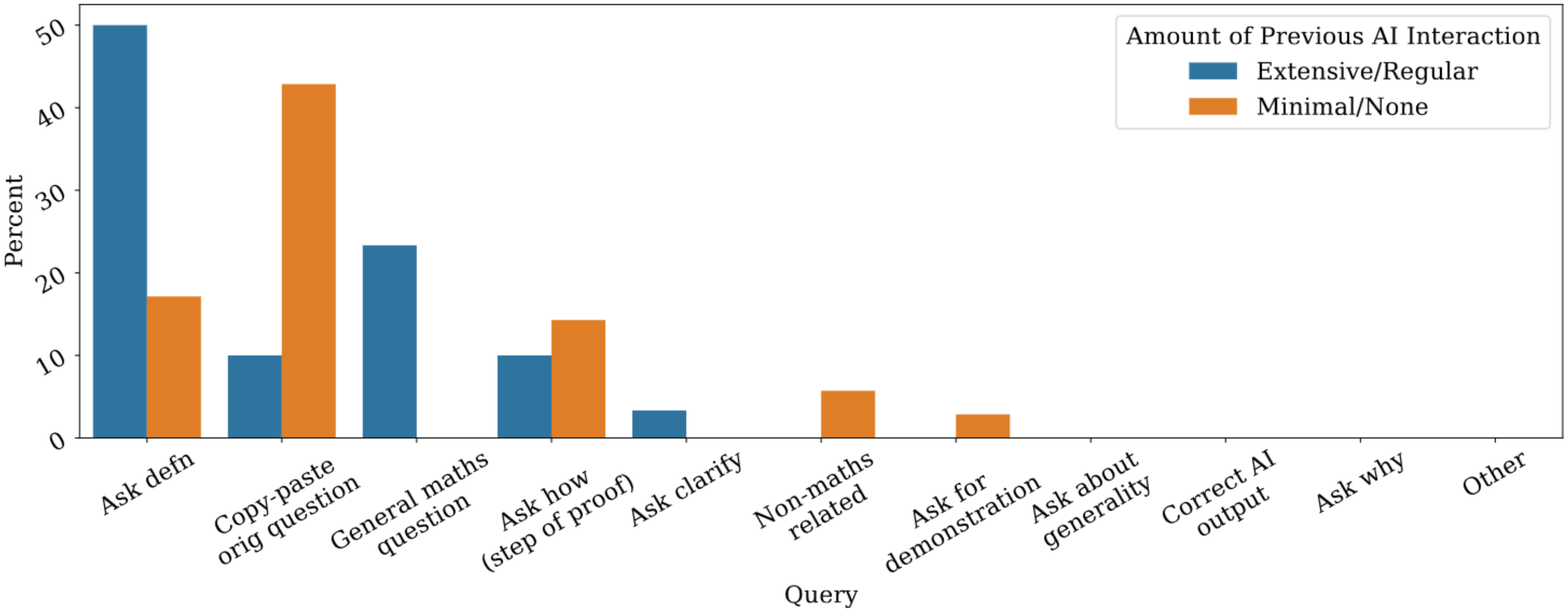


User Studies

1. Observing usage patterns teases out differences between perceived helpfulness and correctness
2. Unconfident participants rated incorrect LLM responses as correct
3. Interactive evaluation of LLM outputs is key

Collins, Jiang, Frider, Wong, Zilka, **B**, Lukasiewicz, Wu, Tenenbaum, Hart, Gowers, Li, Weller, Jamnik. *When Should Algorithms Evaluating language models for mathematics through interactions*. PNAS. 2024.

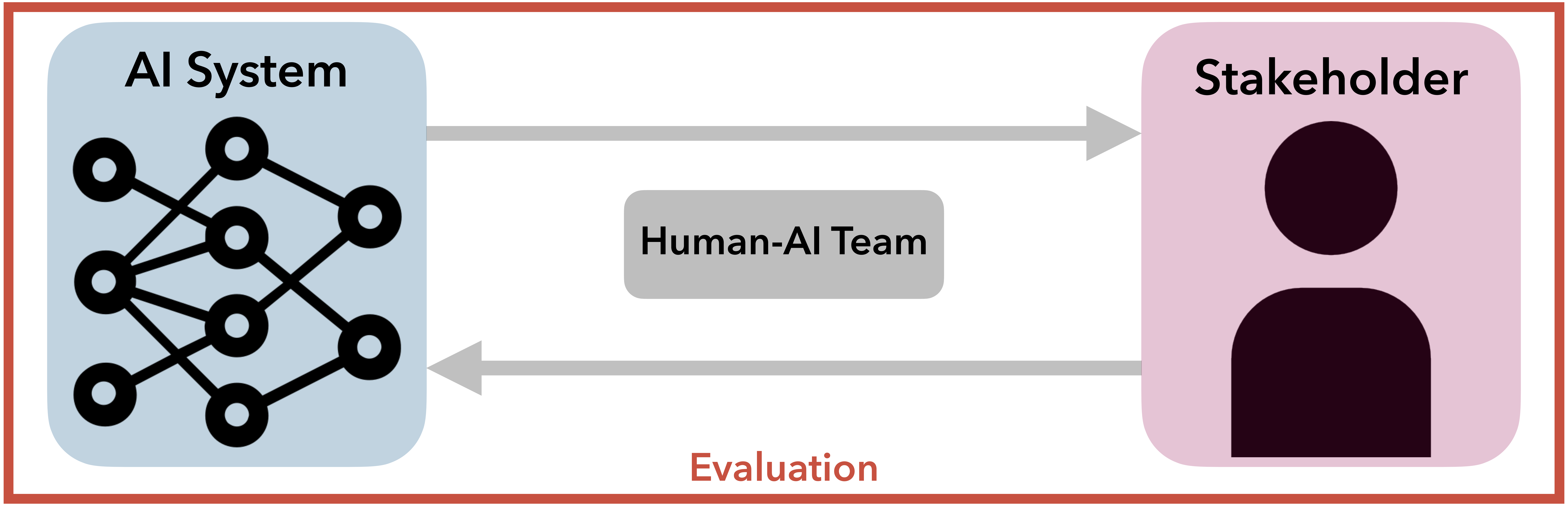
Interactive Human-Centered Evaluation



Regular users of LLMs ask for definitions rather than the query itself

Collins, Jiang, Frider, Wong, Zilka, **B**, Lukasiewicz, Wu, Tenenbaum, Hart, Gowers, Li, Weller, Jamnik. *When Should Algorithms Evaluating language models for mathematics through interactions*. PNAS. 2024.

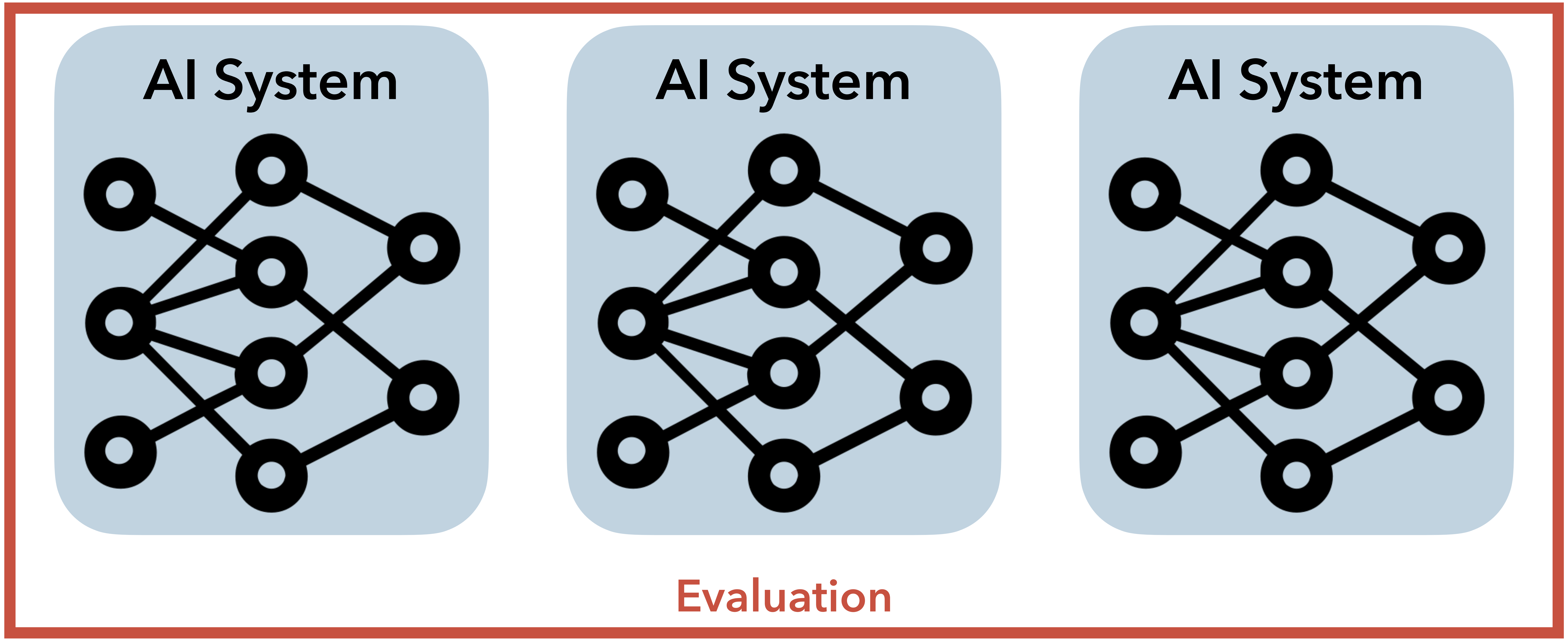
Interactive Human-Centered Evaluation



What would interactive evaluation of LLMs look like for ***humanities****, such as interpreting poetry or critiquing art?*

Collins, Jiang, Frider, Wong, Zilka, **B**, Lukasiewicz, Wu, Tenenbaum, Hart, Gowers, Li, Weller, Jamnik. *When Should Algorithms Evaluating language models for mathematics through interactions*. PNAS. 2024.

Interactive Human-Centered Evaluation

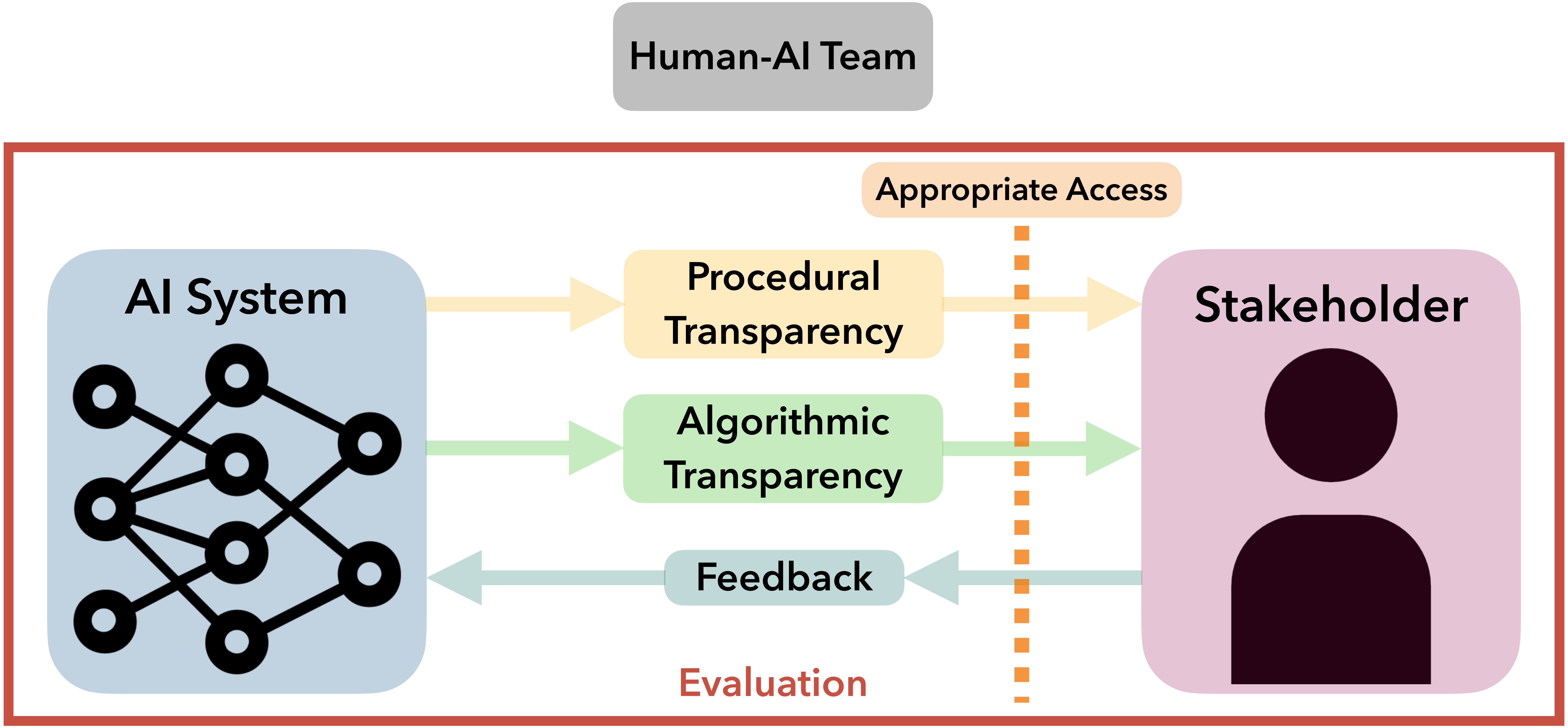


accenture

slalom

AISI AI SAFETY INSTITUTE

How can we catalog how AI systems are deployed to understand their design, governance, and impact in practice?


Human-AI Team
Appropriate Access
AI System
Procedural Transparency
Algorithmic Transparency
Feedback
Stakeholder
Evaluation

Why CHIA?

My research spans multiple disciplines and various research CHIA programmes, including Responsible AI, Social/Interactive AI, and Cognitive AI

Empowering MPhil and PhD students to build **and** deploy AI inspired by their communities is important: practical coursework and rigorous a research

After spending time at Carnegie Mellon, NYU, and Harvard, I find the Cambridge ecosystem unmatched – I want to help CHIA establish itself as a powerhouse for practical human-AI interaction research

Computer Science & Engineering

Isabel Chien — Cambridge
J.M.H Lobato — Cambridge
Mateja Jamnik — Cambridge
Javier Antorán — Cambridge
Katie Collins — Cambridge
Adrian Weller — Cambridge
José Moura — CMU
Valerie Chen — CMU
Ameet Talwalkar — CMU
Hoda Heidari — CMU
Joydeep Ghosh — UT Austin
Shubham Sharma — UT Austin

Riccardo Fogliato — Amazon
Peter Eckersley — PAI
Lama Nachman — Intel
P. Kamalaruban — Turing
Varun Babbar — Duke
Matthew Barker — Trustwise
Dan Ley — Harvard
M. Bilal Zafar — Bochum
Ruchir Puri — IBM
Yunfeng Zhang — Twitter
Vera Liao — Microsoft
Ankur Taly — Google

Elaf Alamhmoud — NYU
Andrew Wilson — NYU
Sanyam Kapoor — NYU
Ilia Sucholutsky — NYU
Albert Jiang — Mistral
Hannah Kirk — Oxford

Psych & CogSci

Bradley Love — UCL
Simone Schnall — Cambridge
Brenden Lake — NYU

Josh Tenenbaum — MIT
Tom Griffiths — Princeton
Guy Davidson — NYU

Design

Kendall Brogle — Turing
Emma Kallina — Cambridge
Becca Ricks — Mozilla
Dorian Peters — Imperial
Malak Sadek — Imperial

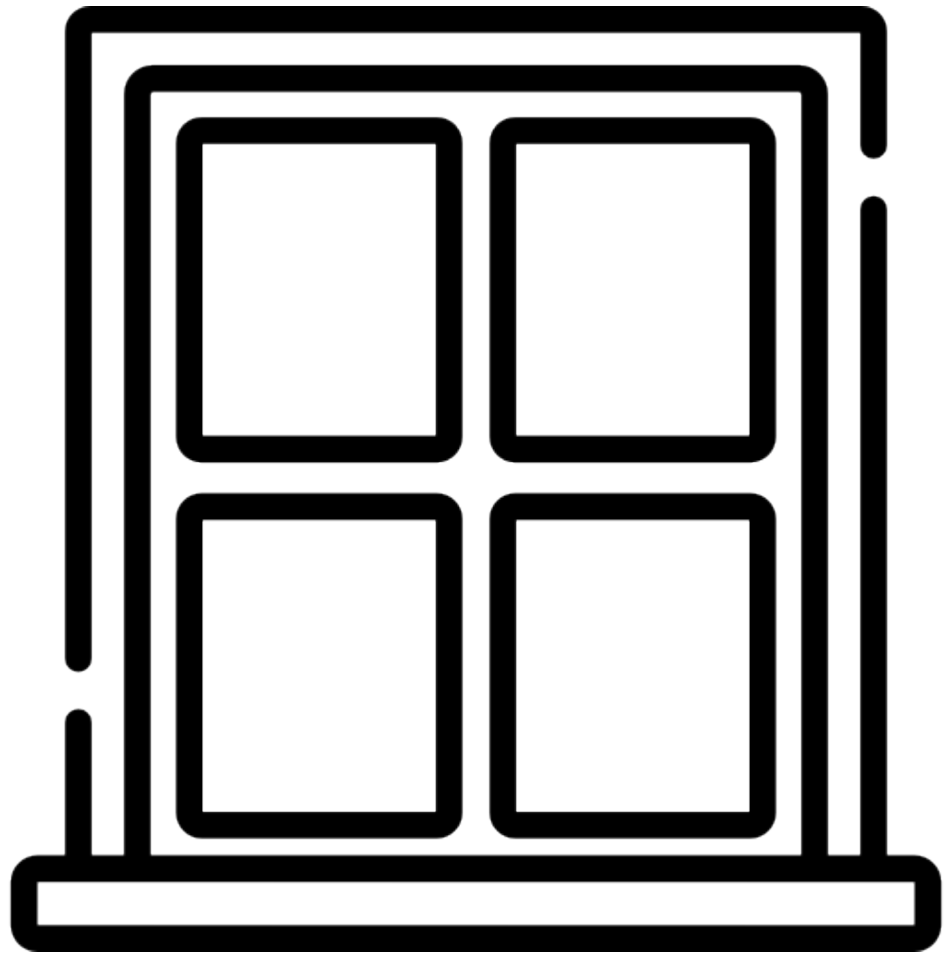
Policy & Law

John Zerilli — Edinburgh
Alice Xiang — Sony AI
Madhu Srikumar — PAI

Holli Sargeant — Cambridge
Karen Yeung — Birmingham
Rotem Medzini — Birmingham

Trustworthy Machine Learning

Transparency, Collaboration, and Evaluation

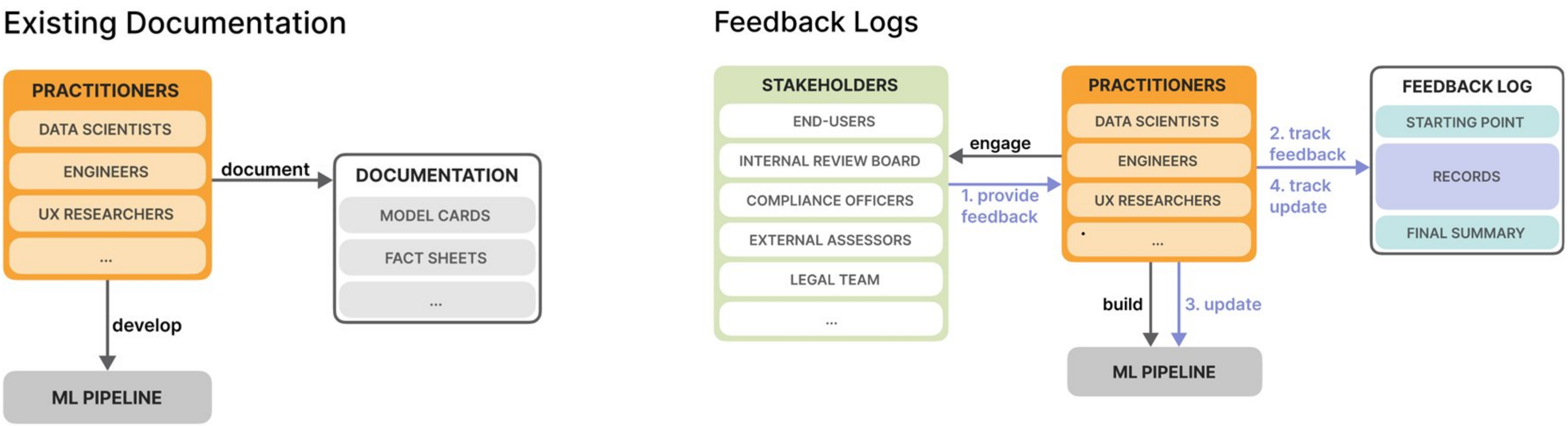


@umangsbhatt

umangbhatt@nyu.edu

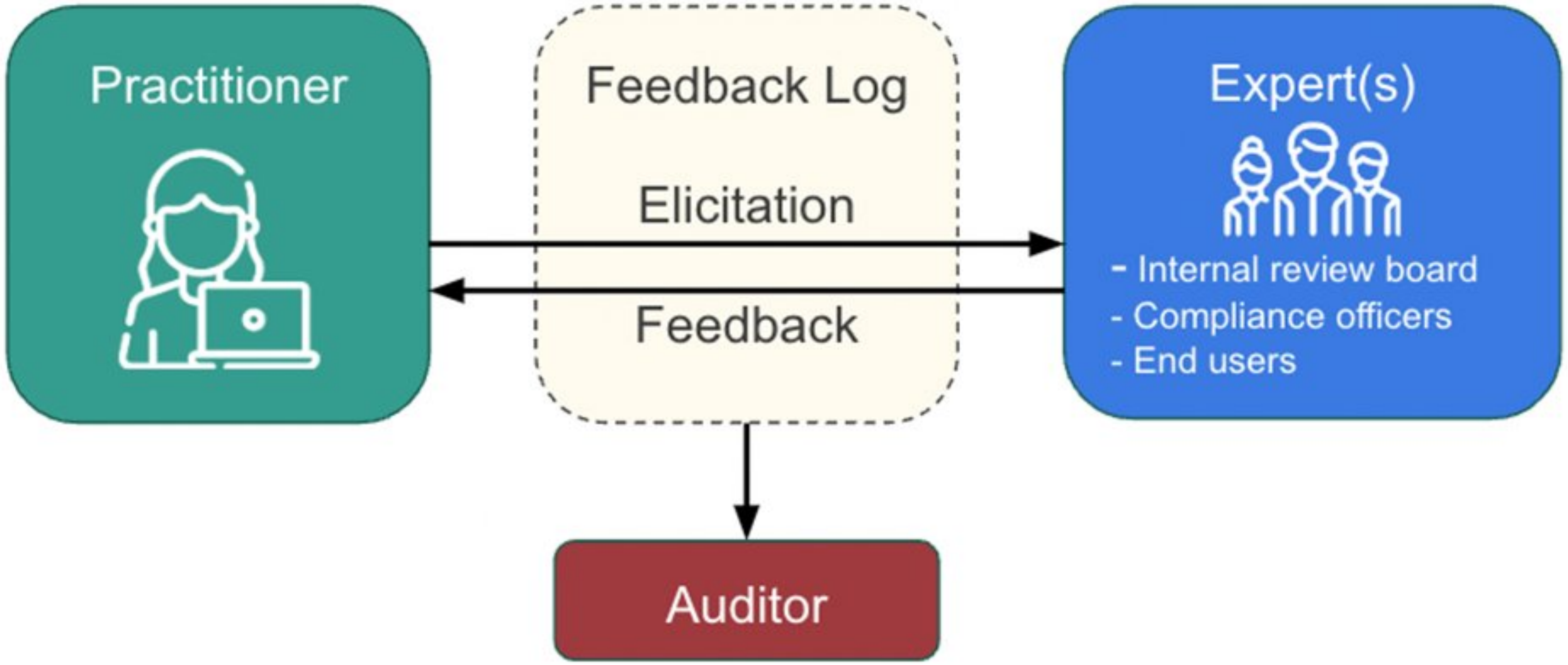
Appendix

FeedbackLogs



Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback*. ACM EAAMO. 2023.

FeedbackLogs



Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback*. ACM EAAMO. 2023.

FeedbackLogs

Starting Point
Data: *Description of the dataset(s) used to train/test/validate the model.*
Models: *Description of the model(s) used and any existing design decisions.*
Metrics: *Description of the metrics used to evaluate the model(s) and their performance.*

Record 1

Elicitation
Who and why? *Which stakeholder(s) are being consulted? What prompted the request for feedback? e.g. legal requirements, poor performance on metrics.*
How? *How is the relevant information presented to them? e.g. model metrics, predictions, prototype.*

Feedback
What? *What insights have been provided by the stakeholder(s)?*

Incorporation

Which?	Where?	When?	Why?	Effect?
Which updates are considered?	*Where in the pipeline did the update occur?*	*When in the pipeline did the update occur?*	*Why has this update been selected?*	*What effect(s) did the update have on the metrics?*
Update 1	x	x	x	x
Update 2	x	x	x	x
...	...	...	...	...

Summary
What? *Summary of the update(s) chosen and their effect(s) on the metric(s).*

Record 2

...

Final Summary
Data: *Description of the dataset(s) used to train/test/validate the model after all updates have been applied.*
Model: *Description of model(s) used and any design changes resulting from the updates.*
Metric performance: *Description of the metrics to evaluate the model(s) and their performance after the above updates.*

Barker, Kallina, Ashok, Collins, Casovan, Weller, Talwalkar, Chen, **B**. *FeedbackLogs: Recording and Incorporating Stakeholder Feedback*. ACM EAAMO. 2023.

Data Scientist

Explanation Evaluation

IJCAI 2020
AAAI 2021

Assess properties of explanations

Methods

Model $f : \mathcal{X} \mapsto \mathcal{Y}$

Explanation Function $g : \mathcal{F} \times \mathcal{X} \mapsto \mathbb{R}$

Problem: *"There are many of candidate explanation methods (LIME, SHAP, etc.) but it is unclear how to decide when to use each."*

Candidate Properties

Sensitivity: Do similar inputs have similar explanations?

$$\mu(f, g, x, r) = \int_{\rho(x,z) \leq r} D(g(f,x), g(f,z))\mathbb{P}_x(z)dz$$

Faithfulness: Does the explanation capture features important for prediction?

$$\mu(f, g, x, S) = \text{corr}\left(\frac{1}{|S|}\sum_{i \in S} g(f,x)_i, f(x) - f(x_{[x_s = \bar{x}_s]})\right)$$

Complexity: Is the explanation digestible?

$$\mu(f, g, x) = H(x) = \mathbb{E}_i\left[-\ln(|g(f,x)_i|)\right]$$

We go on to show how to (A) aggregate multiple explanations into a consensus and (B) how to optimize an explanation for a selected criterion

B, Moura, Weller. *Evaluating and Aggregating Feature-based Model Explanations.* IJCAI. 2020.

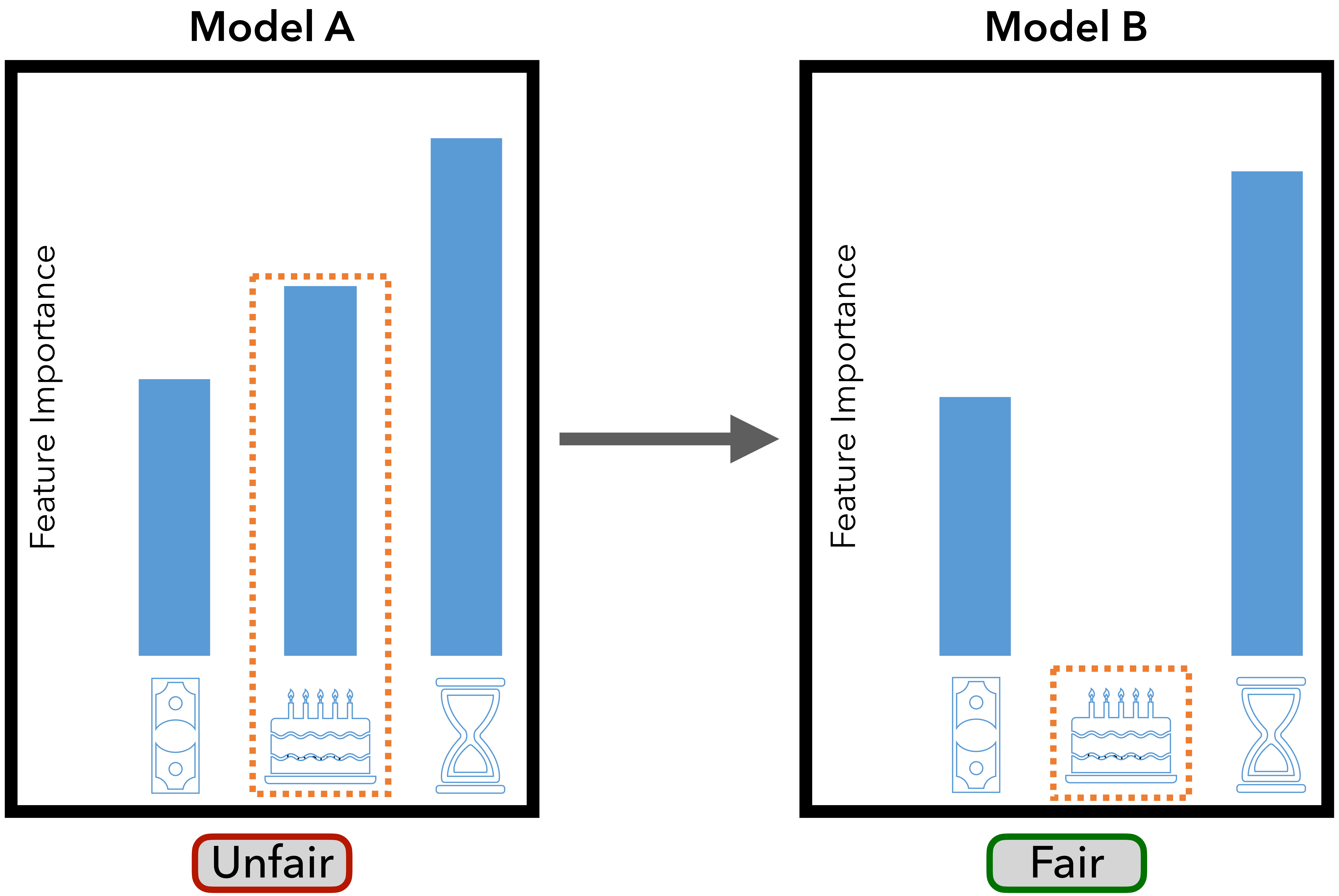
Policy Maker

Explanations of Unfairness

ECAI 2020
AAAI 2022a

Assure model fairness via explanations

Methods



Dimanov, **B**, Jamnik, Weller. *You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods*. ECAI. 2020.

Policy Maker

Explanations of Unfairness

ECAI 2020
AAAI 2022a

Don't assess model fairness via explanations

Methods

Attribution of Sensitive Attribute $g(f, x)_j$

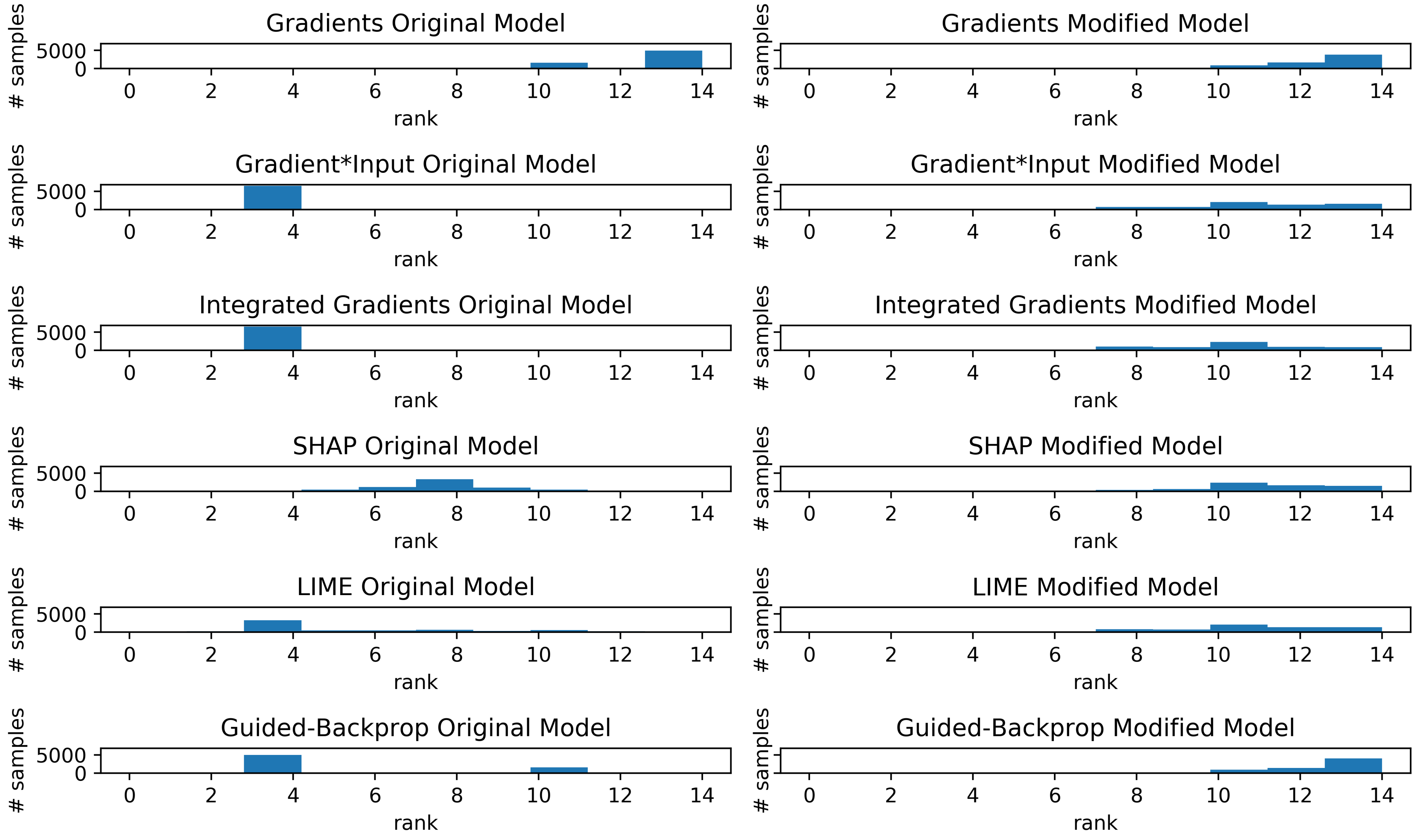
Our Goal $f_\theta \rightarrow f_{\theta+\delta}$

1. Model Similarity $\forall i,\ f_{\theta+\delta}(\mathbf{x}^{(i)}) \approx f_\theta(\mathbf{x}^{(i)})$

2. Low Target Attribution $\forall i,\ |g(f_{\theta+\delta}, \mathbf{x}^{(i)})_j| \ll |g(f_\theta, \mathbf{x}^{(i)})_j|$

Adversarial Explanation Attack

$$\text{argmin}_\delta\ \ L' = L(f_{\theta+\delta}, x, y) + \frac{\alpha}{n} \left|\left| \nabla_{\mathbf{X}_{:,j}} L(f_{\theta+\delta}, x, y) \right|\right|_p$$



Our proposed attack:
1. Decreases relative importance significantly.
2. Generalizes to test points.
3. Transfers across explanation methods.

Heo, Joo, Moon. *Fooling Neural Network interpretations via adversarial model manipulation*. NeurIPS. 2019.
Dimanov, **B**, Jamnik, Weller. *You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods*. ECAI. 2020.

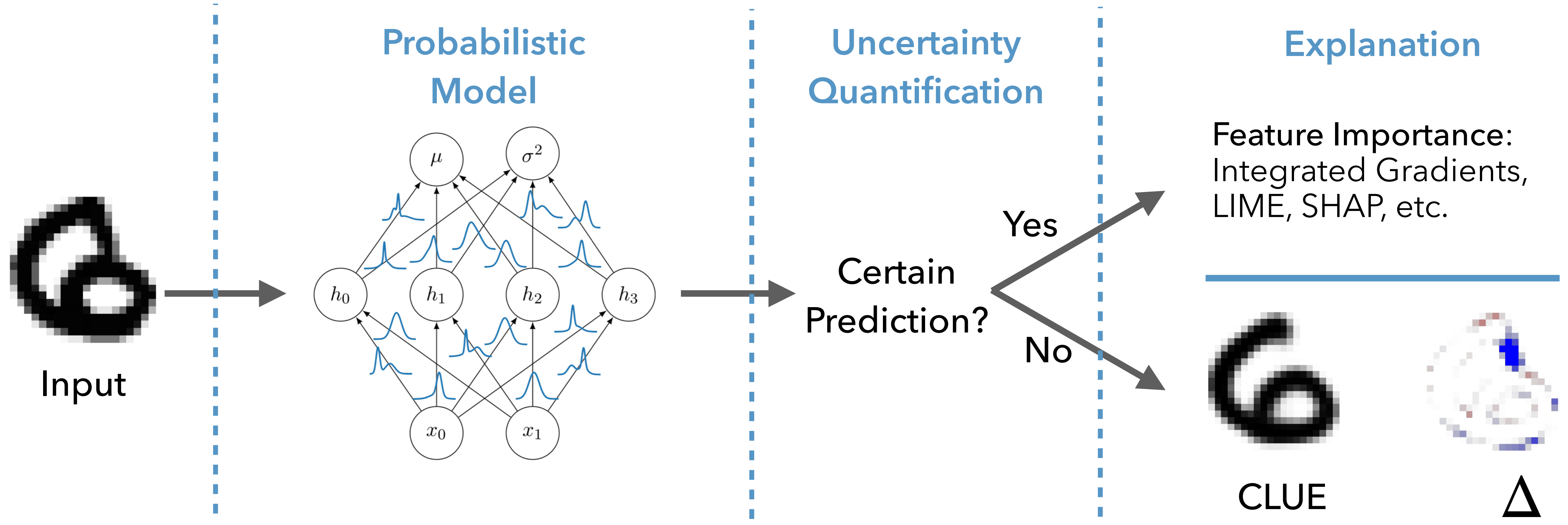
Risk Executive

Explanations of Uncertainty

CLUE: Counterfactual Latent Uncertainty Explanations

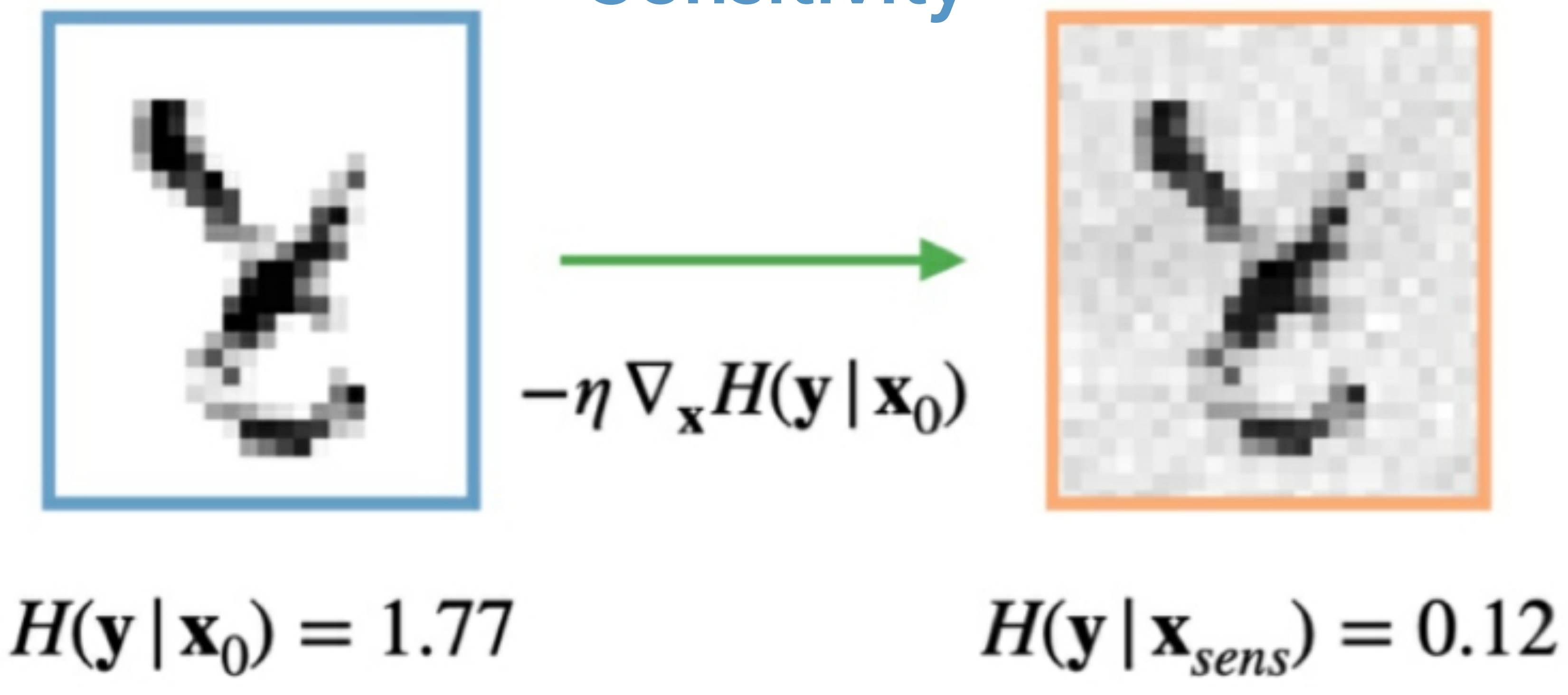
Methods

Question: *"Where in my input does uncertainty about my outcome lie?"*

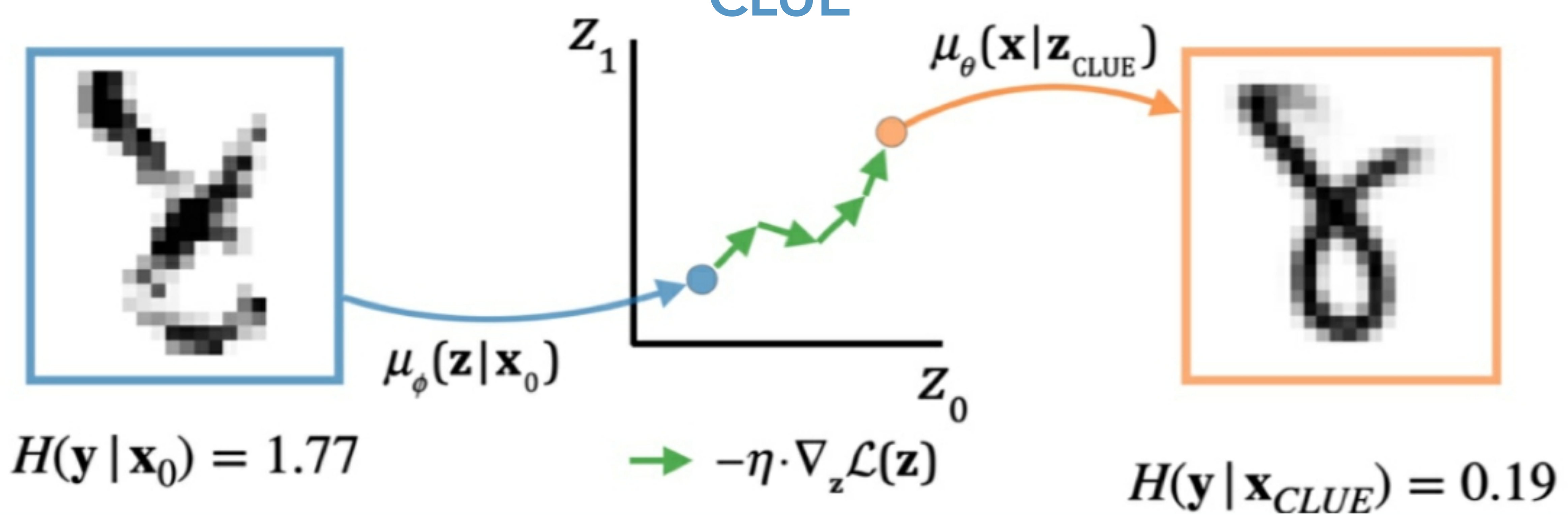


Formulation: *What is the smallest change we need to make to an input, while staying in-distribution, such that our model produces more certain predictions?*

Sensitivity



CLUE



Antoran, **B**, Adel, Weller, Hernandez-Lobato. *Getting a CLUE: A Method for Explaining Uncertainty Estimates.* ICLR. 2021.

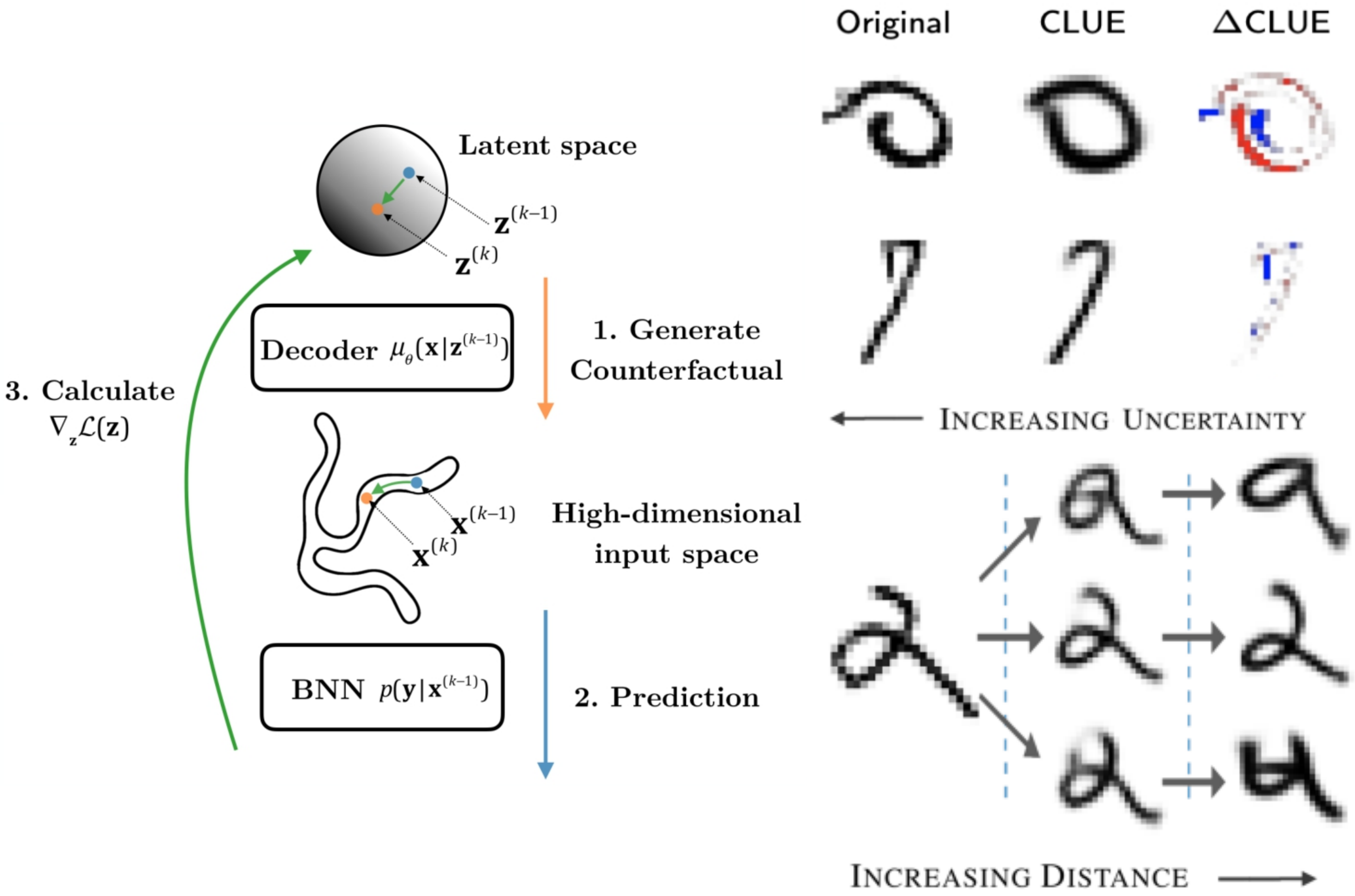
Ley, **B**, Weller. *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates.* AAAI. 2022.

Risk Executive

Explanations of Uncertainty

CLUE: Counterfactual Latent Uncertainty Explanations

Methods

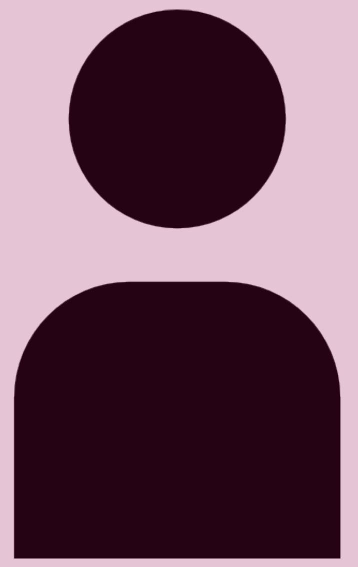


Antoran, **B**, Adel, Weller, Hernandez-Lobato. *Getting a CLUE: A Method for Explaining Uncertainty Estimates.* ICLR. 2021.

Ley, **B**, Weller. *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates.* AAAI. 2022.

CLUE: Counterfactual Latent Uncertainty Explanations

User Studies



Risk Executive

Explanations of Uncertainty

Forward Simulation: Users are shown context examples and are tasked with predicting model behavior on new datapoint.

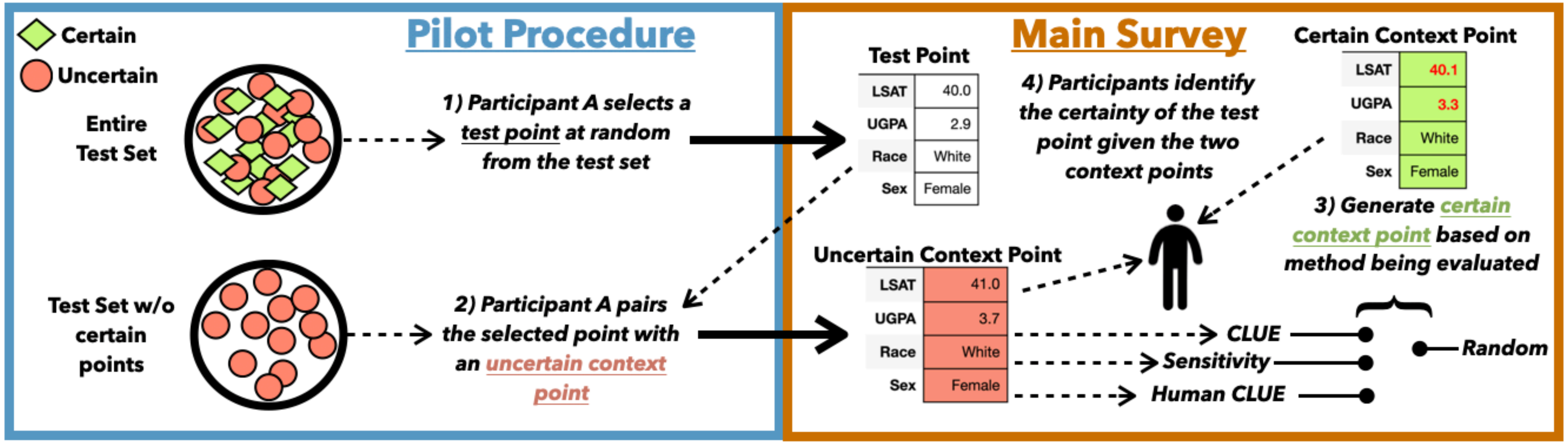
	Uncertain
Age	Less than 25
Race	Caucasian
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	Yes
Prior Convictions	1
Days Served	0

	Certain
Age	Less than 25
Race	African-American
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	No
Prior Convictions	0
Days Served	0

	?
Age	Less than 25
Race	Hispanic
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	No
Prior Convictions	0
Days Served	0

	Combined	LSAT	COMPAS
CLUE	**82.22**	**83.33**	**81.11**
Human CLUE	62.22	61.11	63.33
Random	61.67	62.22	61.11
Local Sensitivity	52.78	56.67	48.89

CLUE outperforms other approaches with statistical significance.
(Using Nemenyi test for average ranks across test questions)



Antoran, **B**, Adel, Weller, Hernandez-Lobato. *Getting a CLUE: A Method for Explaining Uncertainty Estimates.* ICLR. 2021.
Ley, **B**, Weller. *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates.* AAAI. 2022.

Radiologist

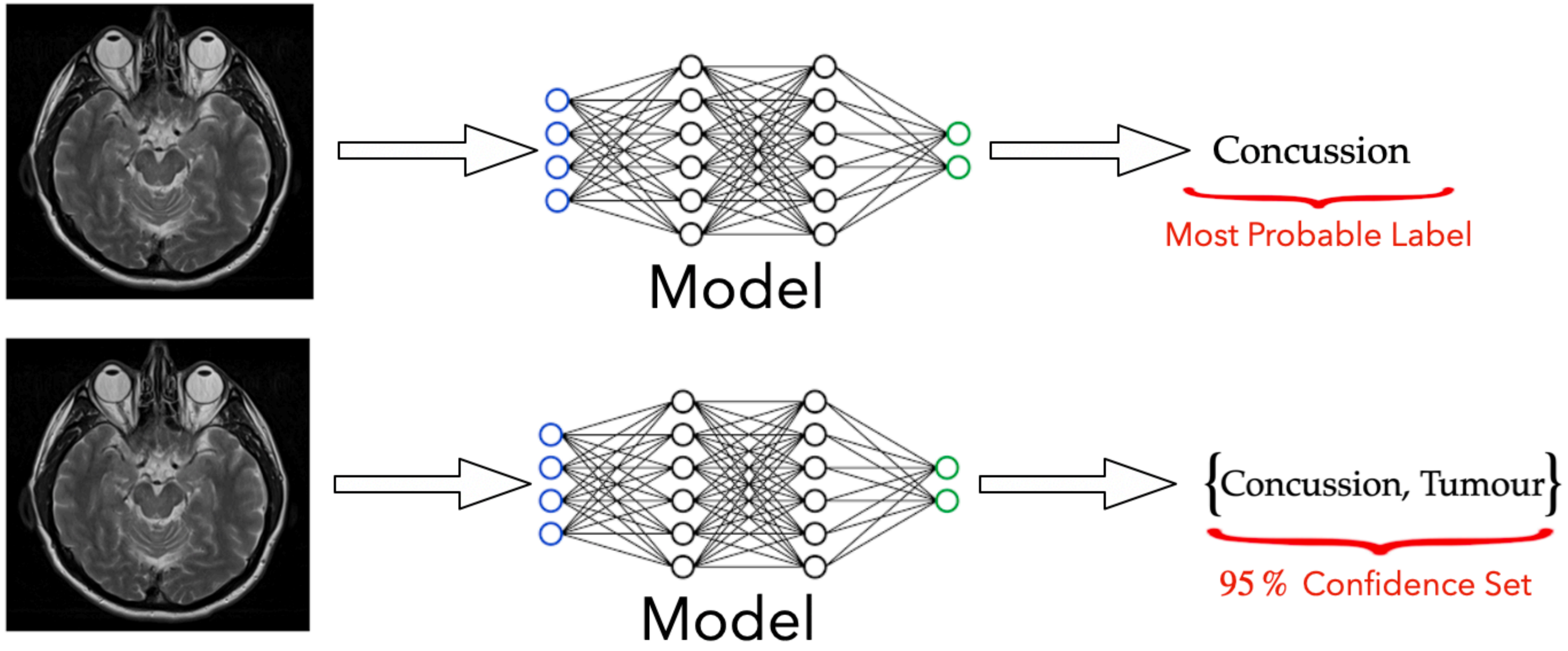
Prediction Sets

IJCAI 2022

Generate prediction sets for experts

Methods

Question: *"What other outcomes are probable?"*



Top-1 Classifier

Set Valued Classifier

Prediction Set $\Gamma(x) = \{y \in \mathcal{Y} \,|\, P(y\,|\,x) \geq \tau\}$

Conformal Prediction $FNR \leq \alpha \equiv P\big(y \notin \Gamma(x)\big) \leq \alpha$

Risk Controlling Prediction Sets $P\Big(\underbrace{\mathbb{E}[L(y, \Gamma(x))]}_{\text{Risk}} \leq \alpha\Big) \geq 1 - \delta$

Vovk, Gammerman, Shafer. Algorithms in the Real World. 2005

Bates, Angelopoulos, Lei, Malik, Jordan. *Distribution-Free, Risk-Controlling Prediction Sets*. Journal of the ACM. 202.

Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.

Generate prediction sets for experts

User Studies

Radiologist

Prediction Sets

IJCAI 2022

Question: *Do prediction sets improve human-machine team performance?*

A CP Scheme!

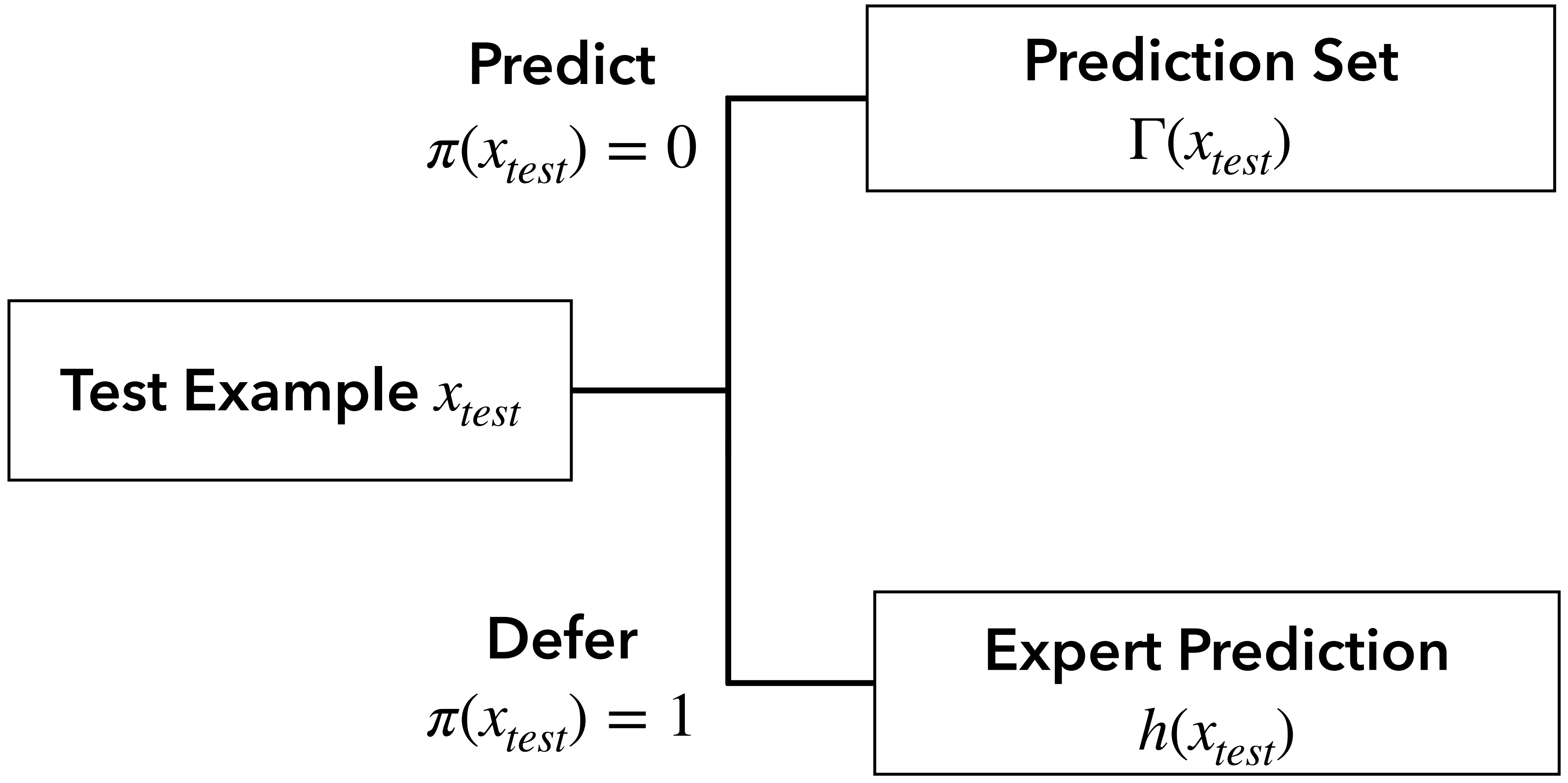
For CIFAR-100:

1. Prediction sets are perceived to be more useful ✓
2. Users trust prediction sets more than Top-1 classifiers ✓

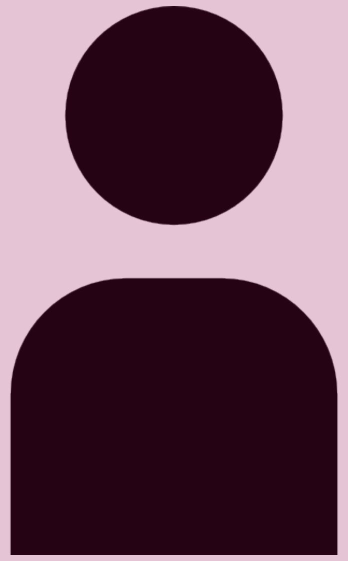
Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.05	0.76 ± 0.05	0.999	0.000
Reported Utility	5.43 ± 0.69	6.94 ± 0.69	**0.003**	1.160
Reported Confidence	7.21 ± 0.55	7.88 ± 0.29	0.082	0.674
Reported Trust in Model	5.87 ± 0.81	8.00 ± 0.69	$<$ **0.001**	1.487

Observation: *Some prediction sets can be quite large, rendering them useless to experts!*

Idea: *Learn a deferral policy $\pi(x) \in \{0,1\}$ and reduce prediction set size on remaining examples*



Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams.* IJCAI. 2022.



Radiologist

Prediction Sets

IJCAI 2022

Generate prediction sets for experts

User Studies

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.08	0.67 ± 0.05	**0.003**	0.832
Reported Utility	7.93 ± 0.39	6.32 ± 0.60	$<$ **0.001**	1.138
Reported Confidence	7.31 ± 0.29	7.28 ± 0.29	0.862	0.046
Reported Trust in Model	8.00 ± 0.45	6.87 ± 0.61	**0.006**	0.754

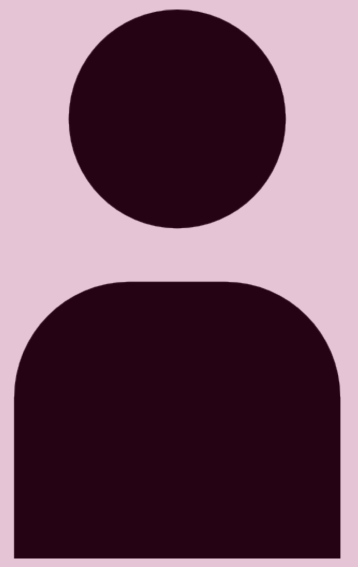
Using our deferral plus prediction set scheme, we achieve:

1. Higher perceived utility ✓
2. Higher reported trust ✓
3. Higher team accuracy ✓

We also (A) prove that set size is reduced for the non-deferred examples and (B) optimize for additional set properties (e.g., sets with similar labels).

Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.

Student



Personalize Access

Learning Personalized Decision Support Policies

Methods

Question: *"When is it appropriate to provide decision support (e.g. ML model predictions) to a specific decision-maker?"*

Forms of support | **Decision-maker**

x_t

a_1 = **None**

a_2 = **ML Prediction**

$\pi_t^{Alice}(x_t) = a_2$

$\tilde{y}_t = h_{Alice}(x_t, a_2)$

Update π_{t+1}^{Alice} using $\ell(\tilde{y}_t, y_t)$

Formulation: *For an unseen decision-maker, which available form of decision support would improve their decision outcome performance the most?*

Set Up

We select a form of support $a_t \in A$ using a decision support policy $\pi_t : X \rightarrow \Delta(A)$

The decision-maker makes the final prediction: $\tilde{y}_t = h(x_t, a_t)$

Performance differs under each form of support: $r_{A_i}(x; h) = \mathbb{E}_{y|x}[\ell(y, h(x, A_i))]$

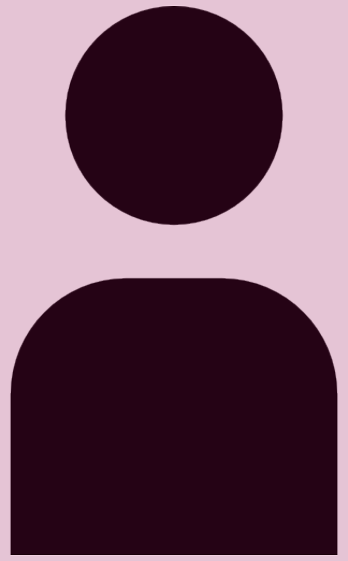
Core Idea of THREAD

Learn policy π_t using a exisiting contextual bandits techniques

Include cost of a_t in the objective

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

Student



Personalize Access

Learning Personalized Decision Support Policies

Methods

MMLU Task: *60 questions from 4 categories*

Computer Science, Elementary Math, Biology, Foreign Policy

Expertise Profiles

1. Invariant: $r_{A_1}(X_j; h) \approx r_{A_2}(X_j; h), \forall j \in [N]$

   equally good (or bad) with or without LLM support

2. Varying: $r_{A_1}(X_j; h) \leq r_{A_2}(X_j; h)$ and $r_{A_2}(X_k; h) \leq r_{A_1}(X_k; h)$

   better for some topics with LLM support

3. Strictly Better: $r_{A_1}(X_j; h) \leq r_{A_2}(X_j; h), \forall j \in [N]$

   strictly better with (or without) LLM support

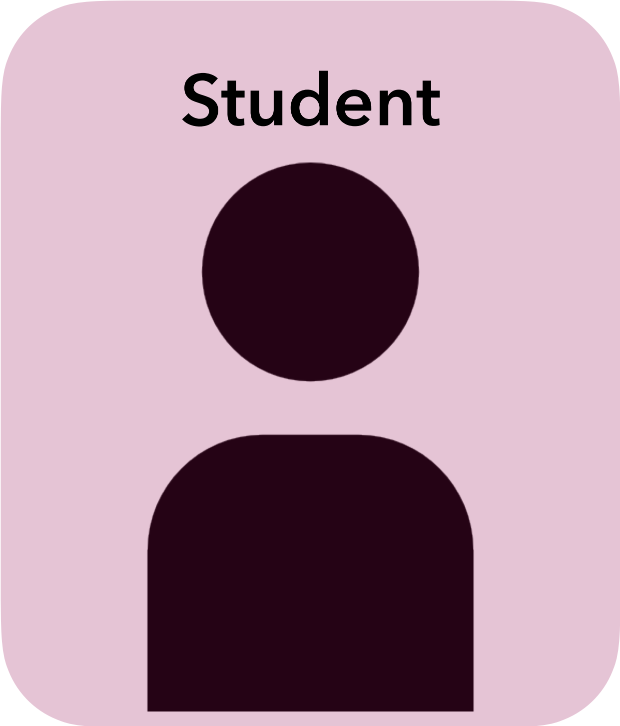
Excess loss over optimal loss

MMLU

Algorithm	Invariant	Strictly Better	Varying
H-Only	0.01 ± 0.01	0.18 ± 0.17	0.22 ± 0.12
H-LLM	0.01 ± 0.01	0.18 ± 0.21	0.12 ± 0.17
Population	0.00 ± 0.02	0.19 ± 0.07	0.12 ± 0.09
THREAD-LinUCB	0.00 ± 0.01	0.12 ± 0.03	0.07 ± 0.04
THREAD-KNN	0.01 ± 0.01	$\mathbf{0.05 \pm 0.03}$	$\mathbf{0.05 \pm 0.03}$

If a decision-maker benefits from having support some of the time, we can learn their policy online

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

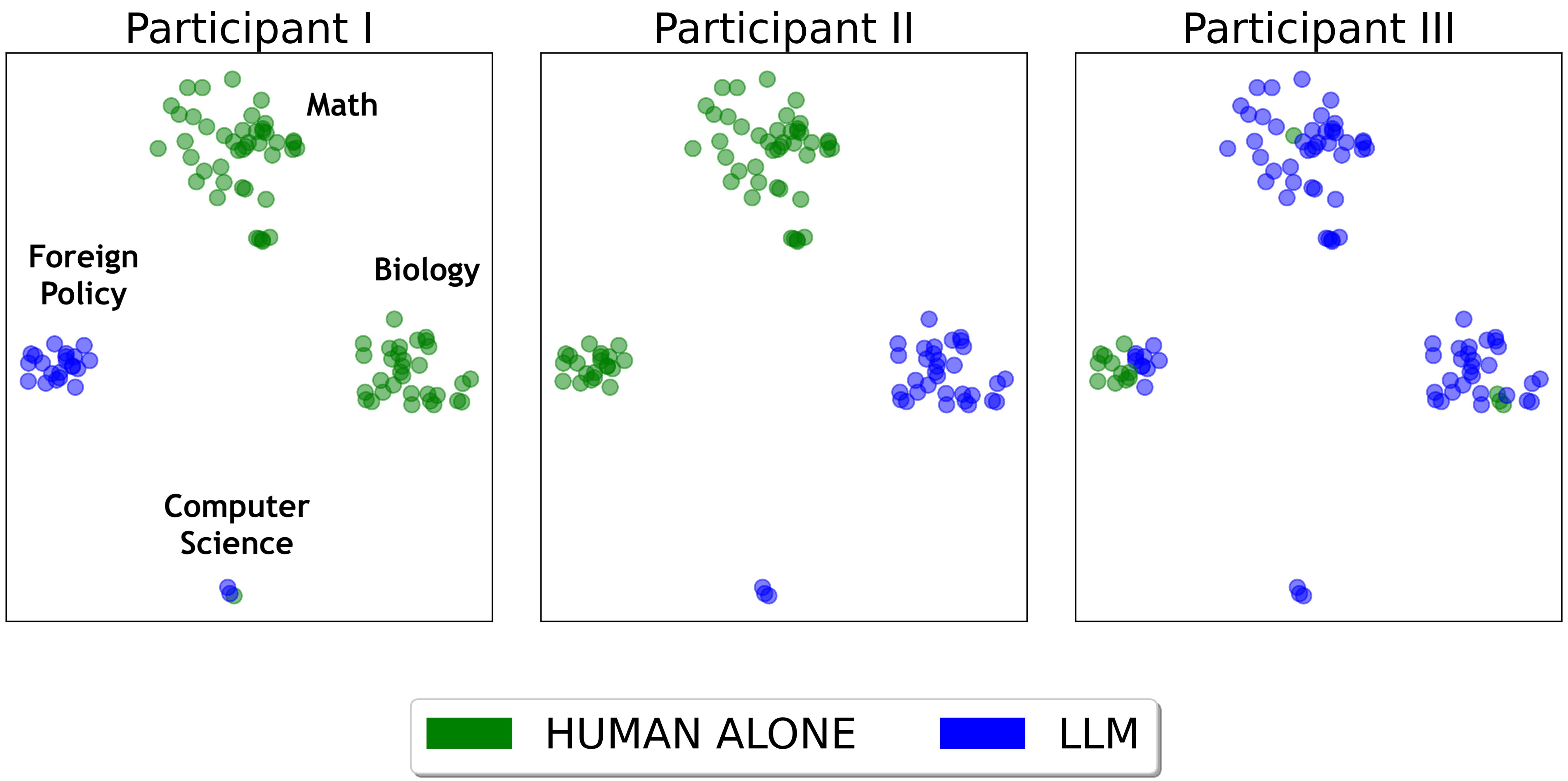


Personalize Access

Learning Personalized Decision Support Policies

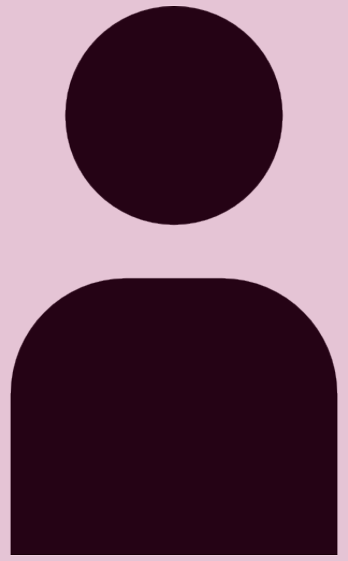
User Studies

Interactive Evaluation: Users interact with our tool, **Modiste**, which uses THREAD to learn when users require support online.



B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

Student



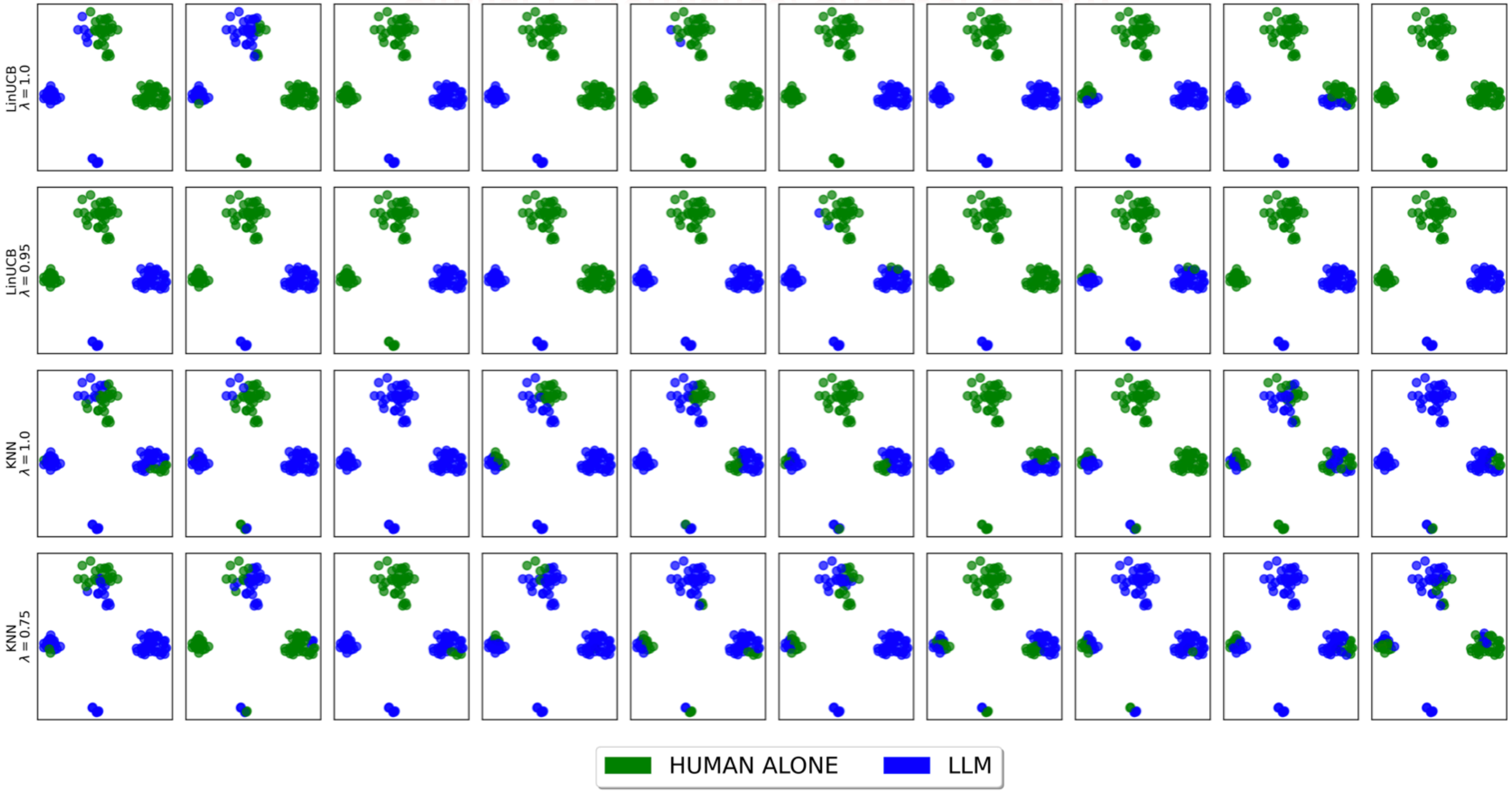
Personalize Access

Learning Personalized Decision Support Policies

User Studies

Interactive Evaluation: Users interact with our tool, **Modiste**, which uses THREAD to learn when users require support online.

Similar Performance, Cheaper Cost!!!



B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

Algorithmic resignation is the *deliberate* and *informed* disengagement from AI assistance in certain scenarios.

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.

Algorithmic resignation extends beyond the disuse of AI systems.

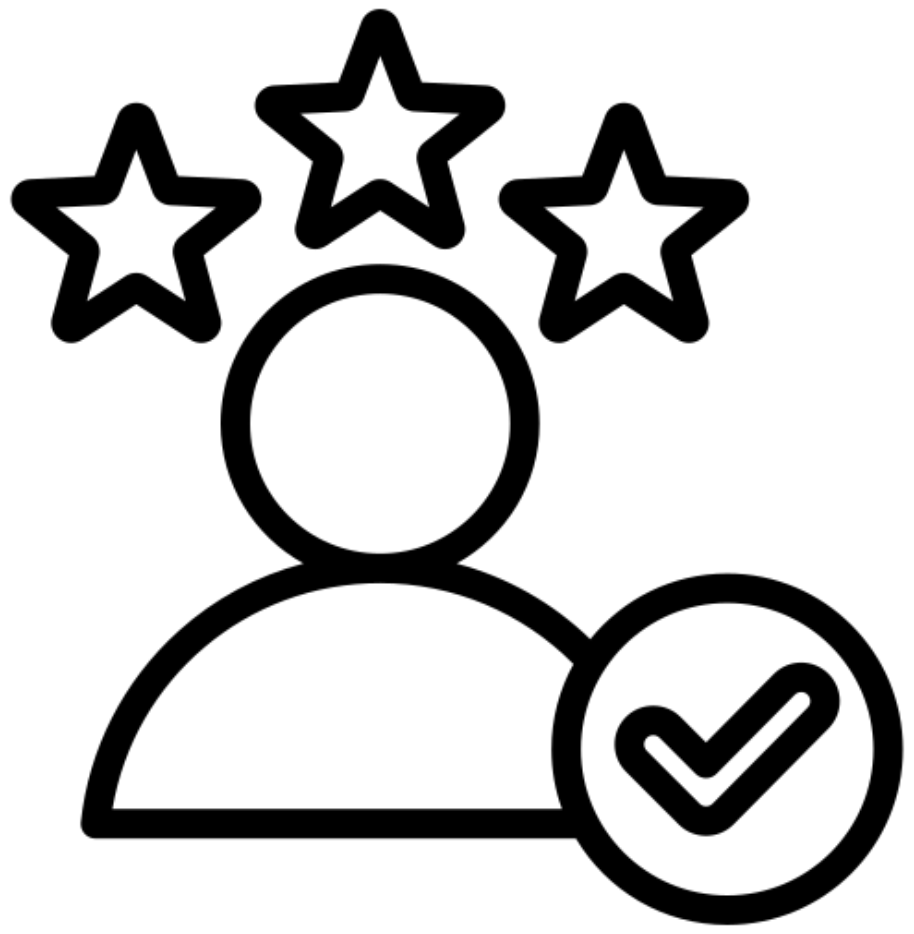
It is about embedding **governance** mechanisms directly within AI systems, guiding when and how these systems should be used or abstained from.

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.

B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

Benefits of Algorithmic Resignation

Economic Efficiency

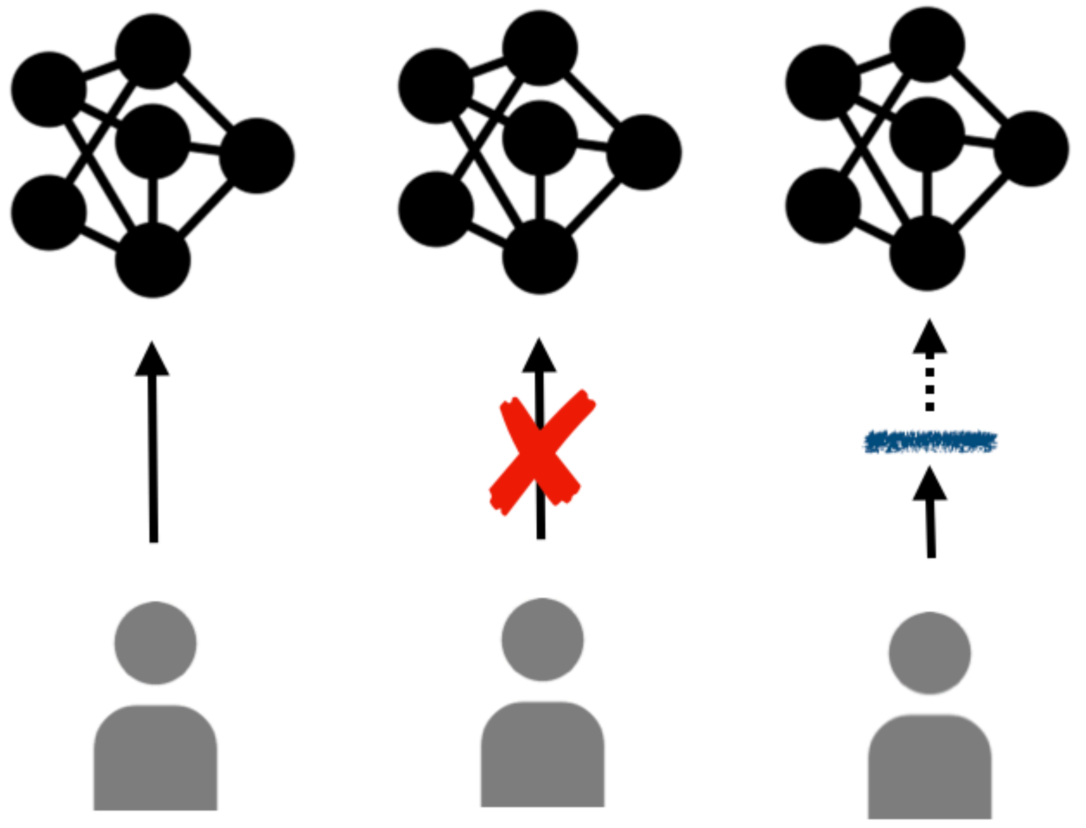


Reputational Gain

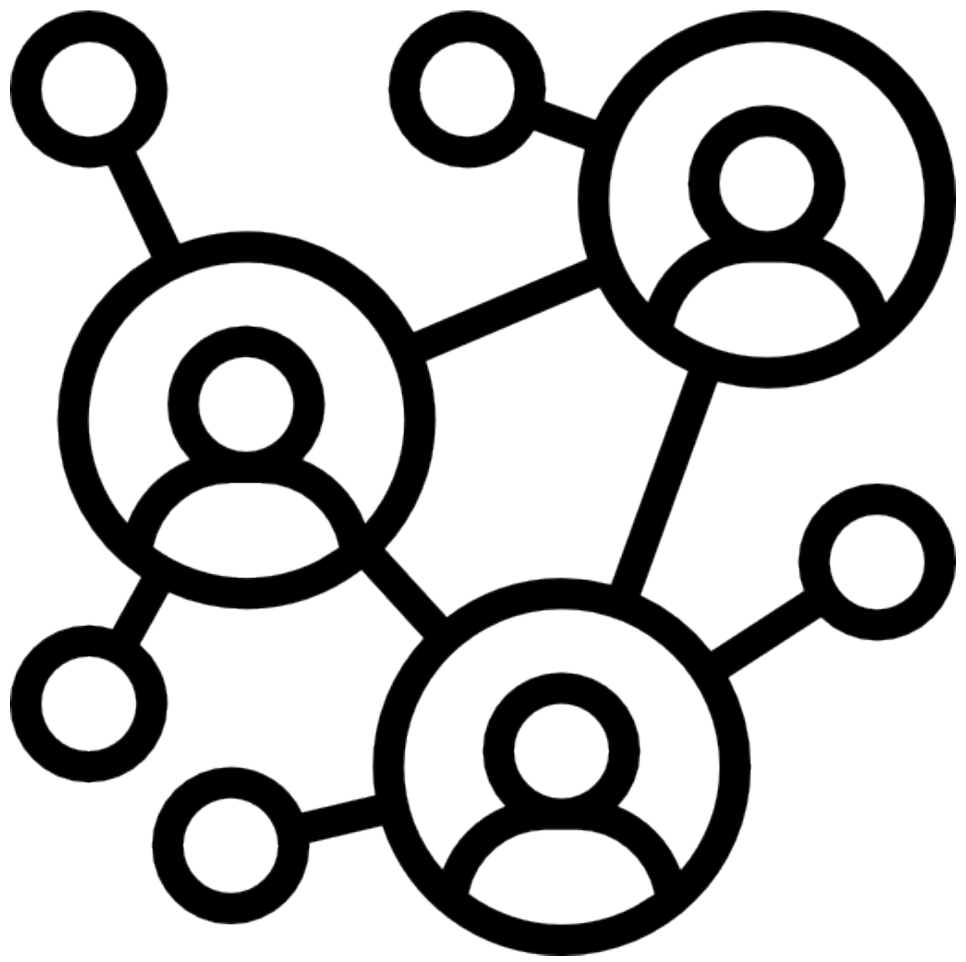
Legal Compliance

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.

Considerations for Algorithmic Resignation



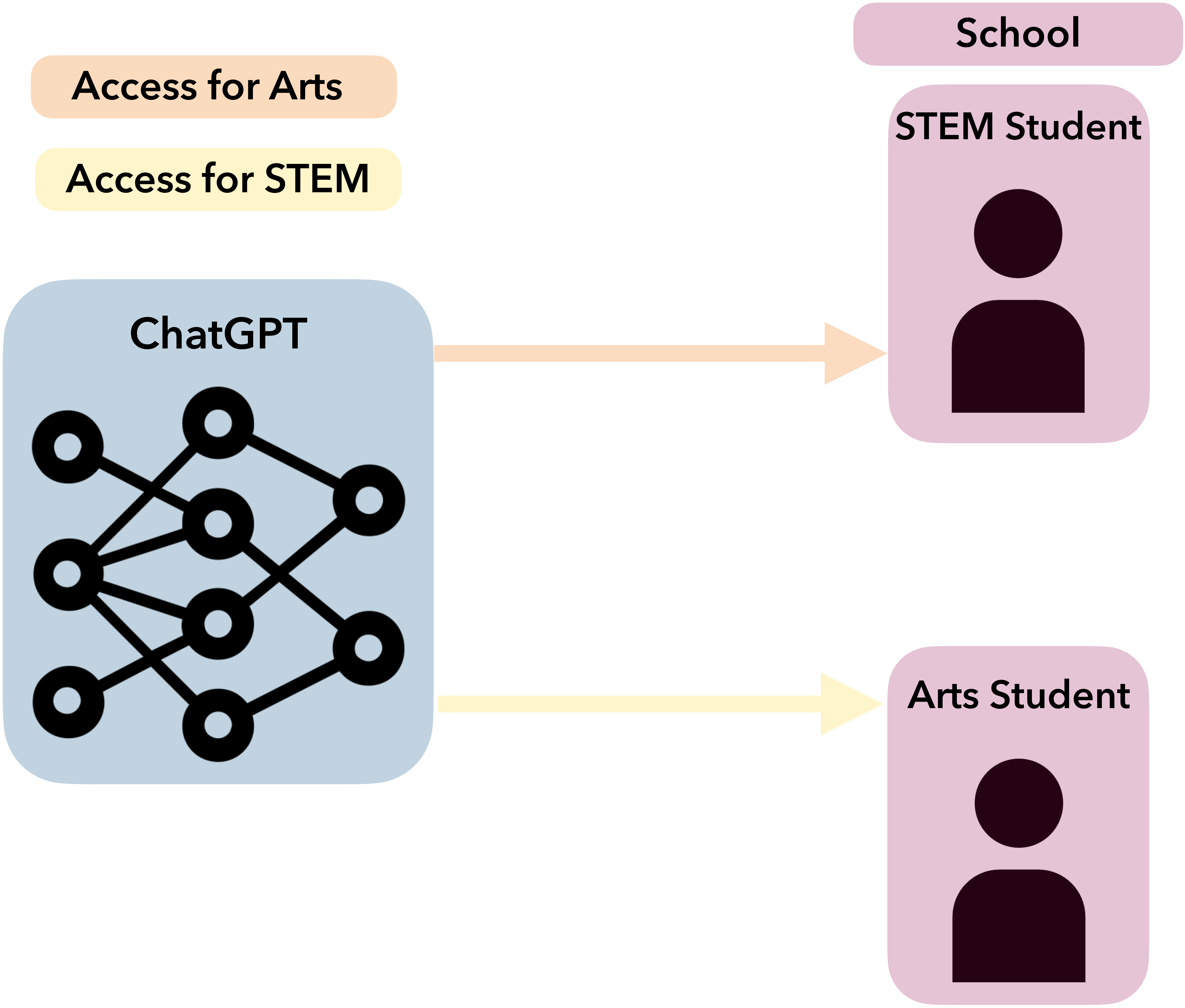
Friction over Resignation



Stakeholder Incentives

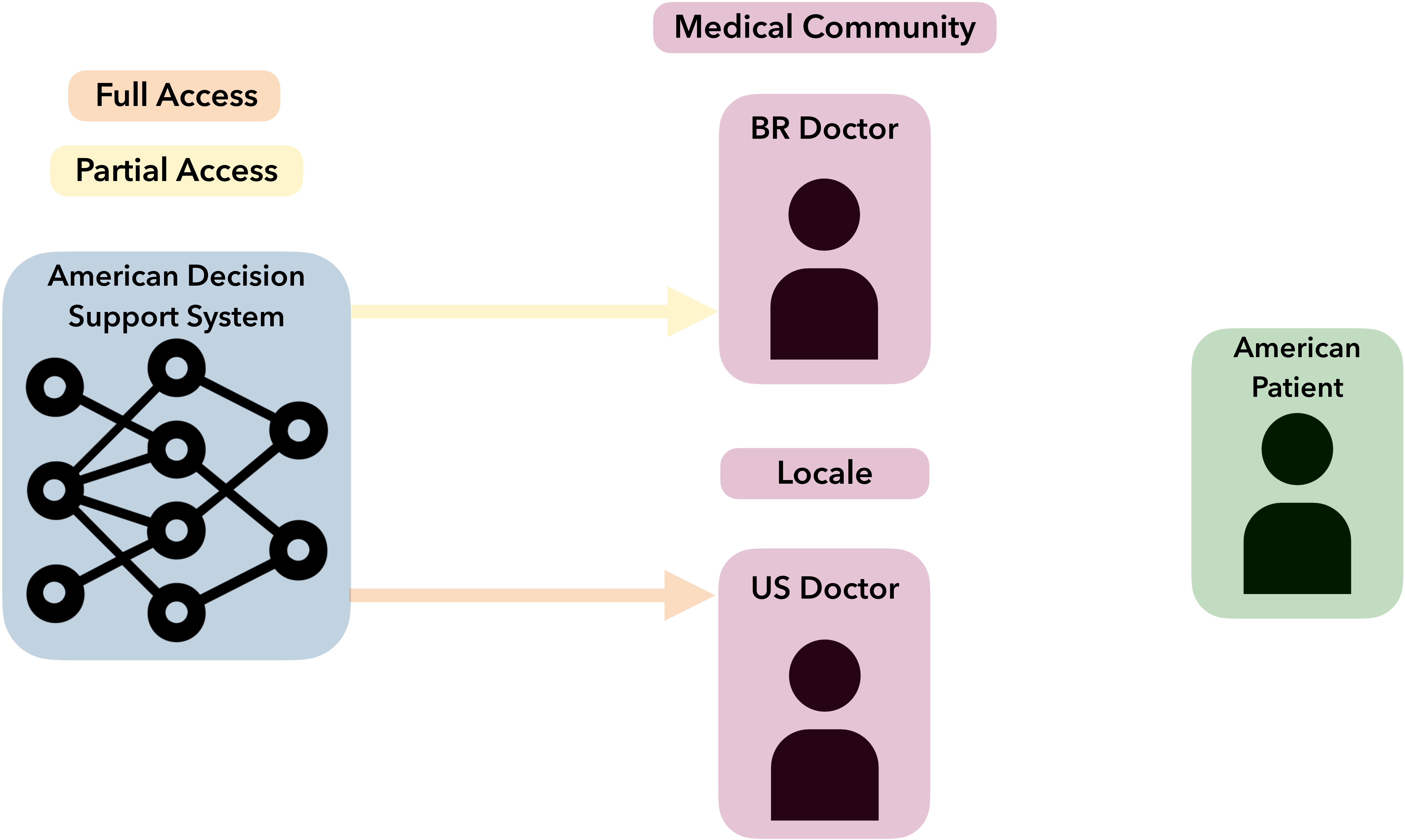
Level of Engagement

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.

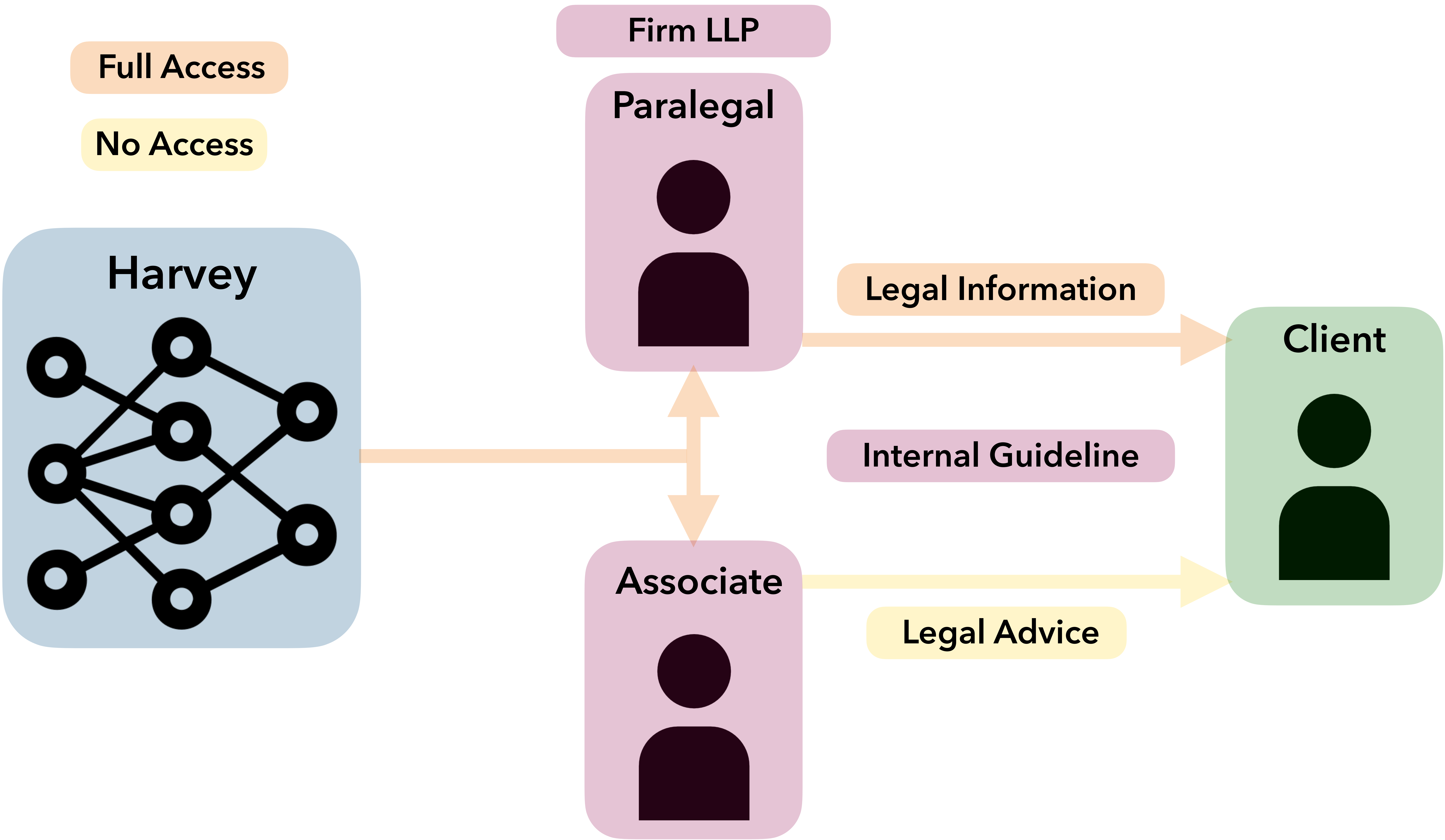


1. Different students will need different levels of support
2. Access to support can be learned over a series of interactions
3. Access may be complementary to expertise

B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.



B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.



B*, Sargeant*. *When Should Algorithms Resign?* IEEE Computer (Forthcoming). 2024.