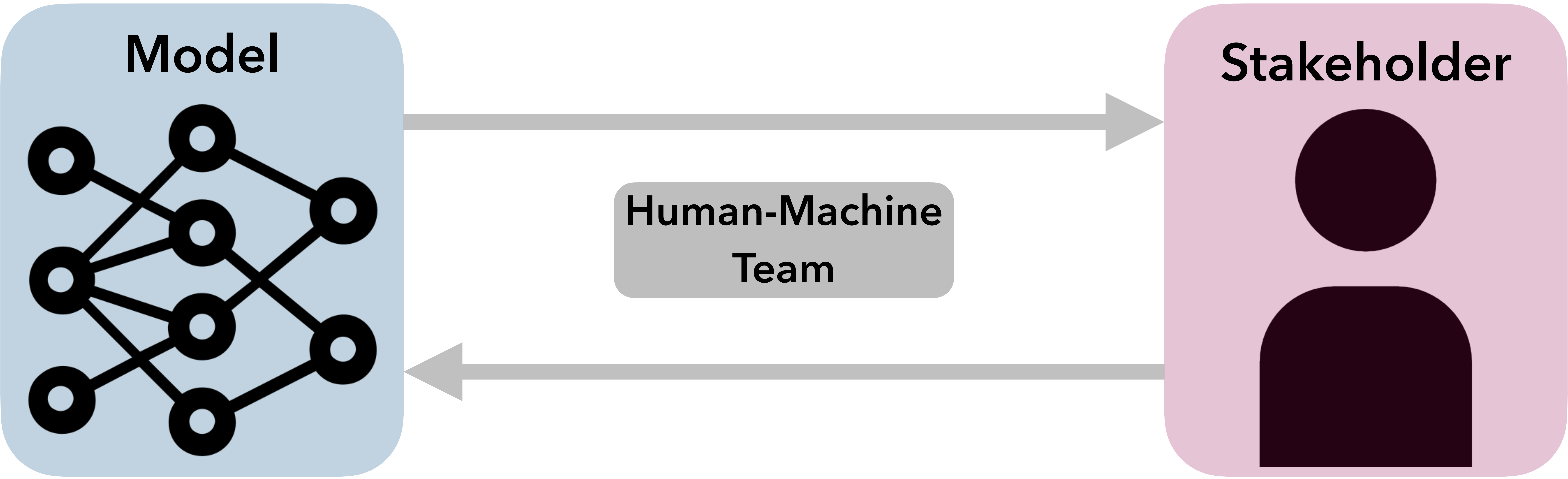
Trustworthy Machine Learning

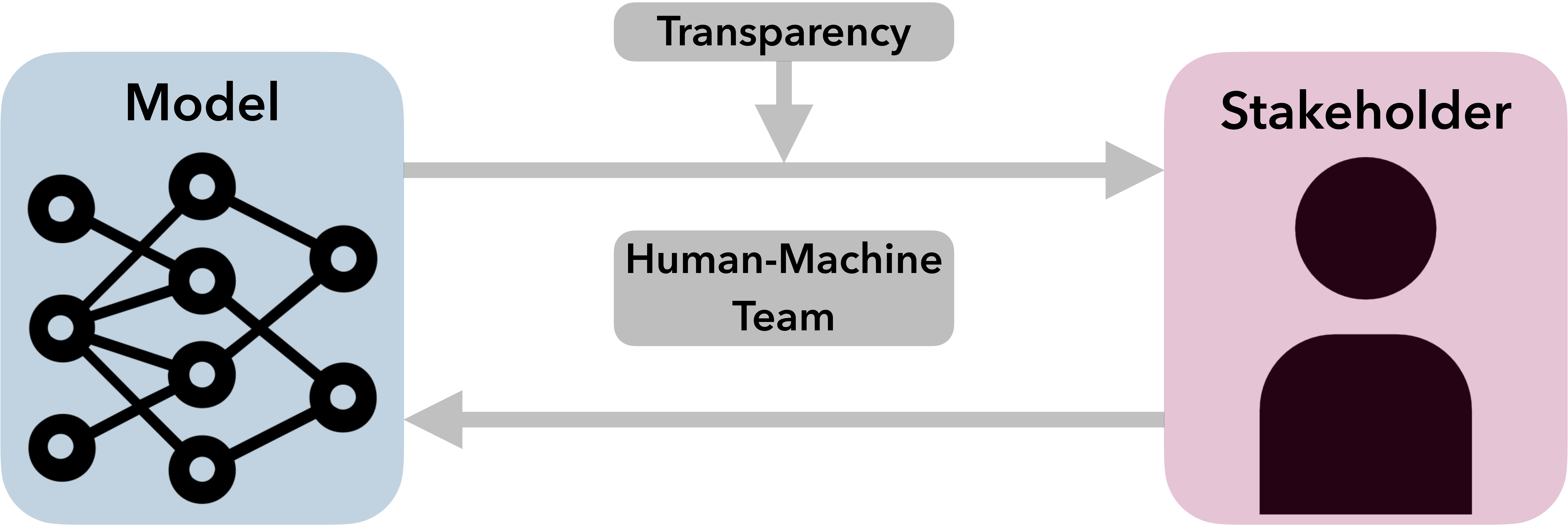
From Algorithmic Transparency to Decision Support

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

@umangsbhatt
umangbhatt@nyu.edu

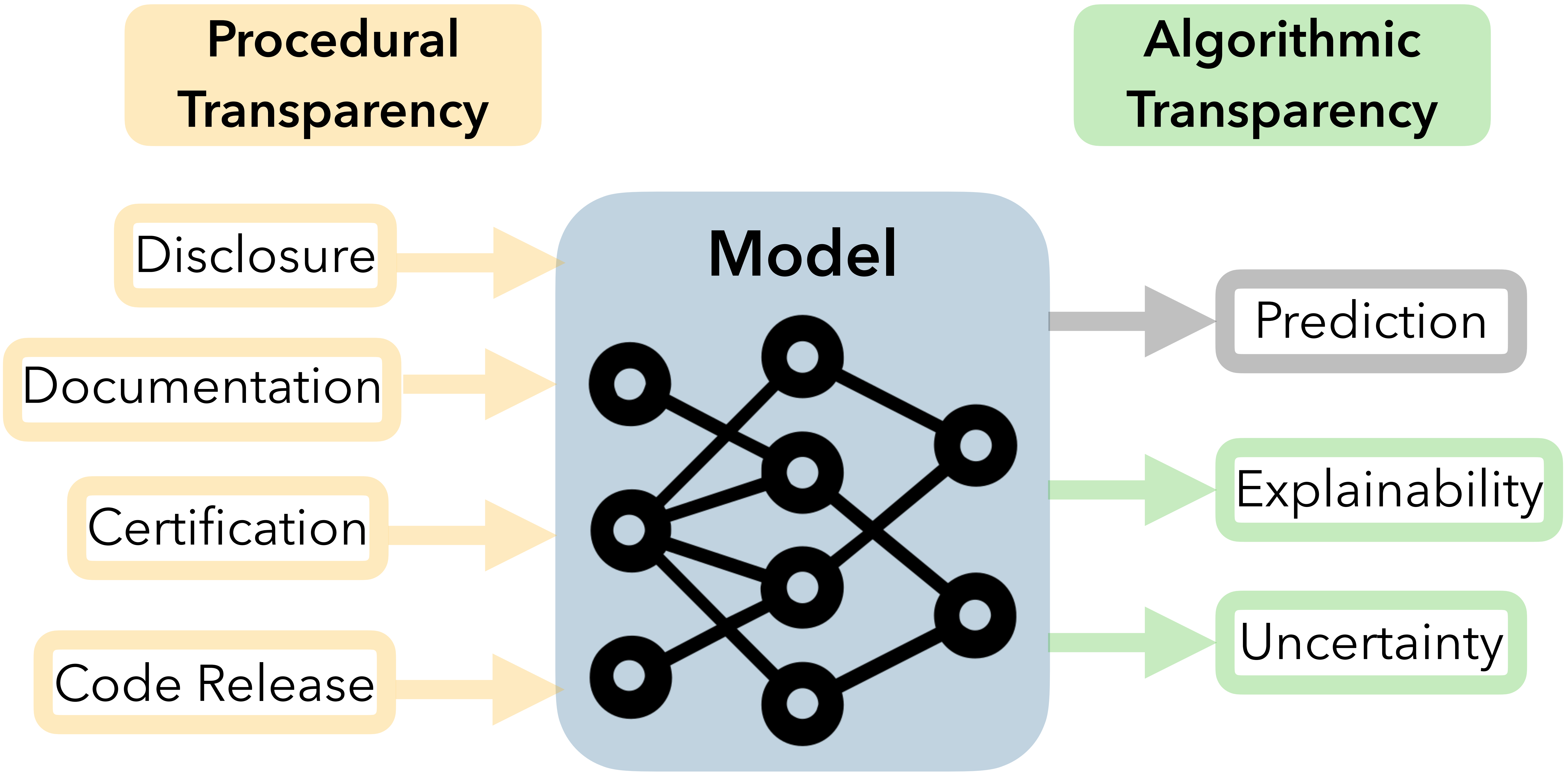
NYU
The Alan Turing Institute
CFI


Model
Human-Machine
Team
Stakeholder

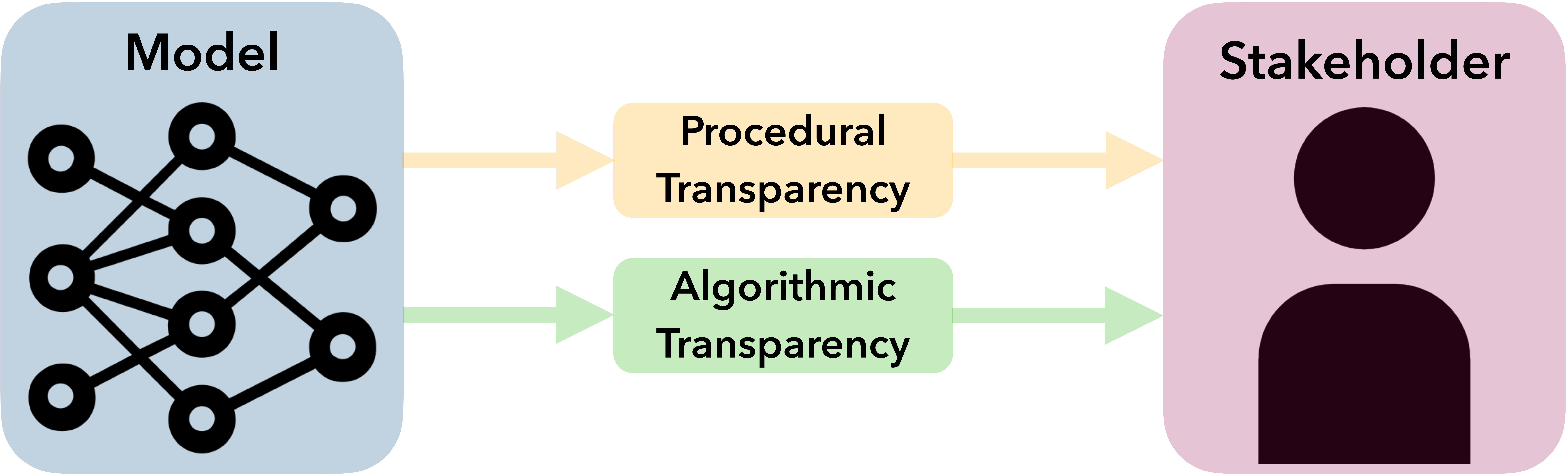


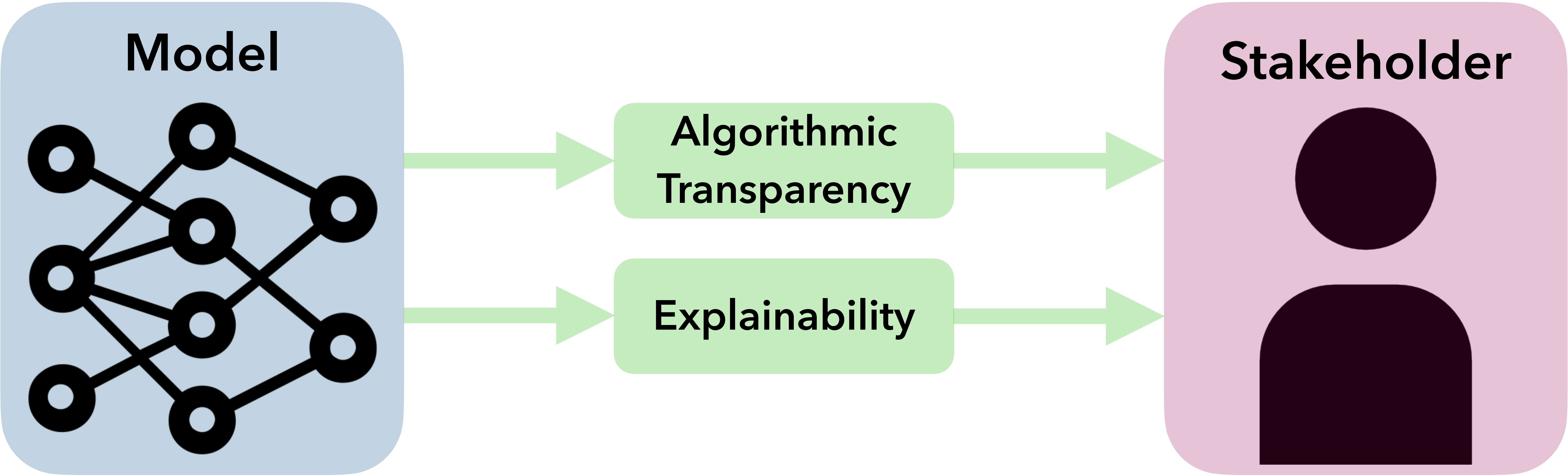
Transparency means providing stakeholders with *relevant* information about how a model works

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



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


Model
Procedural Transparency
Algorithmic Transparency
Stakeholder



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

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

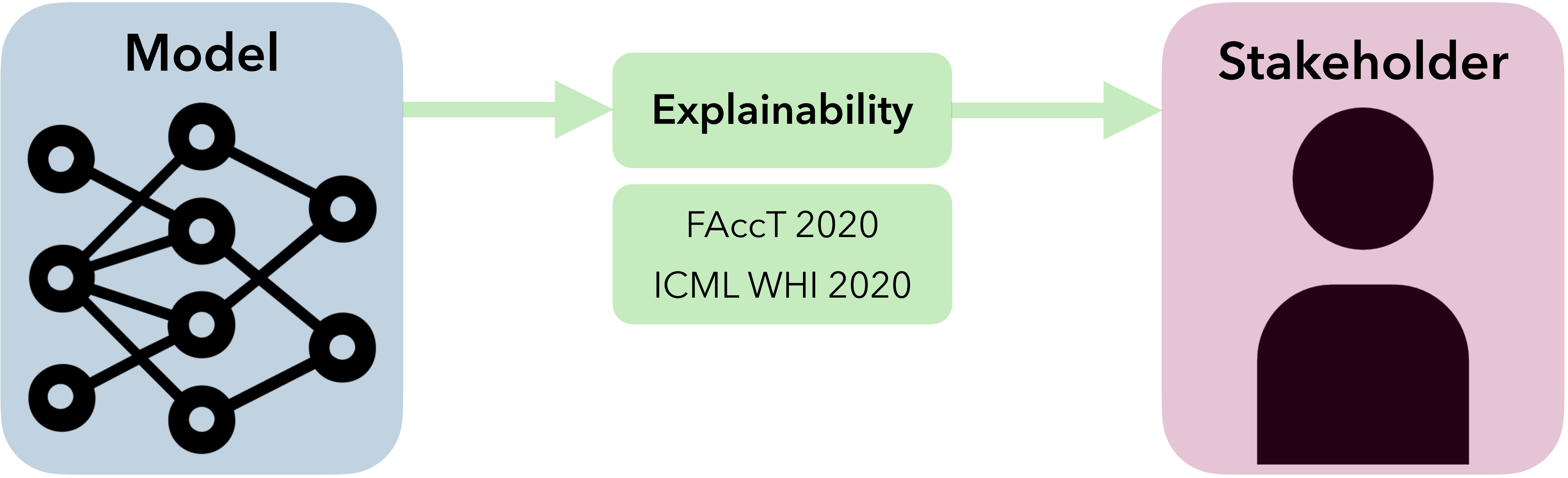
Research Style

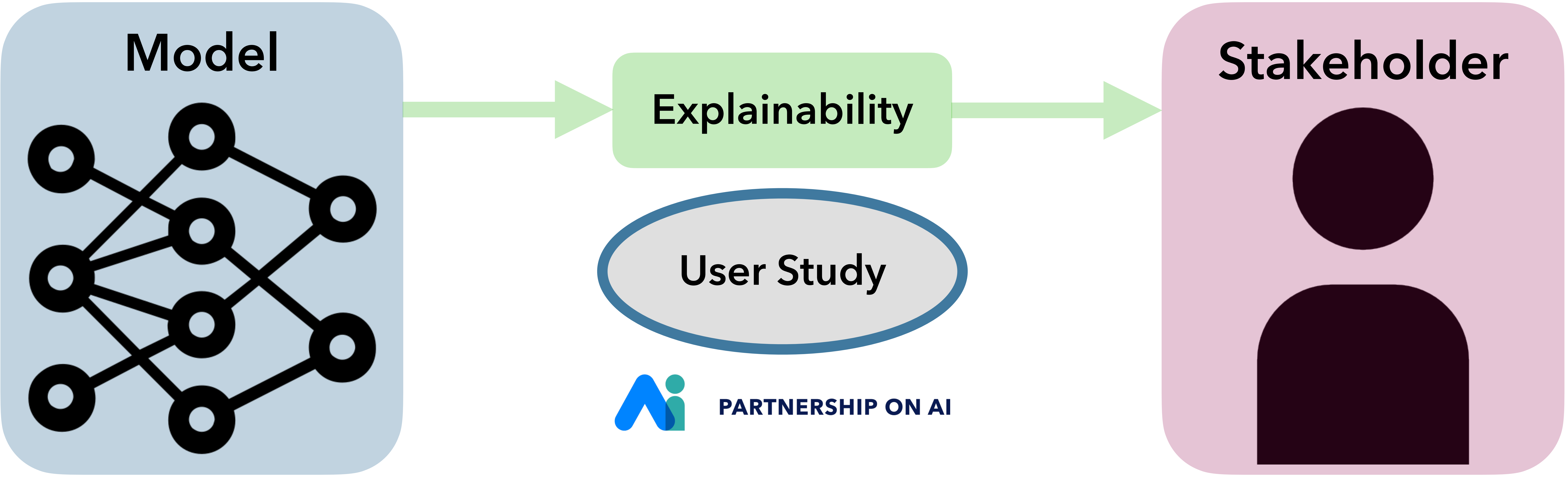
Convenings

Methods

User Studies

Chapter 3


Model
Explainability
FAccT 2020
ICML WHI 2020
Stakeholder

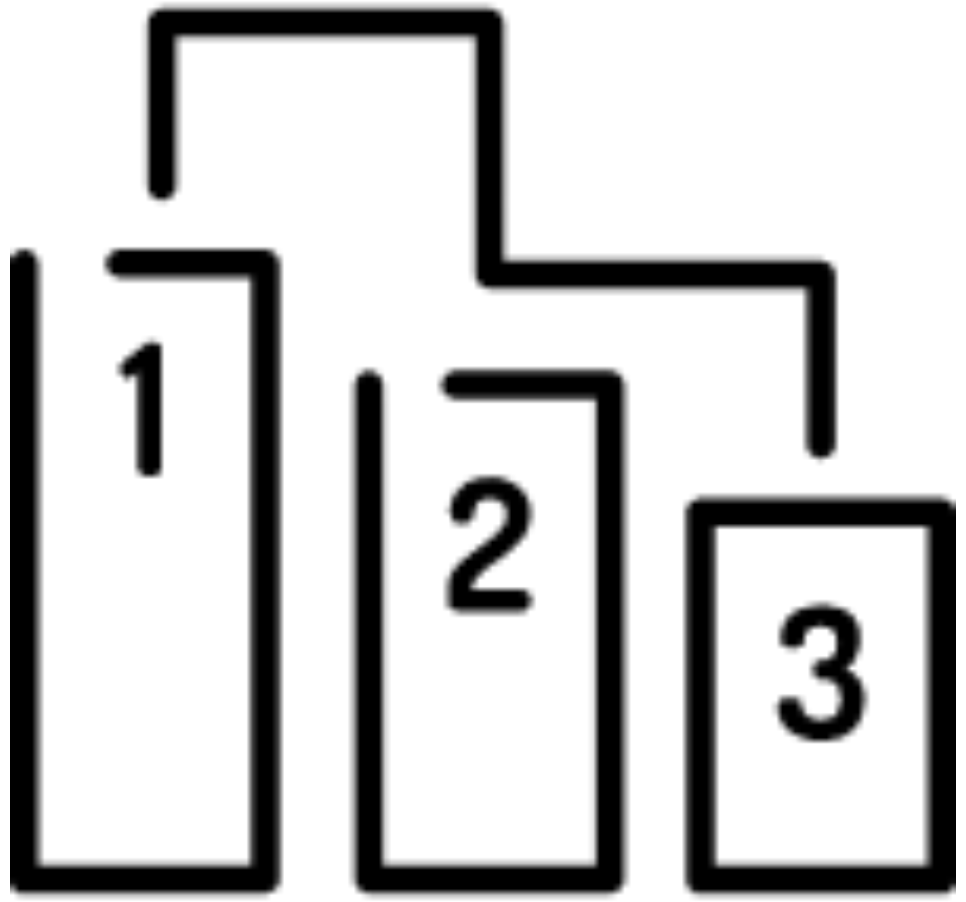


Goal: understand how explainability methods are used in *practice*

Approach: 30min to 2hr *semi-structured* interviews with 50 individuals from 30 organizations

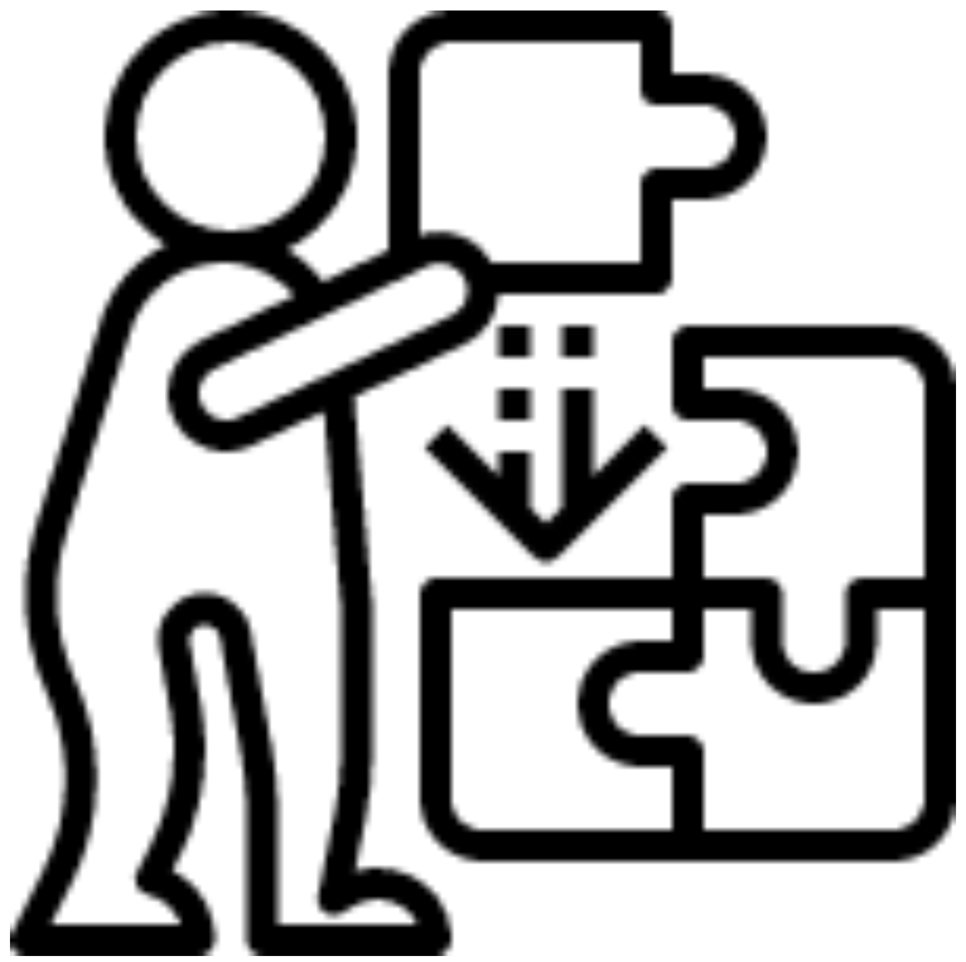
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Popular Explanation Styles



Feature Importance

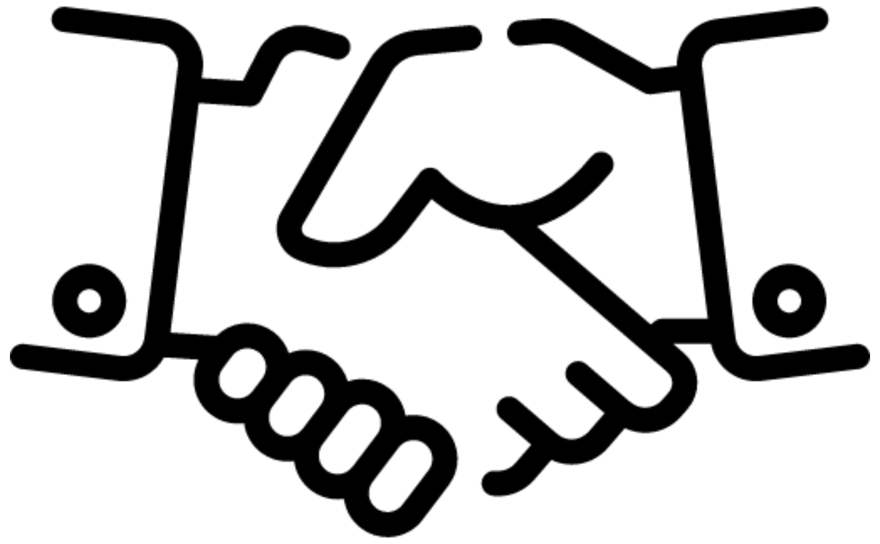
Sample Importance



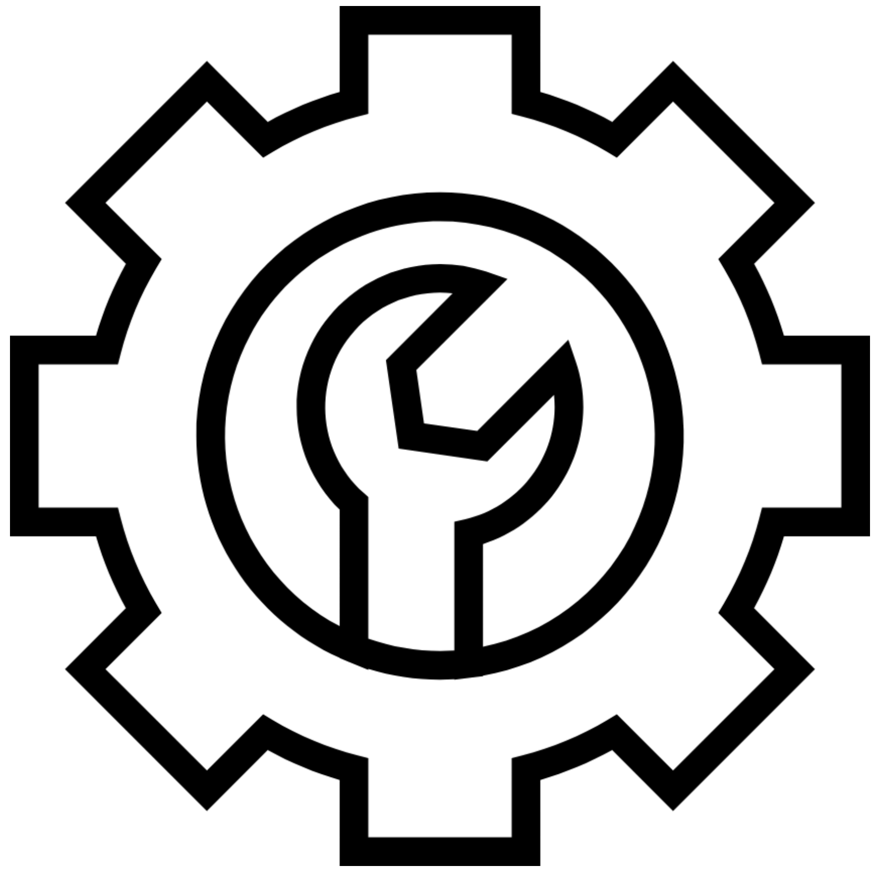
Counterfactuals

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

Common Explanation Stakeholders



Executives



Engineers

End Users

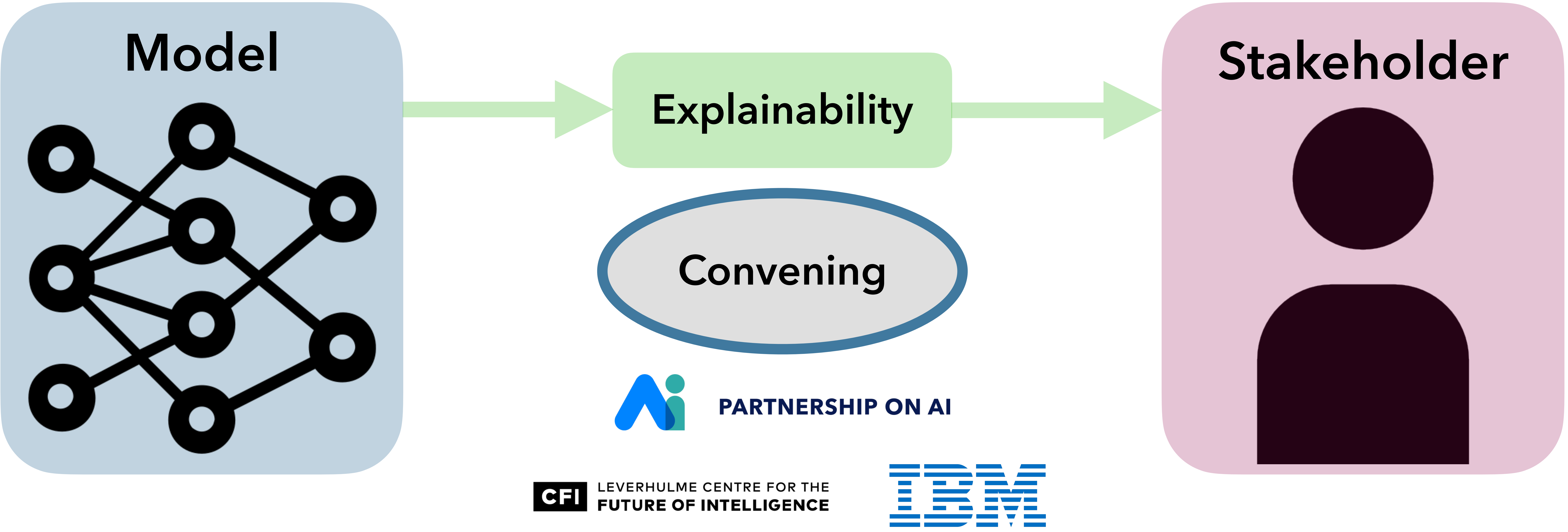
Regulators

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

Findings

1. Explainability is used for **debugging** internally
2. **Goals** of explainability are not clearly defined within organizations
3. Technical **limitations** make explainability hard to deploy in real-time

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



Goal: facilitate an *inter-stakeholder* conversation around explainability

Conclusion: *Community engagement* and *context consideration* are important factors in deploying explainability thoughtfully

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

Community Engagement

1. *In which **context** will this explanation be used?*

2. *How should the explanation be **evaluated**? Both quantitatively and qualitatively…*

3. *Can we prevent data misuse and preferential treatment by involving **affected groups** in the development process?*

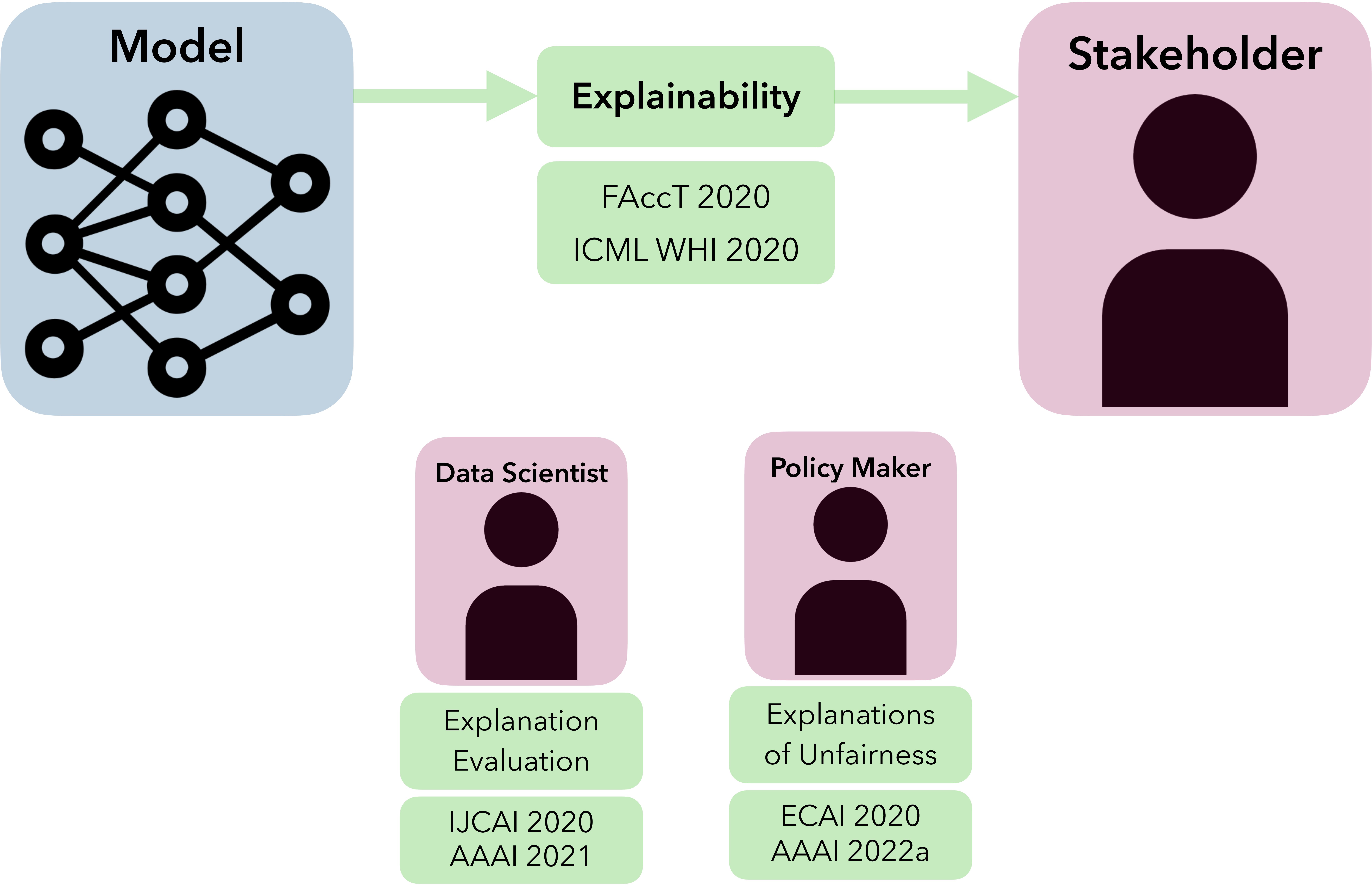
4. *Can we **educate** stakeholders regarding the functionalities and limitations of explainable machine learning?*

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

Deploying Explainability

1. *How does **uncertainty** in the model's predictions and explanation technique affect the resulting explanations?*
2. *How can stakeholders **interact** with the resulting explanations?*
3. *How, if at all, will stakeholder **behavior** change as a result of the explanation shown?*
4. *Over **time**, how will the model and explanations adapt to changes in stakeholder behavior?*

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


Model
Explainability
FAccT 2020
ICML WHI 2020
Stakeholder
Data Scientist
Explanation Evaluation
IJCAI 2020
AAAI 2021
Policy Maker
Explanations of Unfairness
ECAI 2020
AAAI 2022a

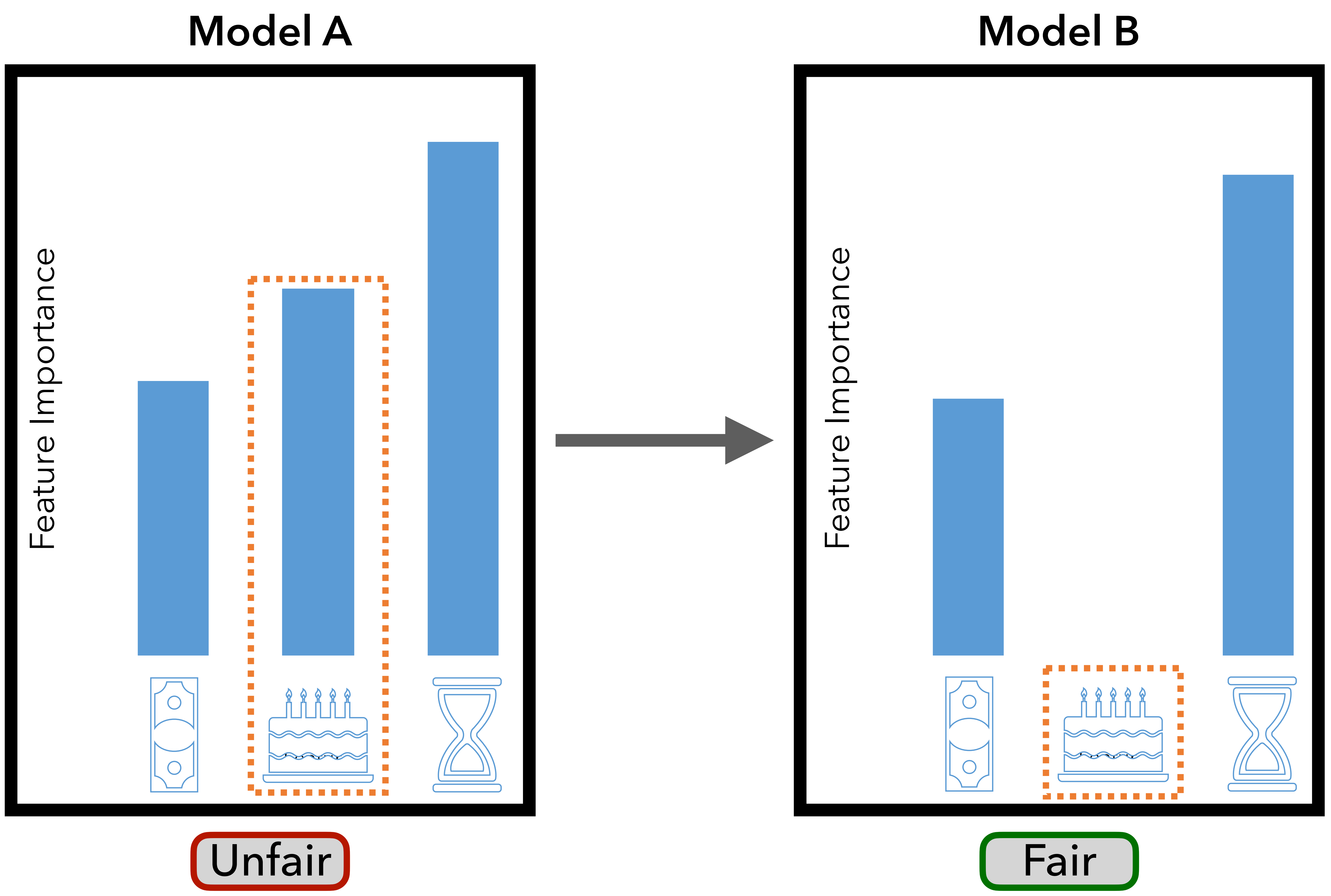
Policy Maker

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 assume model fairness via explanations

Methods

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

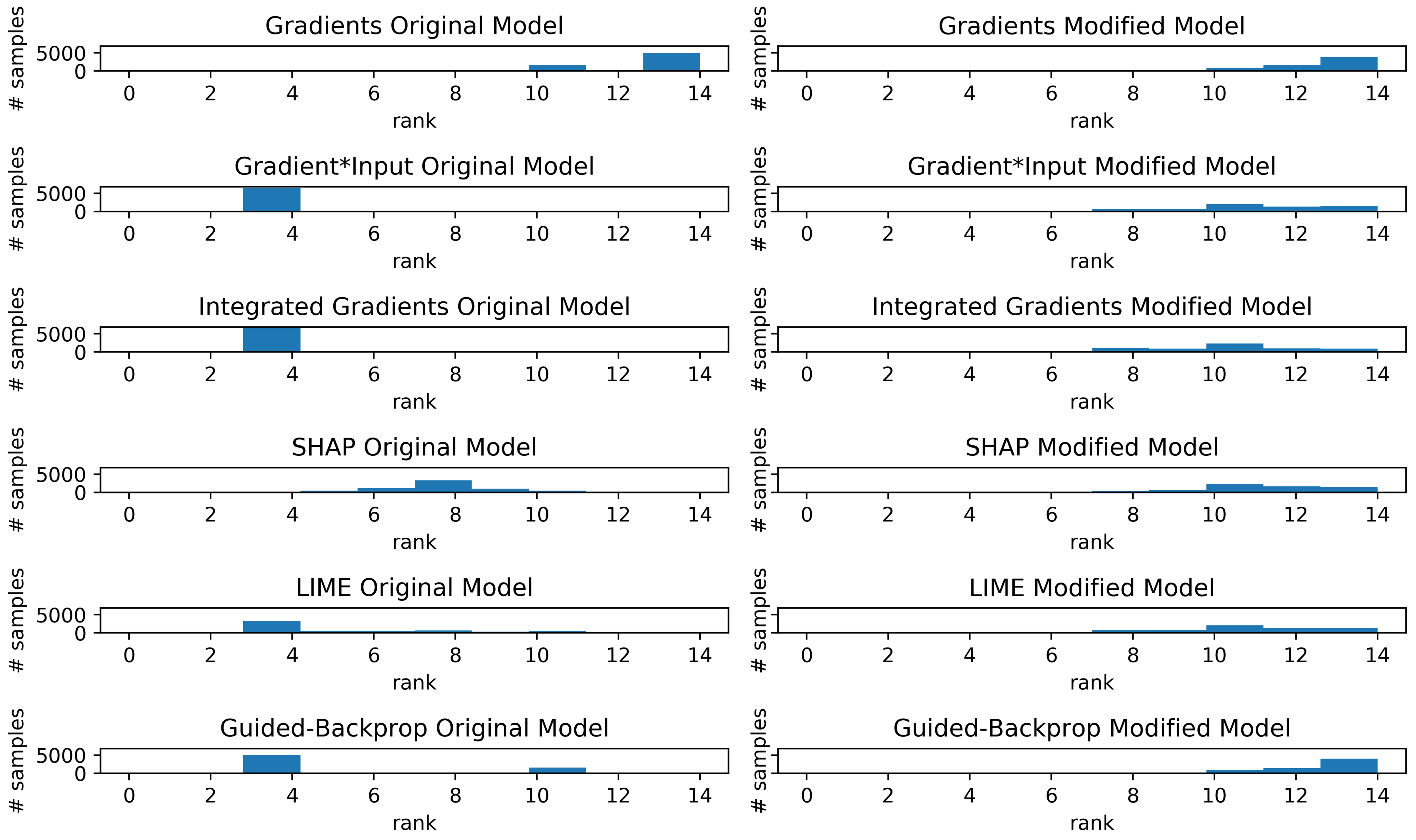
Our Goal $f_\theta \rightarrow f_{\theta+\delta}$

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.

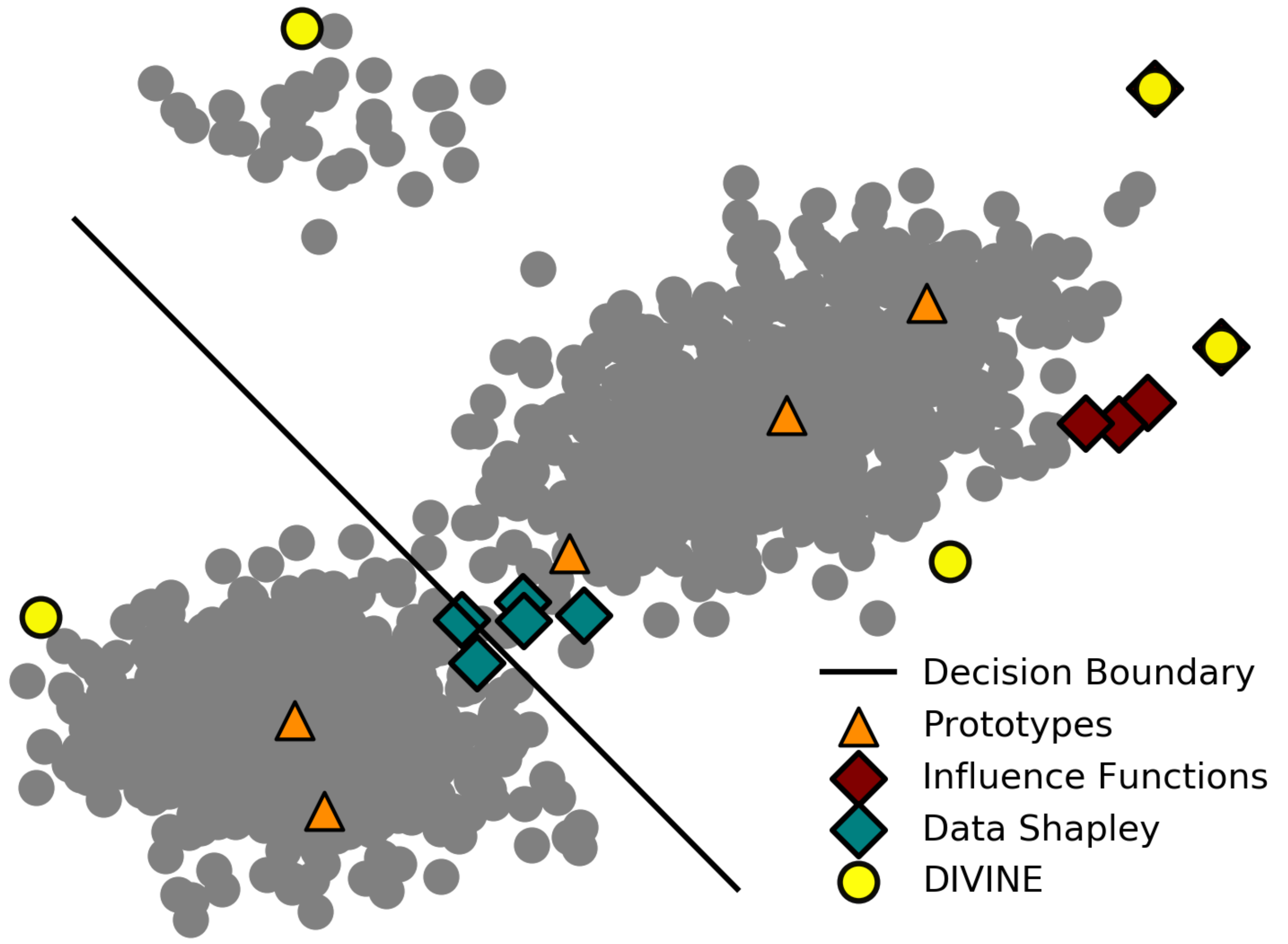
Data Scientist

Explanation Evaluation

DIVINE: DIVerse INfluEntial Training Points

Methods

Question: *"Which training points are important to a specific prediction?"*



Formulation: *Can we find a set of m training points that are not only influential to the model but also diverse in input space?*

I. Measuring Influence

$$f_{\text{loss}}(\theta) = \sum_{i=1}^{n} l(x_i, y_i; \theta)$$

$$I_i = f(\hat{\theta}_i) - f(\hat{\theta})$$

$$I(S) = \sum_{i \in S} I_i$$

II. Measuring Diversity

Submodular $R(S)$

$$R_{\text{SR}}(S) = \kappa - \sum_{u,v \in S} \phi(u, v)$$

III. Optimizing for Both

$$\max_{S \in D, |S| = m} I(S) + \gamma R(S)$$

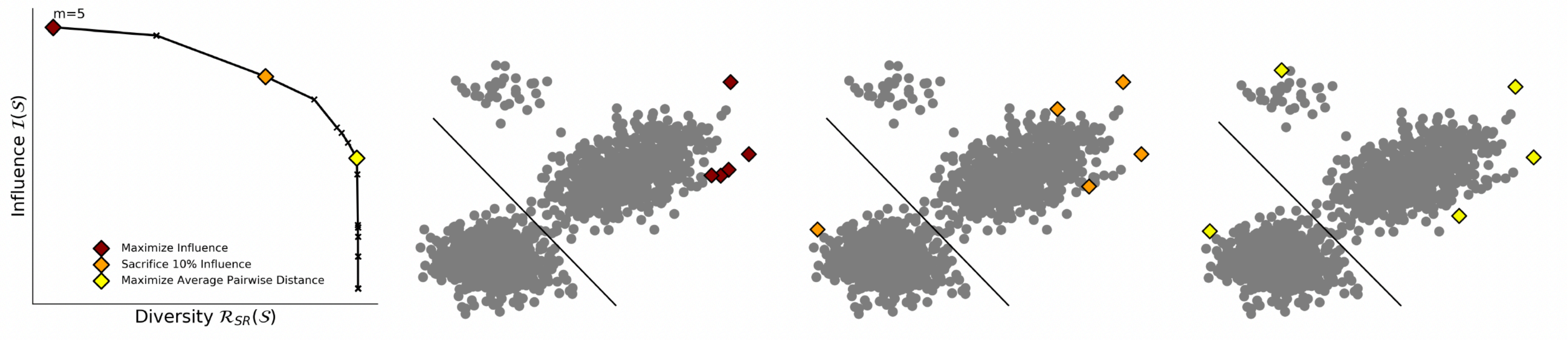
B, Chien, Zafar, Weller. *DIVINE: DIVerse INfluEntial Training Points*. Under Review. 2022.

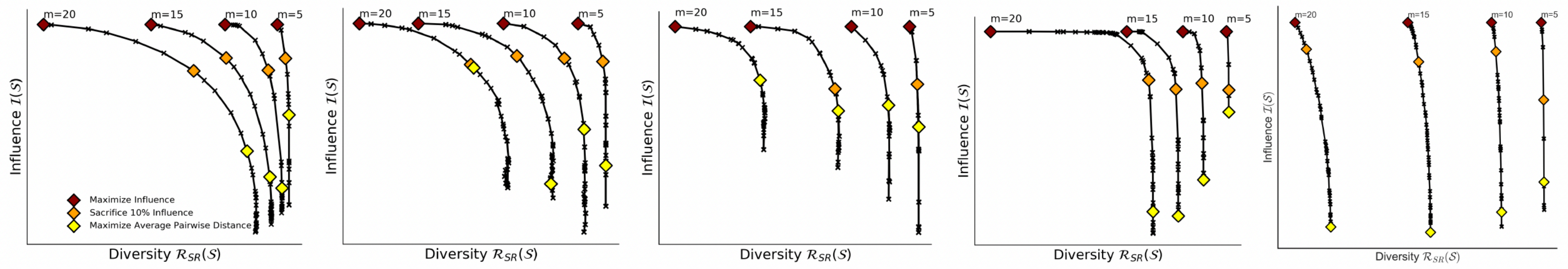
Data Scientist

Explanation Evaluation

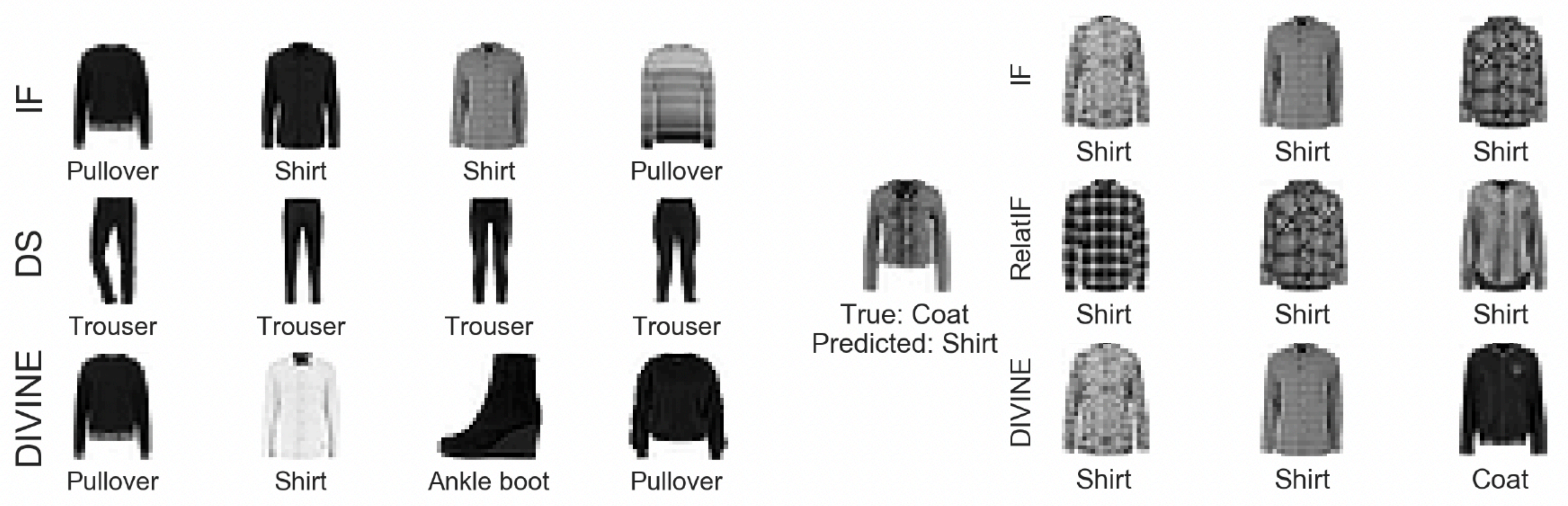
DIVINE: DIVerse INfluEntial Training Points

Methods





(a) Synthetic (b) LSAT (c) COMPAS (d) Adult (e) FashionMNIST



B, Chien, Zafar, Weller. *DIVINE: DIVerse INfluEntial Training Points*. Under Review. 2022.

Data Scientist

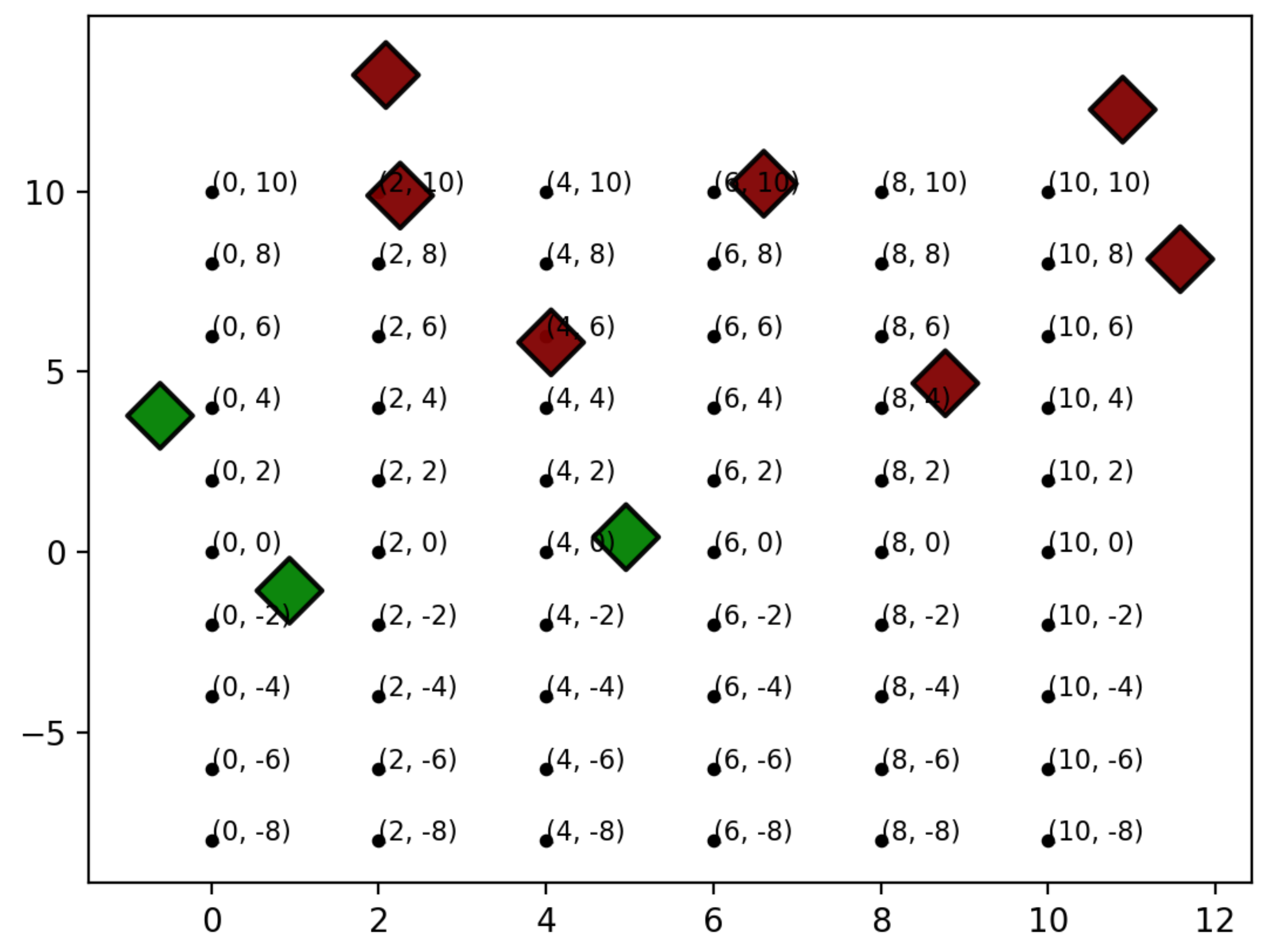
Explanation Evaluation

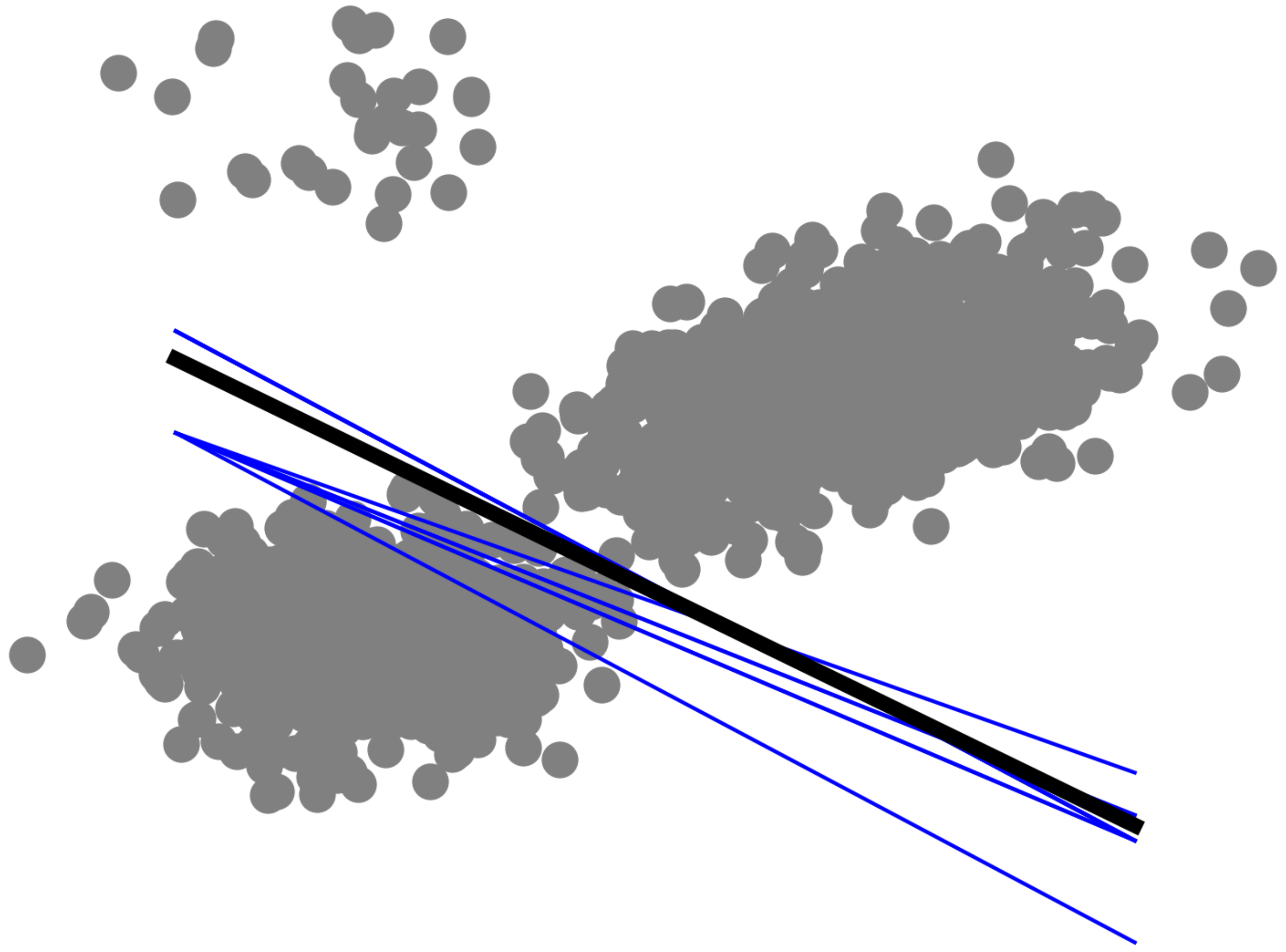
DIVINE: DIVerse INfluEntial Training Points

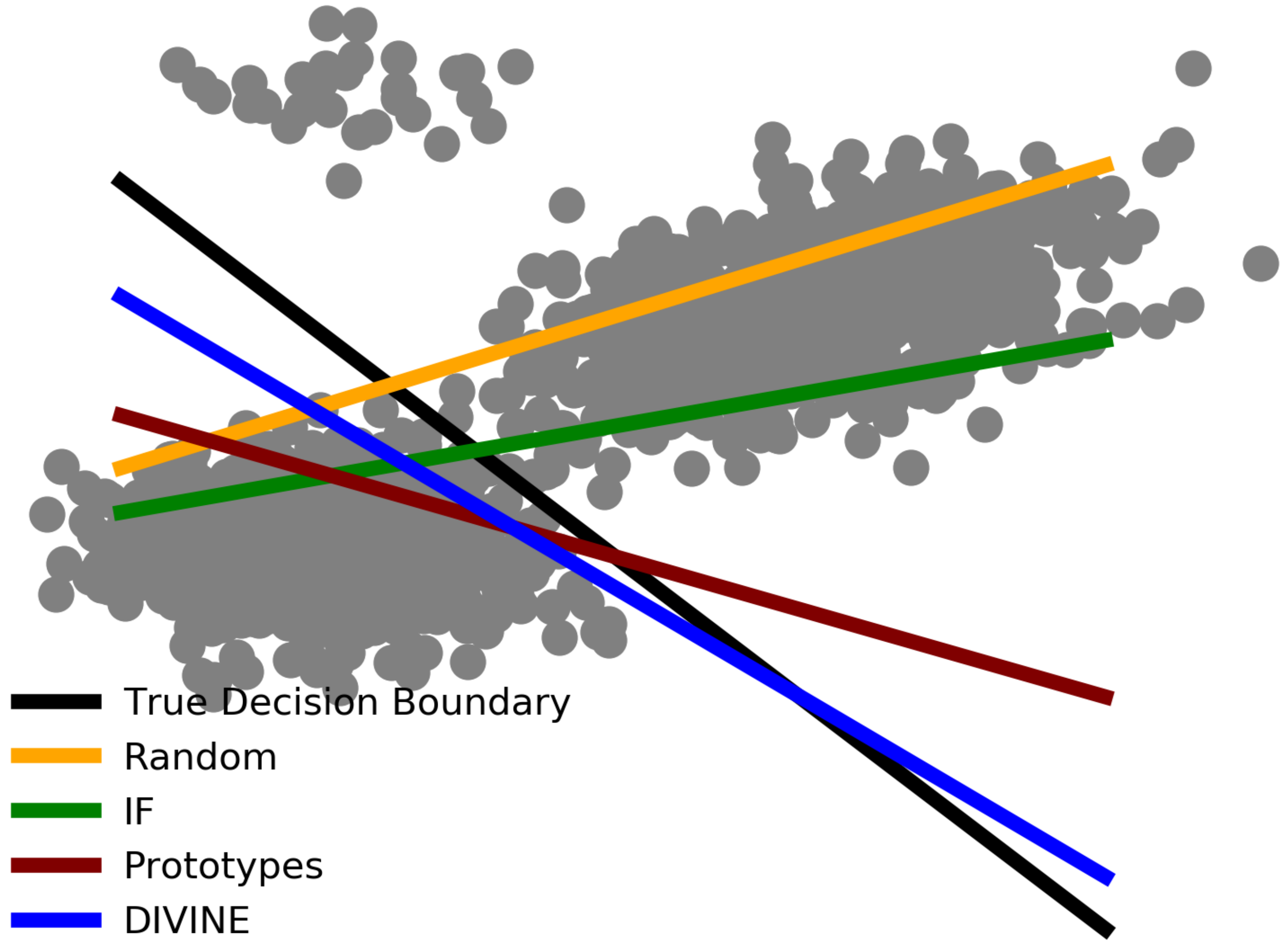
User Studies

Task Simulatability: Users how well a user can reason about an **entire** model given an explanation.

We show sets of points to a user and ask them to draw a decision boundary for each. Users decide upon a decision boundary by selecting two endpoints, which we then translate into a line.

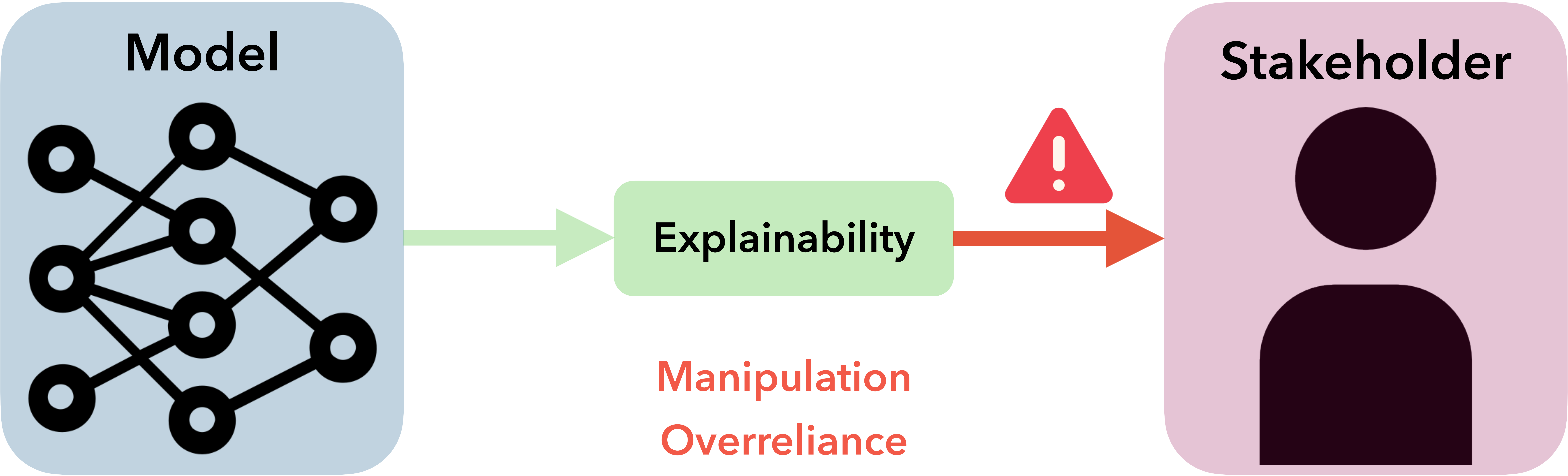




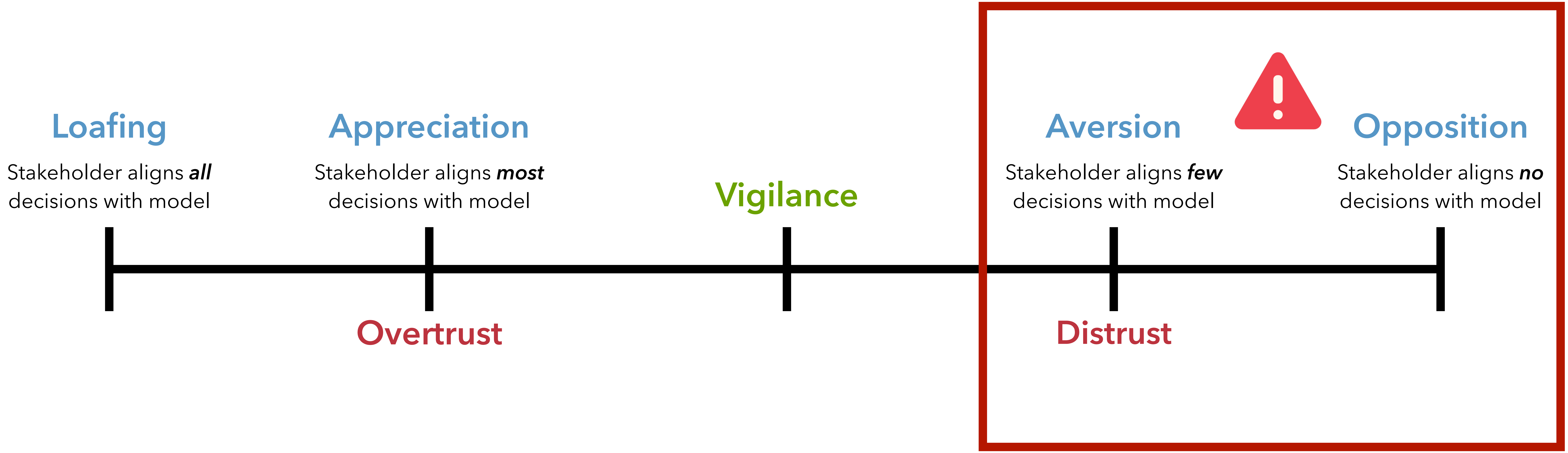


Upon calculating the cosine similarity between the true and user-drawn decision boundaries, we find that DIVINE points were considerably more helpful to users.

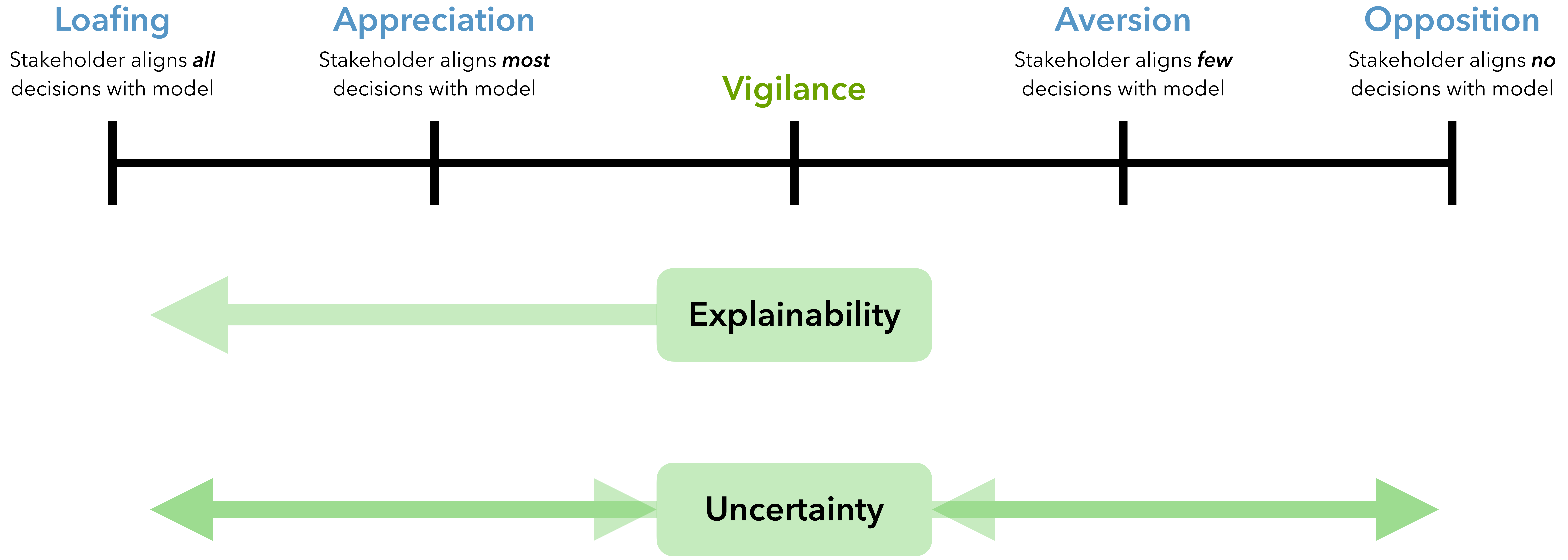
B, Chien, Zafar, Weller. *DIVINE: DIVerse INfluEntial Training Points*. Under Review. 2022.



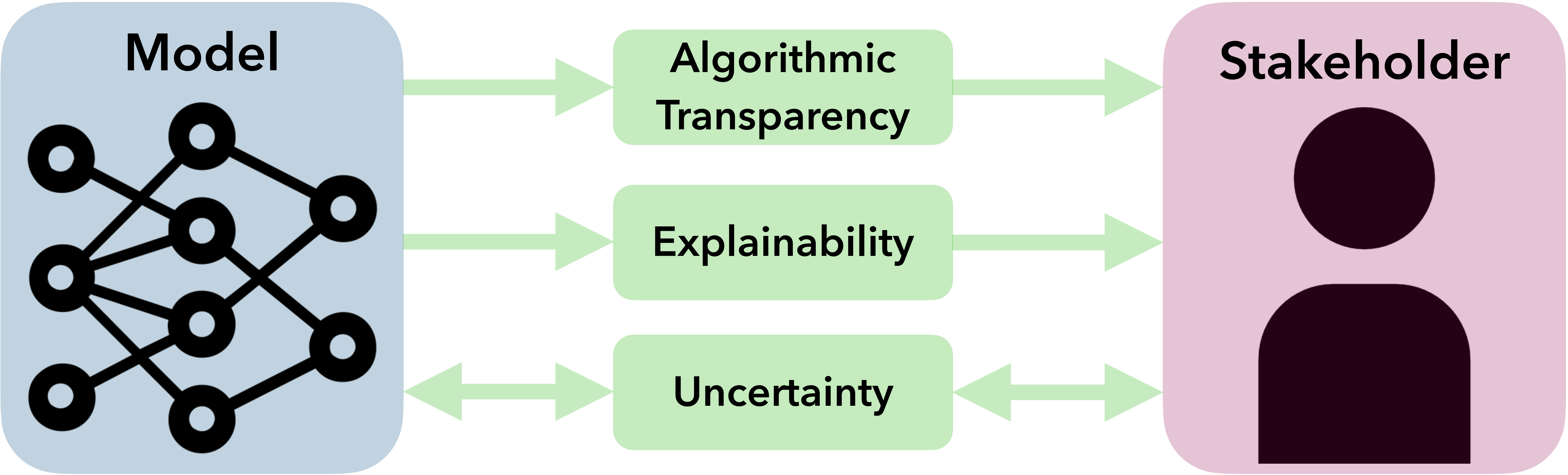
Weller. *Transparency: Motivations and Challenges*. Chapter 2 in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 2019
Buçinca, Malaya, Gajos. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. CSCW. 2021.
Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence*. Patterns. 2022.



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.

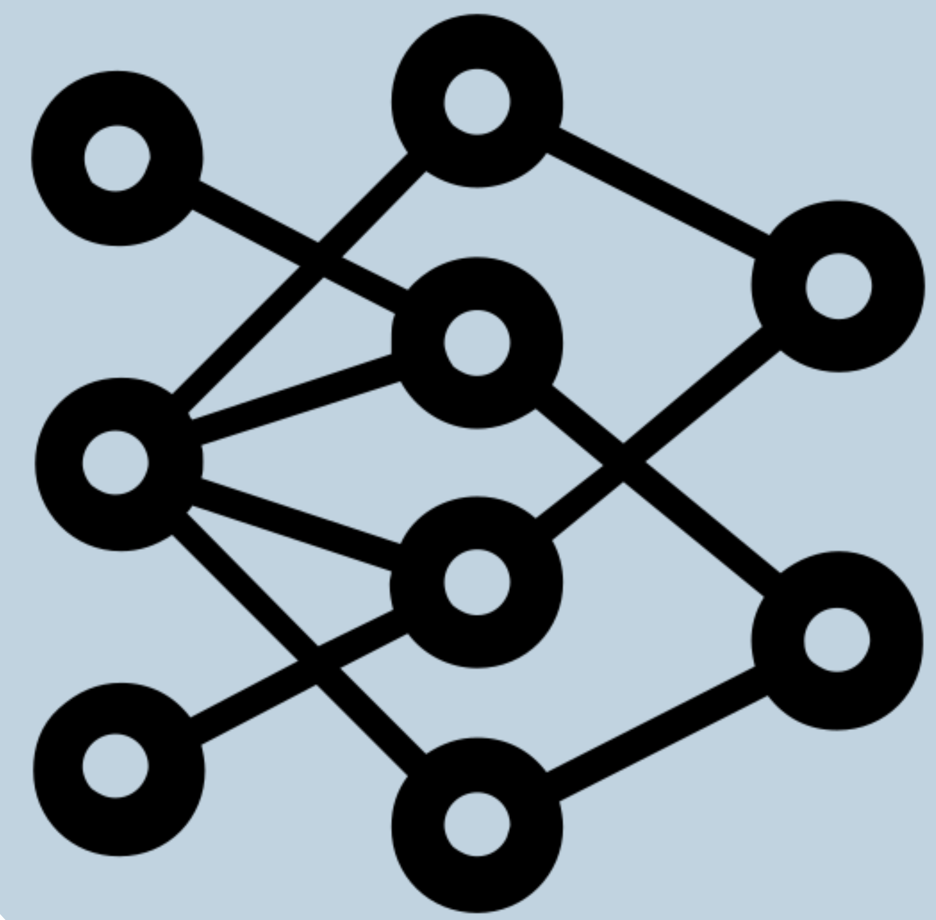


Buçinca, Malaya, Gajos. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. CSCW. 2021.
Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence*. Patterns. 2022.



B, Antoran, Zhang, Liao, Sattigeri, Fogliato, et al. *Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty*. ACM AIES. 2021.
Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence*. Patterns. 2022.

Model

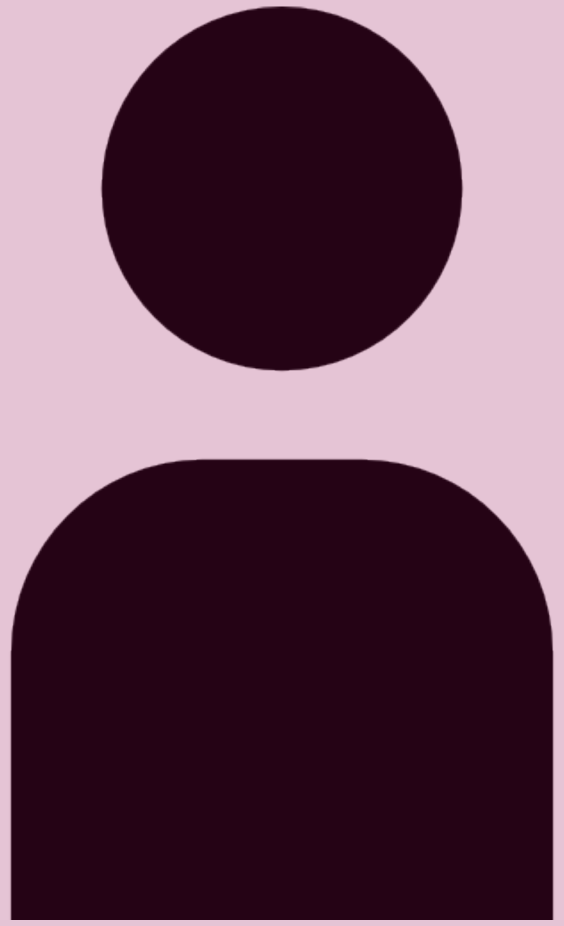


Uncertainty

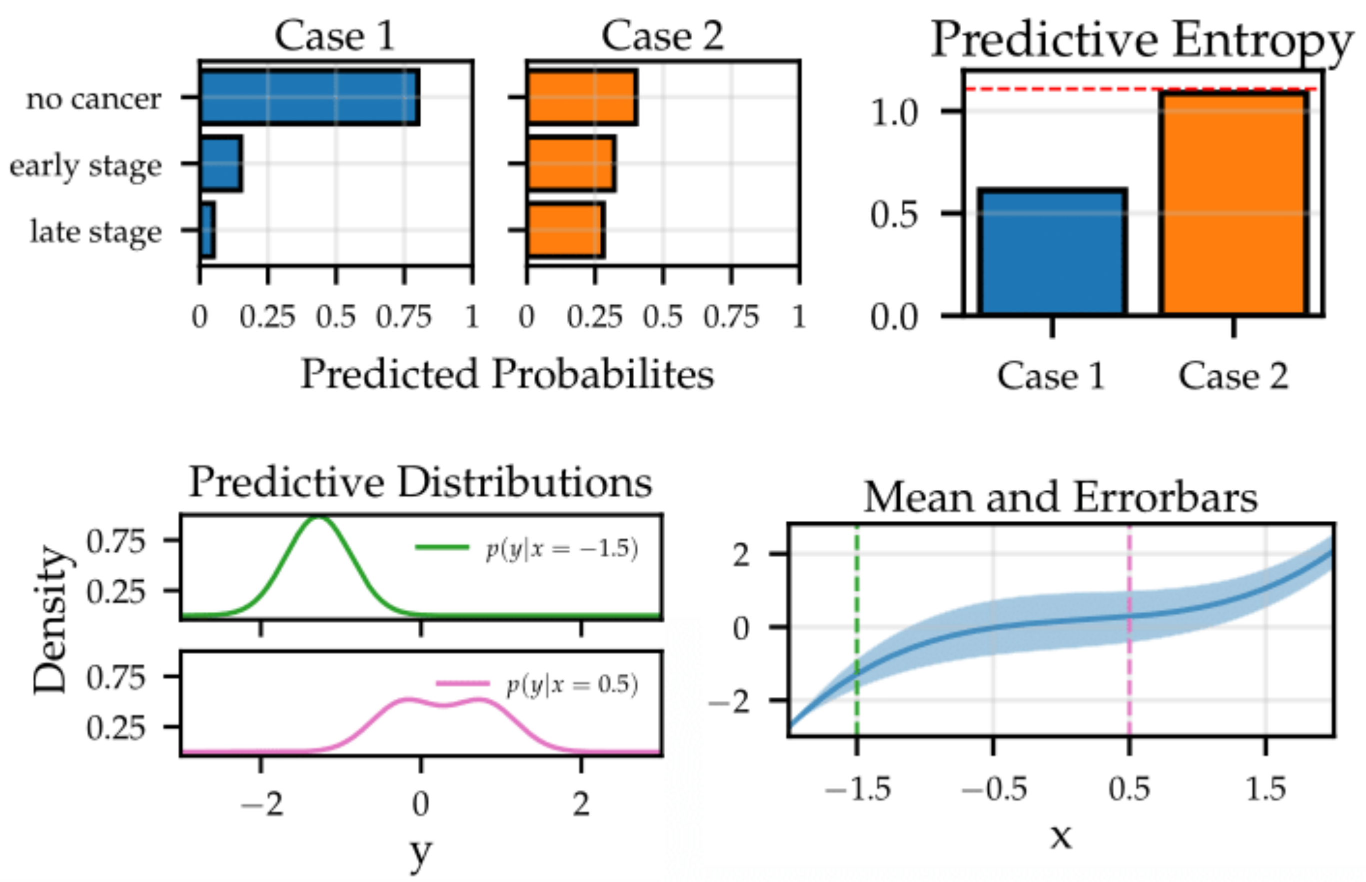
AIES 2021

Convening

Stakeholder



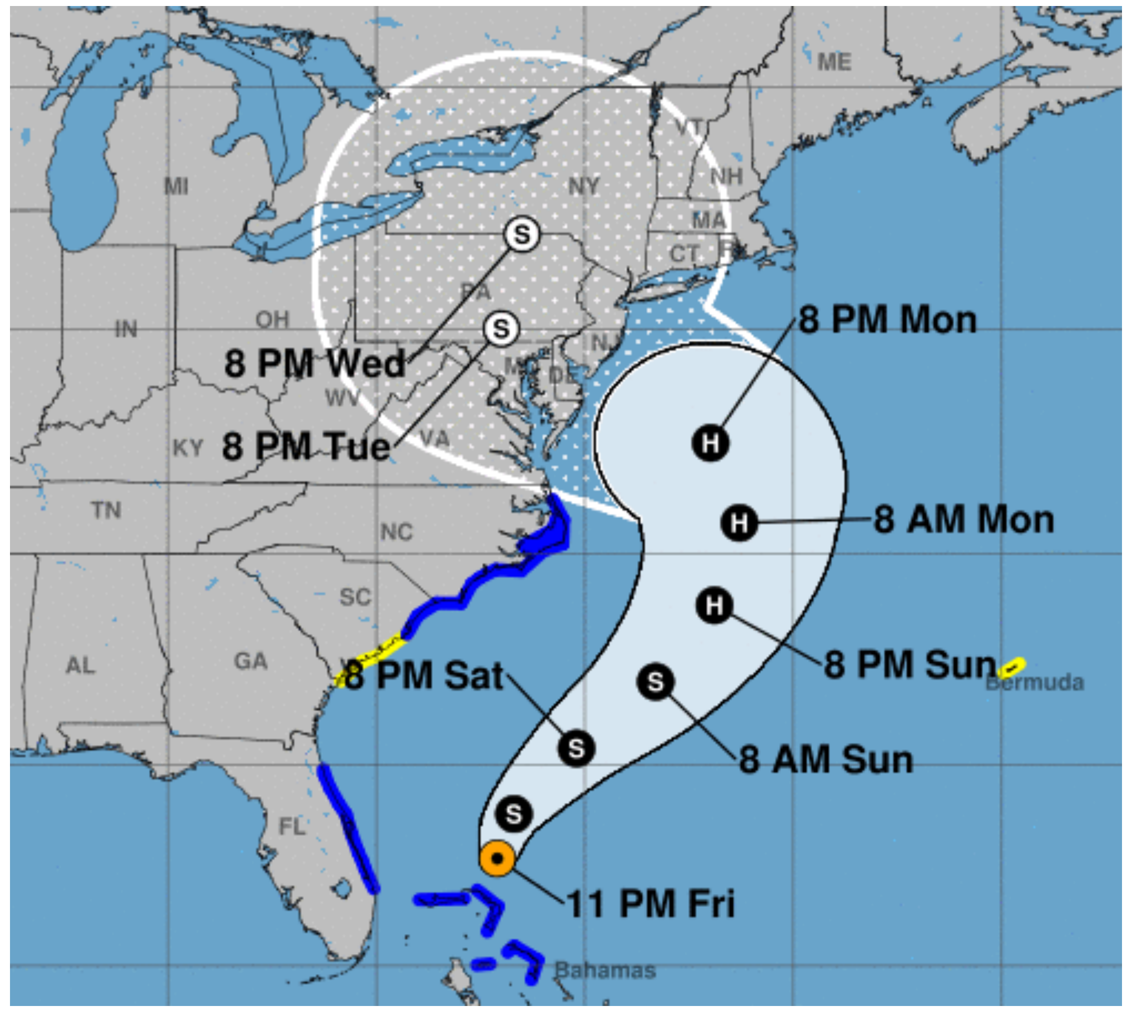
Step 1: Measuring



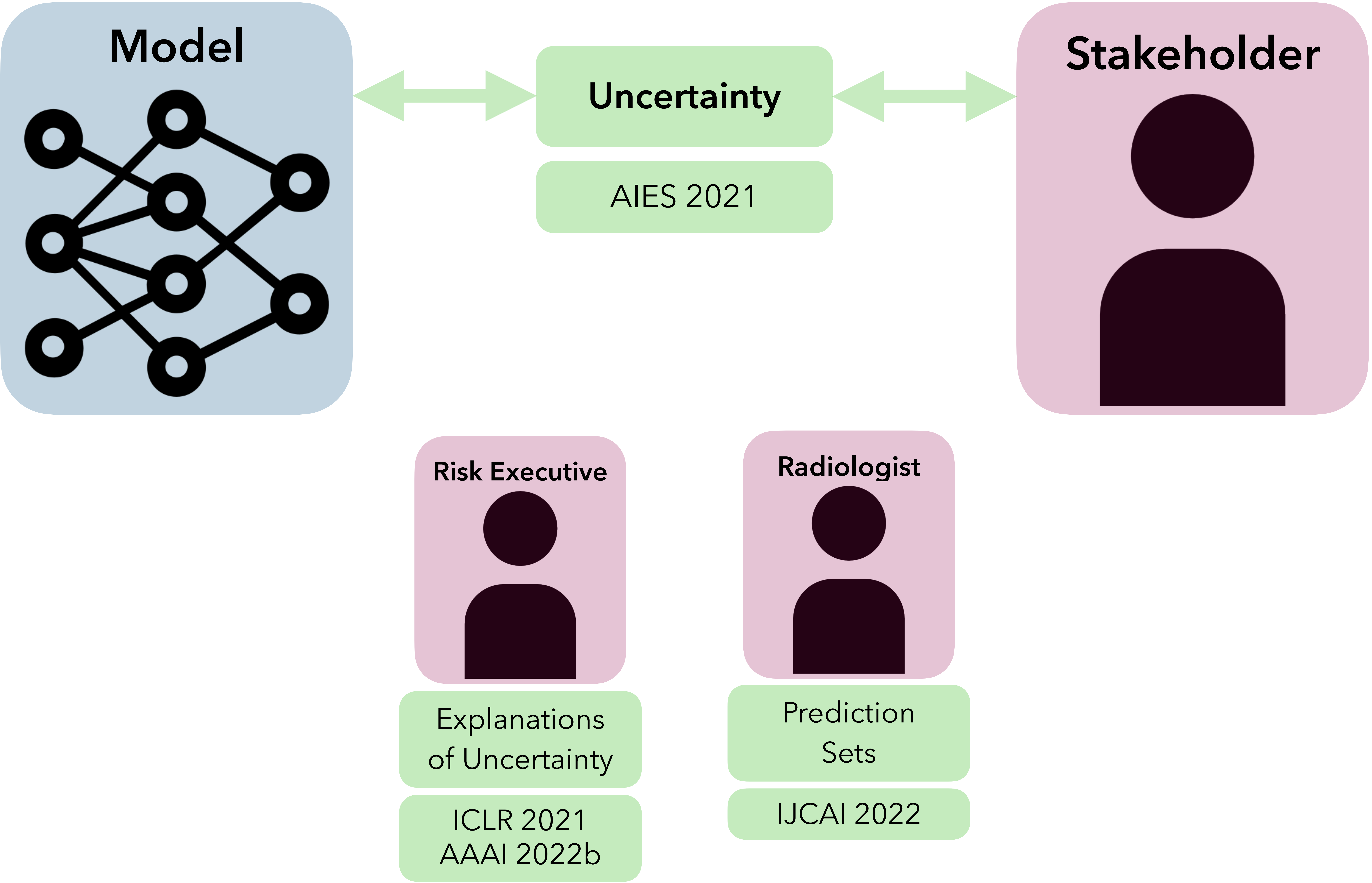
Step 2: Using

- **Fairness:** Measurement and Sampling Bias
- **Decision-Making:** Building Reject Option Classifiers
- **Trust Formation:** Displaying Ability, Benevolence, and Integrity

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.



Model
Uncertainty
AIES 2021
Stakeholder
Risk Executive
Explanations of Uncertainty
ICLR 2021
AAAI 2022b
Radiologist
Prediction Sets
IJCAI 2022

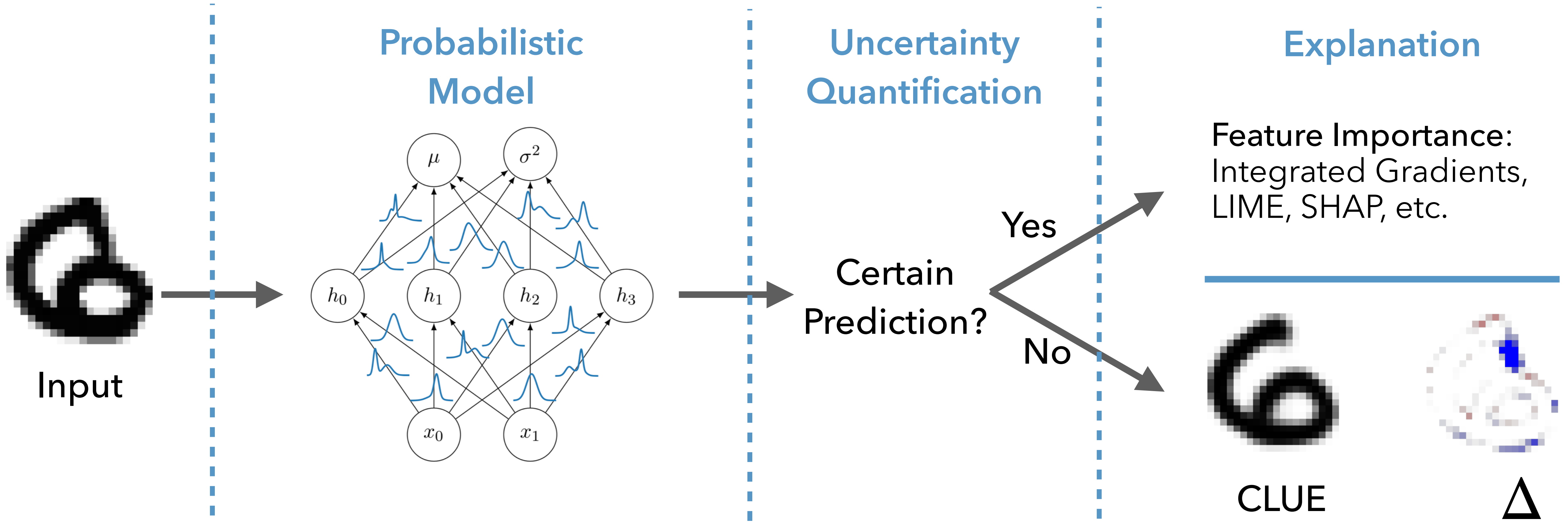
Risk Executive

Explanations of Uncertainty

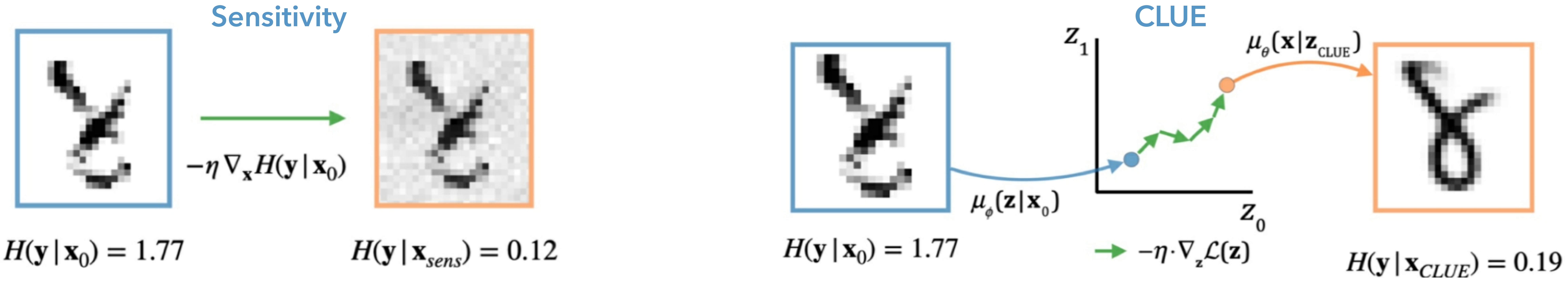
CLUE: Counterfactual Latent Uncertainty Explanations

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?*



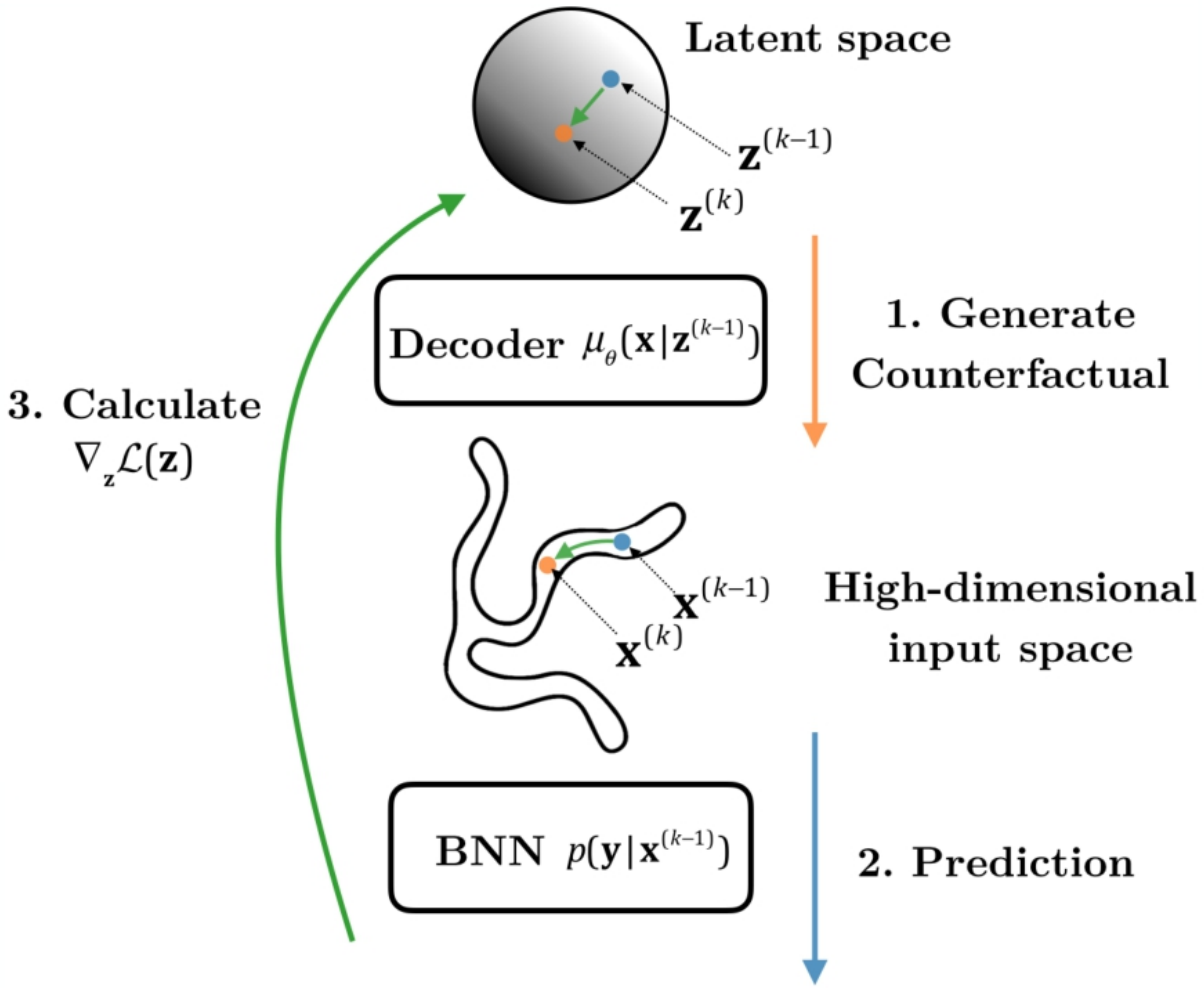
Antoran, **B**, Adel, Weller, Hernandez-Lobato. *Getting a CLUE: A Method for Explaining Uncertainty Estimates.* ICLR. 2021.
Ley, **B**, Weller. *Diverse and Amortised Counterfactual Explanations for Uncertainty Estimates.* AAAI. 2022.

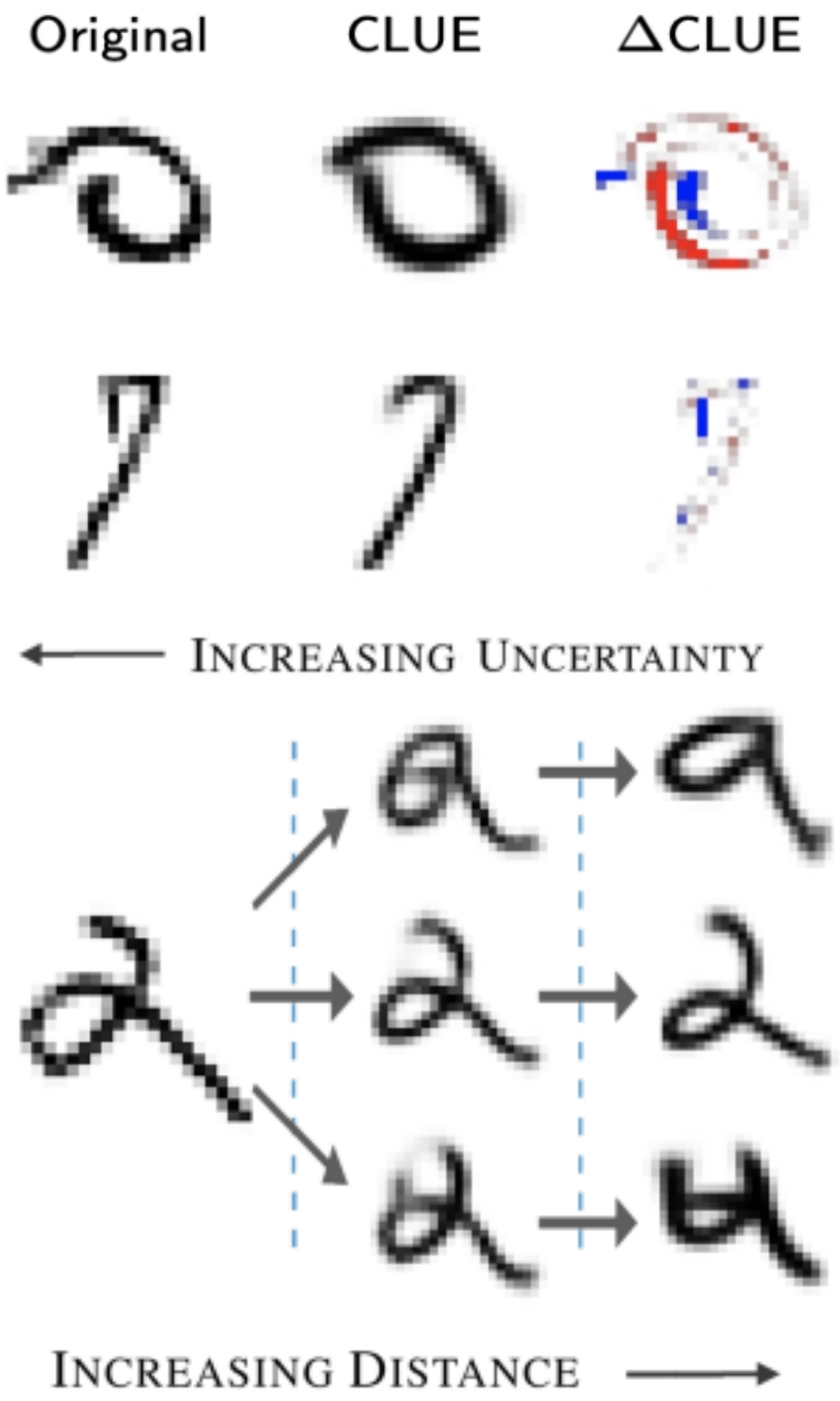
Risk Executive

Explanations of Uncertainty

CLUE: Counterfactual Latent Uncertainty Explanations

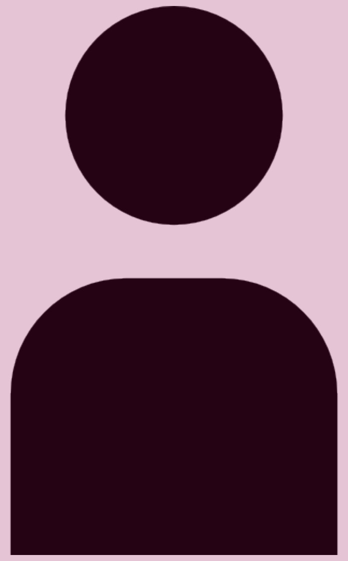
Methods





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.

Risk Executive



Explanations of Uncertainty

CLUE: Counterfactual Latent Uncertainty Explanations

User Studies

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

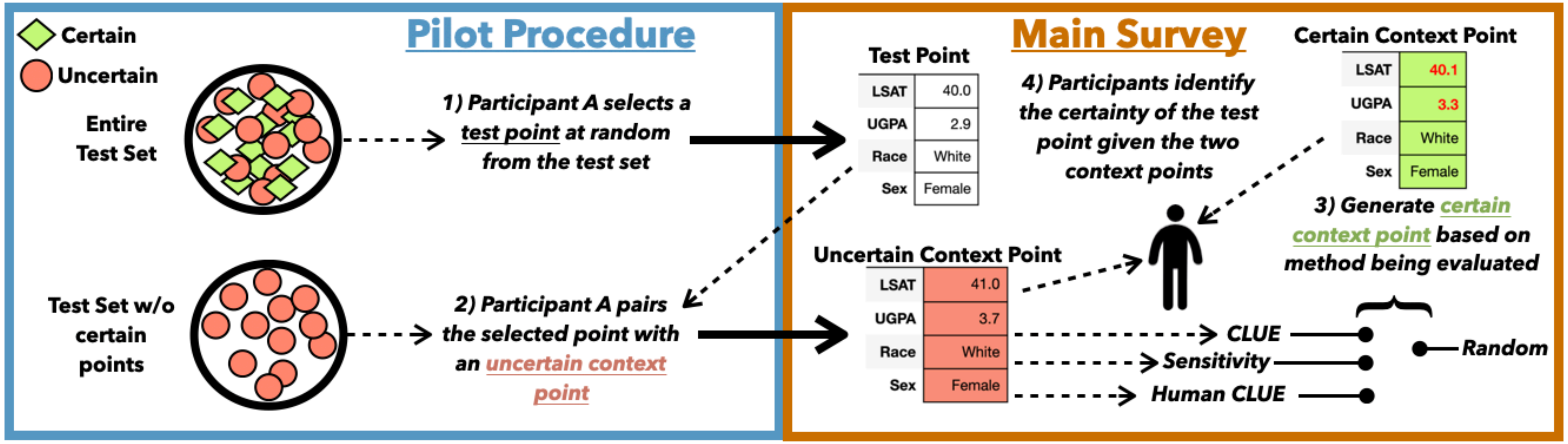
	Uncertain
Age	Less than 25
Race	Caucasian
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	Yes
Prior Convictions	1
Days Served	0

	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

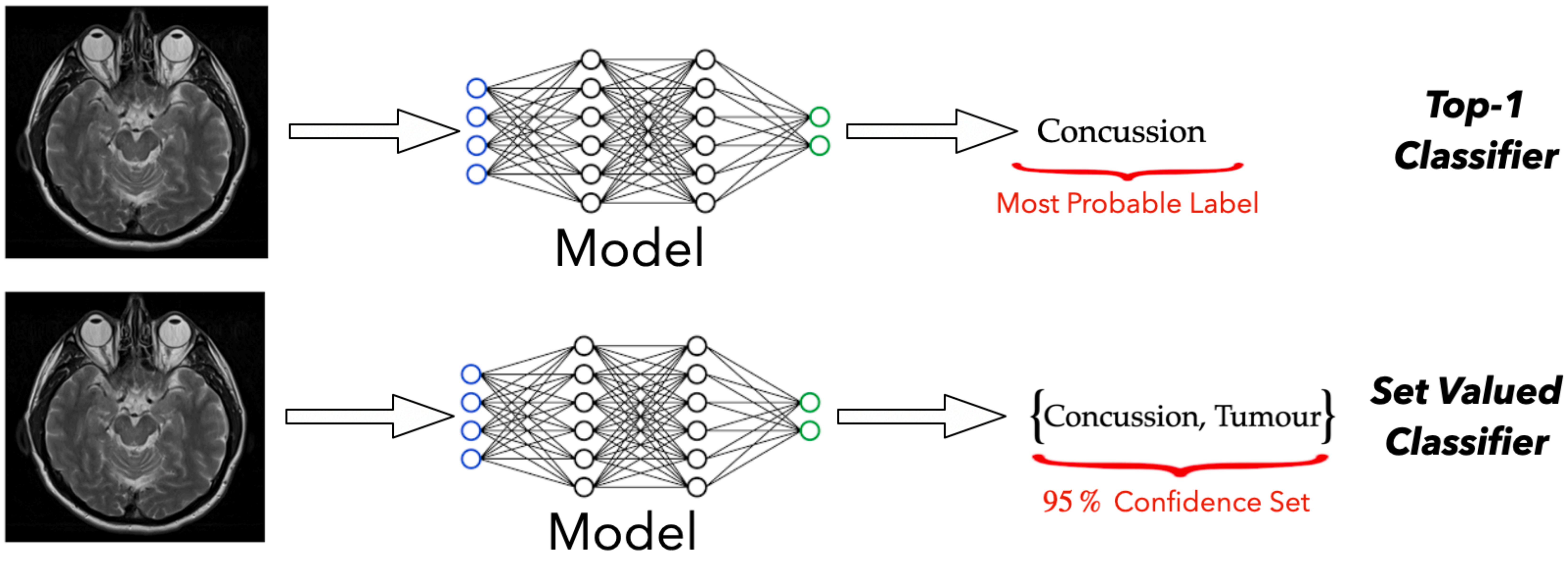
Prediction Sets

IJCAI 2022

Methods

Generate prediction sets for experts

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



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

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

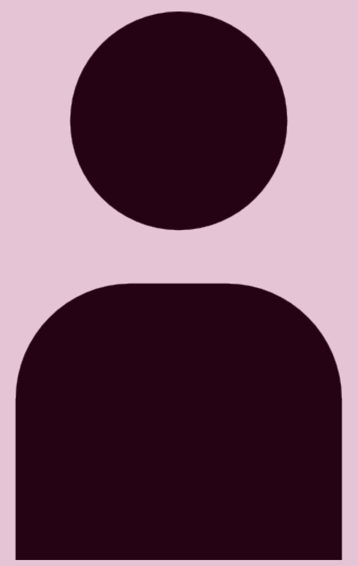
Risk Controlling Prediction Sets $P\Big(\underbrace{\mathbb{E}[L(y, \Gamma(x))]}_{\text{Risk}} \leq \alpha\Big) \geq 1 - \delta$

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.

Radiologist



Prediction Sets

IJCAI 2022

Generate prediction sets for experts

User Studies

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

A CP Scheme!

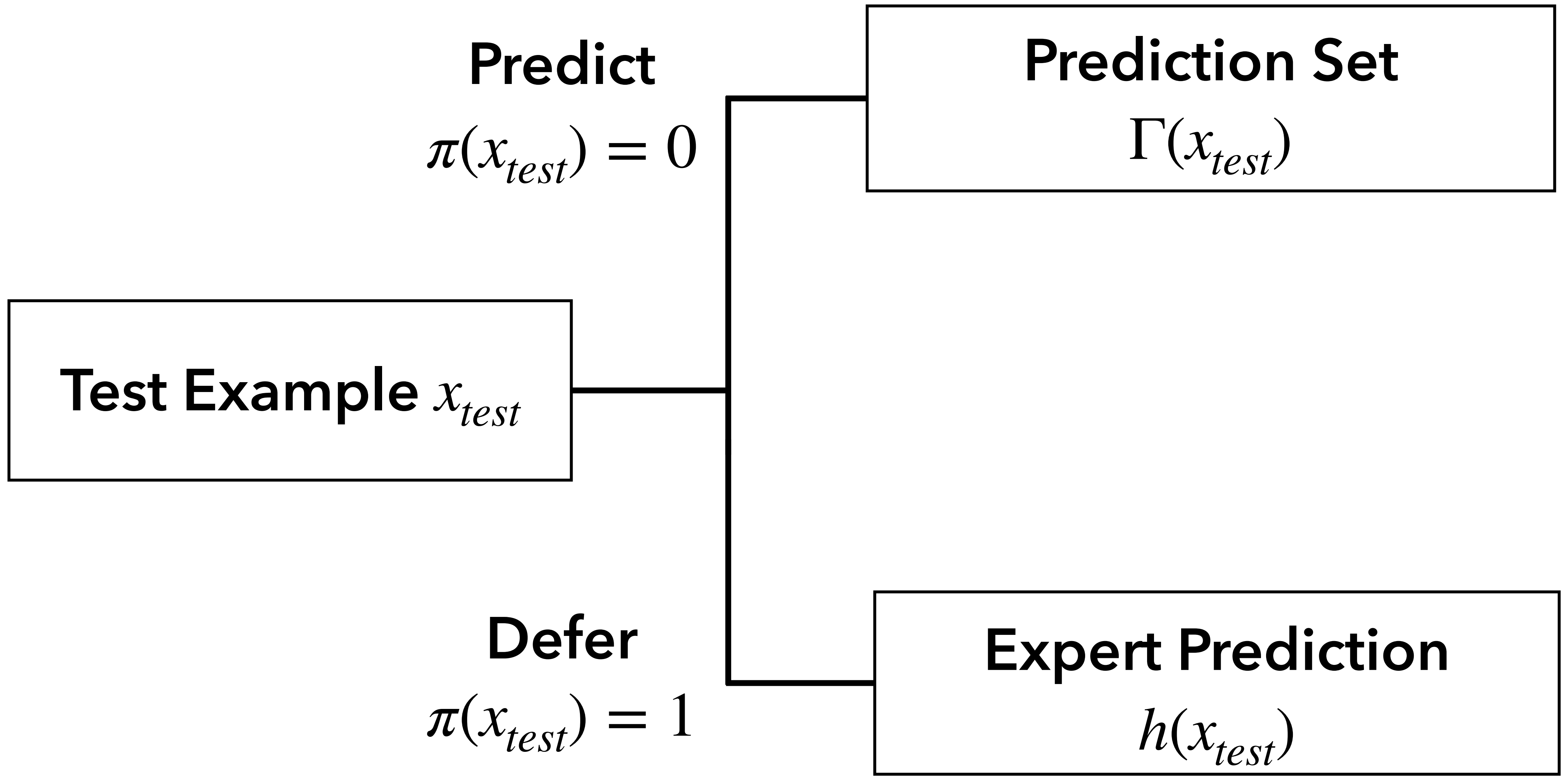
For CIFAR-100:

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	$< \mathbf{0.001}$	1.487

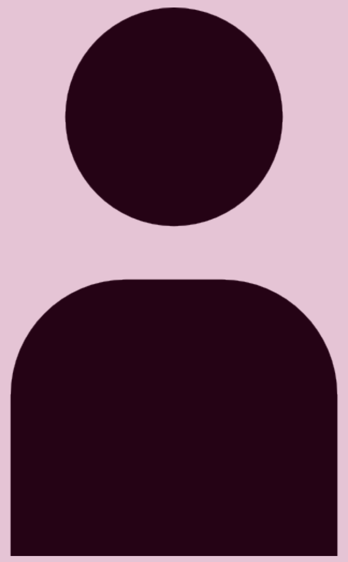
Observation: *Some prediction sets can be quite large, rendering them useless to experts!*

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).

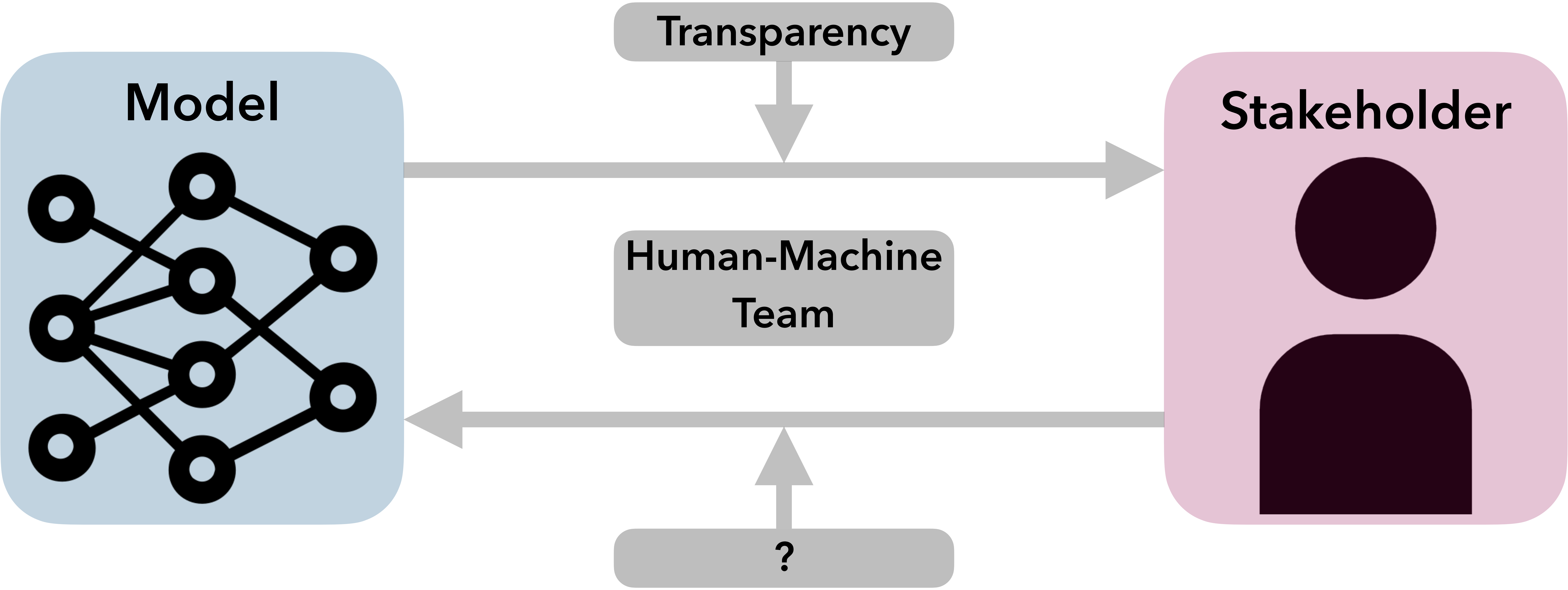
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.

Some Takeaways Thus Far

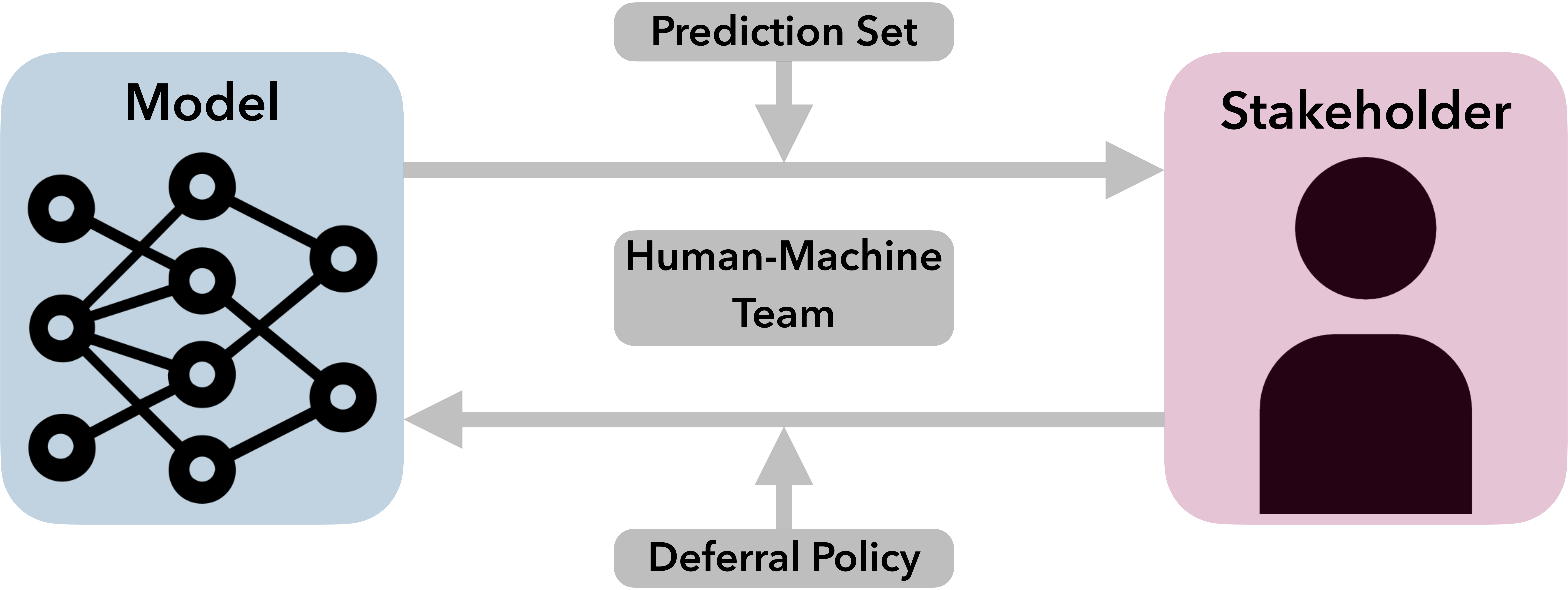
Algorithmic transparency is important but difficult

- Explanations are desirable in theory but are hard to operationalize
- Uncertainty can be treated as a form of transparency that can be used to alter stakeholder interaction with model
- We need to consider the context of transparency carefully to improve outcomes of human-machine teams

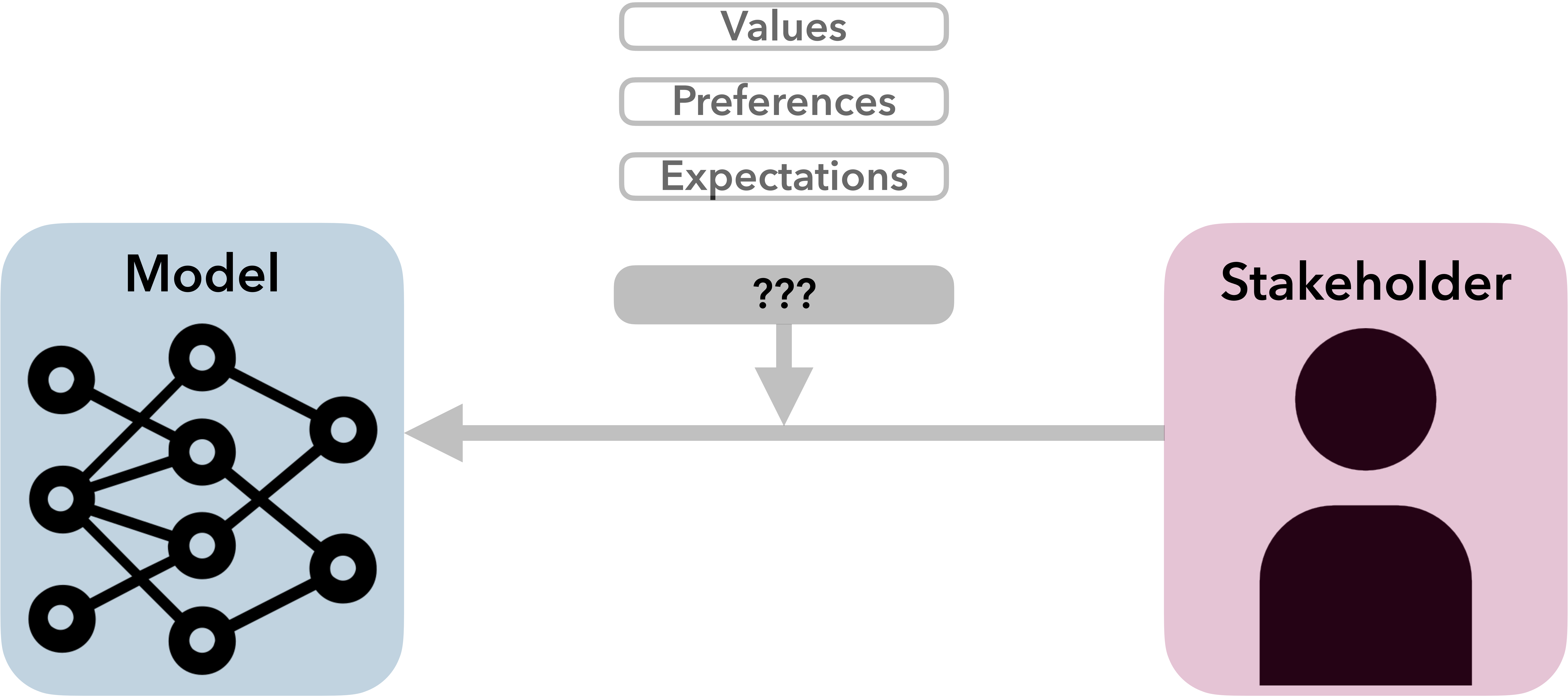
Convening is powerful tool to motivate technical and socio-technical research



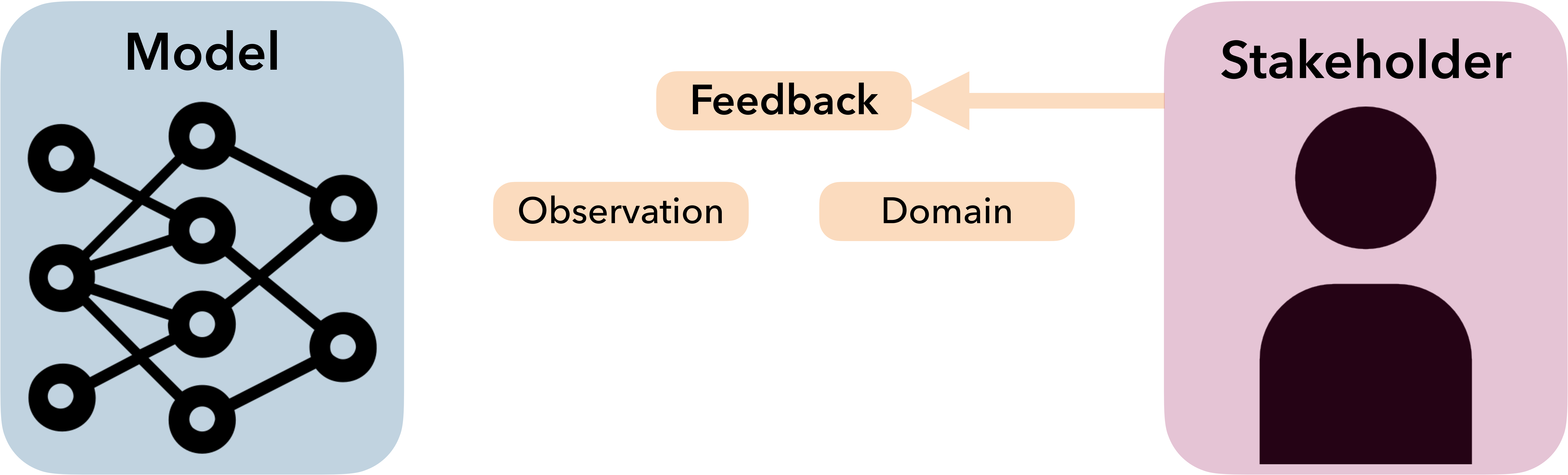
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams.* IJCAI. 2022.
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* Patterns. 2023.



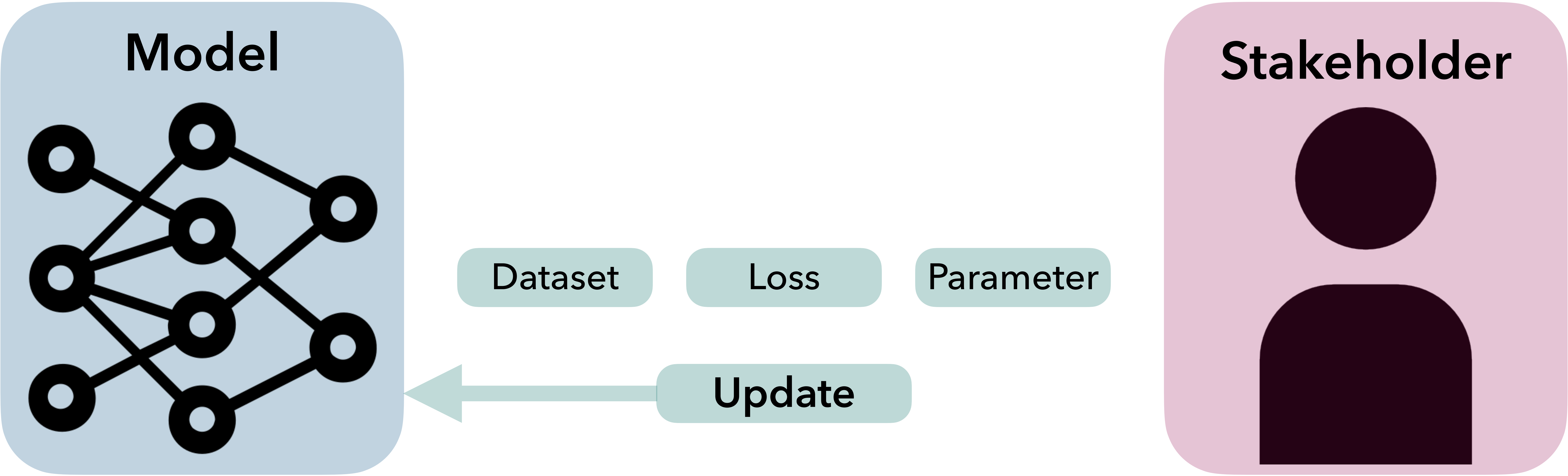
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.



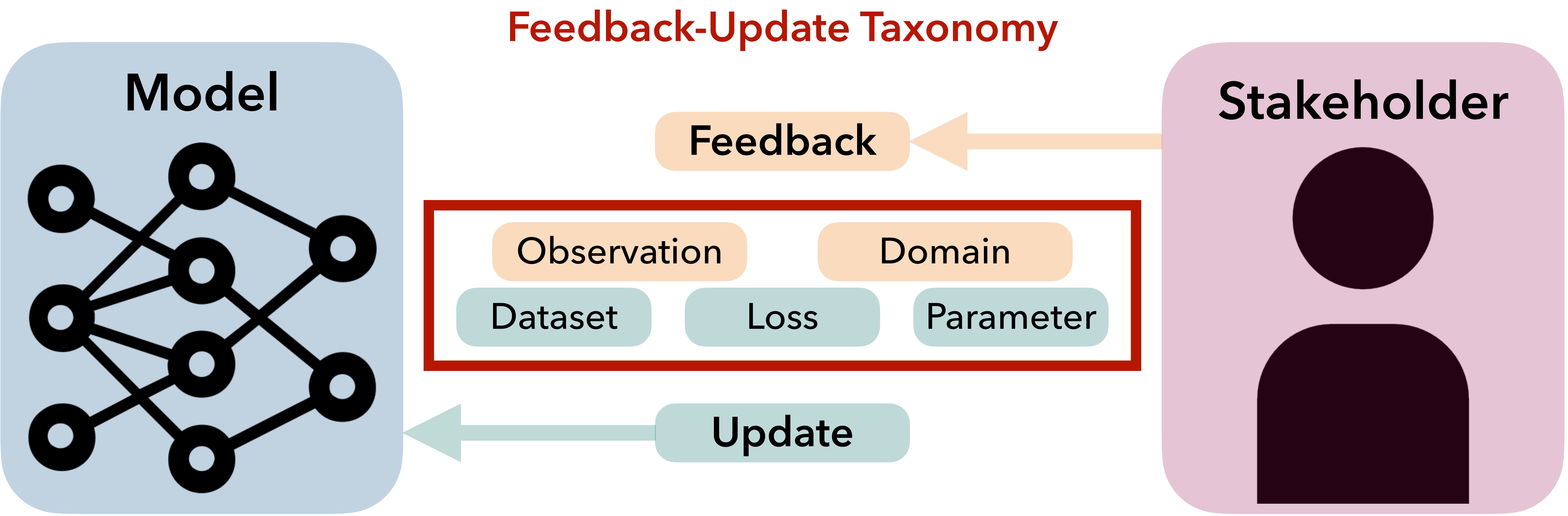
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. Patterns. 2023.



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.

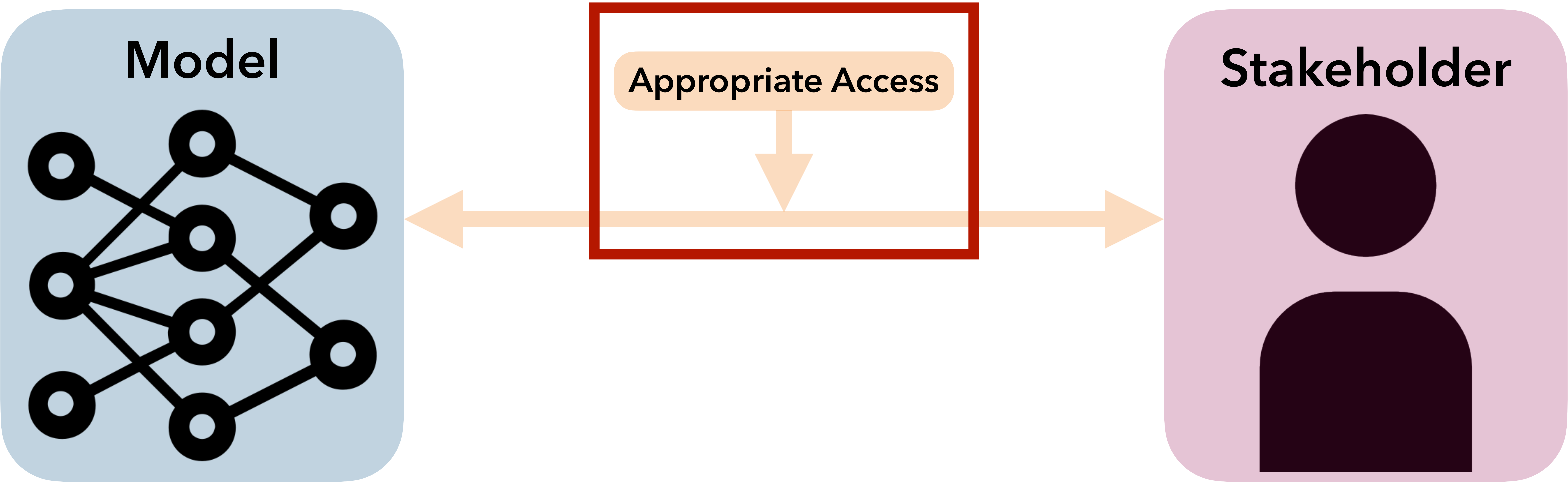


Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. Patterns. 2023.

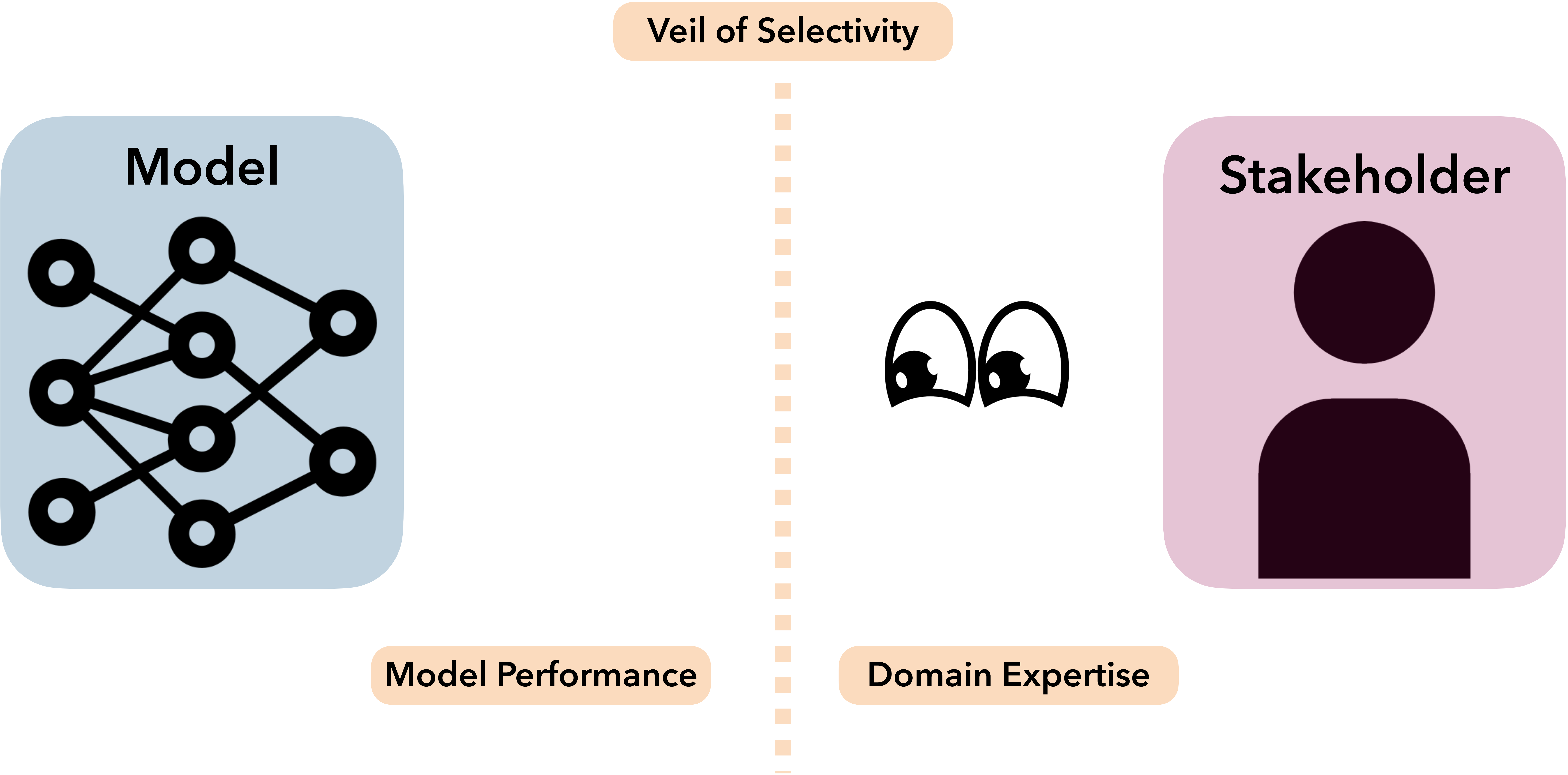


Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. Patterns. 2023.

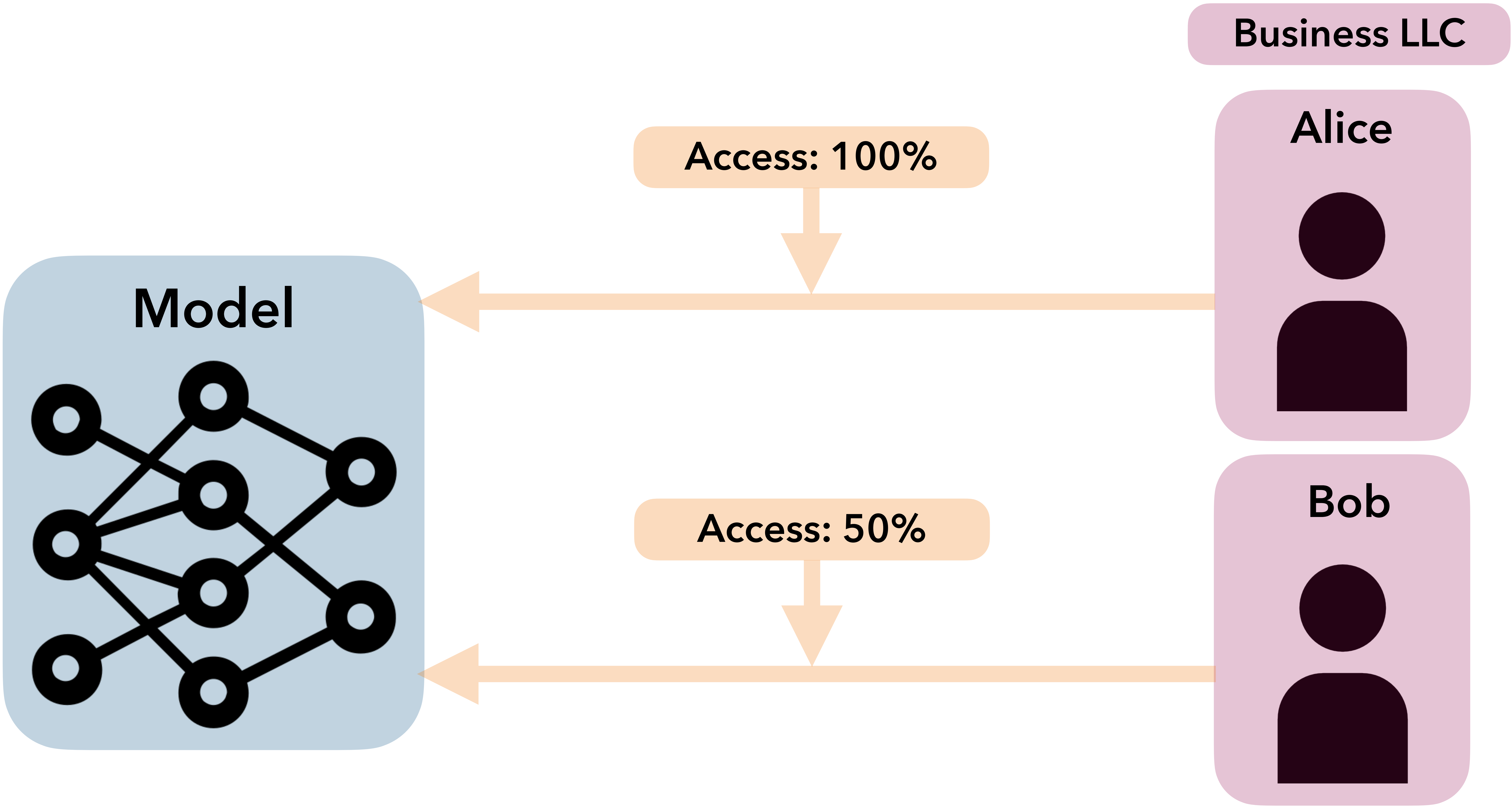
Socio-technical Relationship



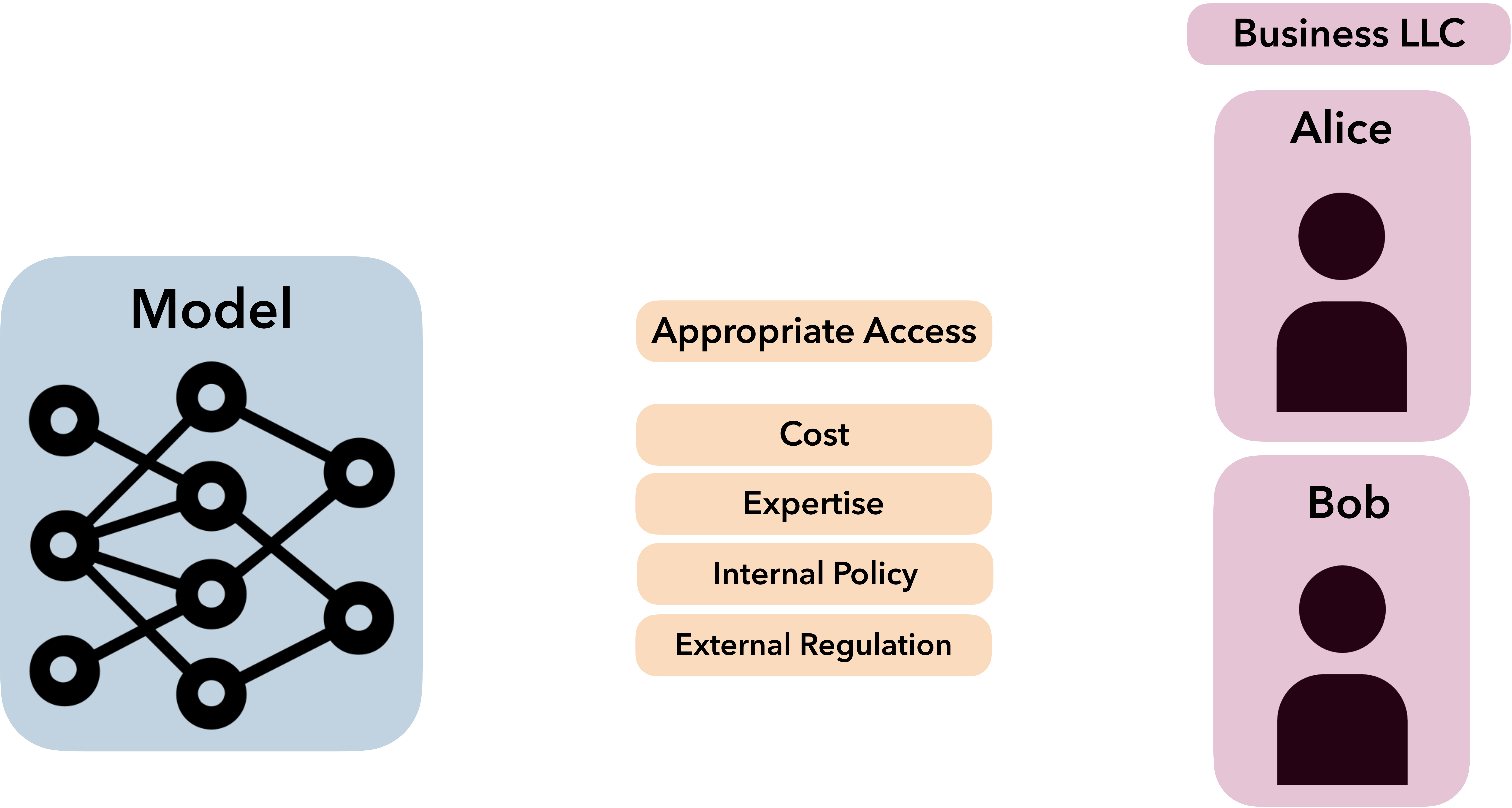
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. Patterns. Cell Press 2023.
B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.



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



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



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

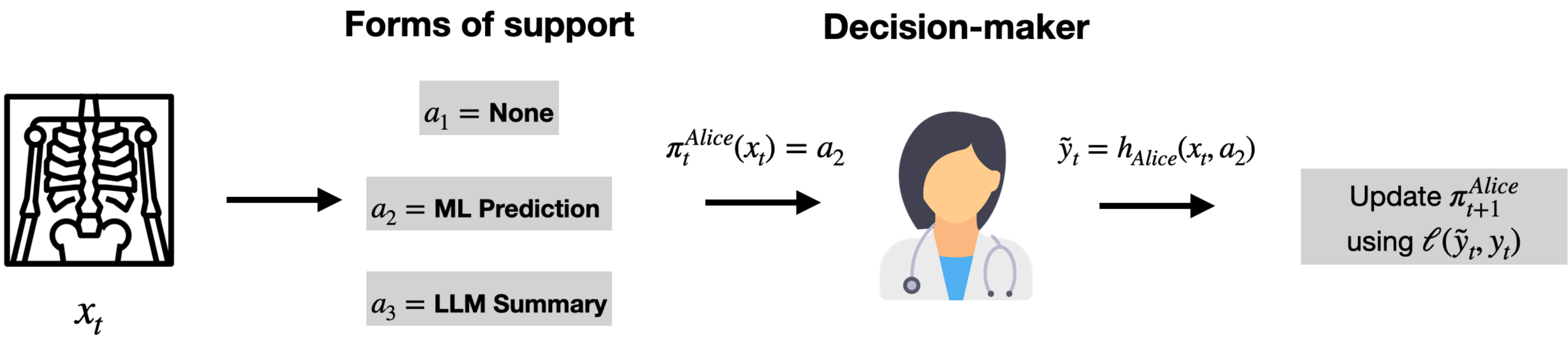
Decision Maker

Personalize Access

Methods

Learning Personalized Decision Support Policies

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



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 $\pi_t : X \to \Delta(A)$

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.

Decision Maker

Personalize Access

Learning Personalized Decision Support Policies

Methods

Expertise Profiles

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

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

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

CIFAR10 Task: *3 forms of support (None, Model, or Expert Consensus) and 5 classes*

MMLU Task: *2 forms of support (None or LLM) and 4 categories*

CIFAR

Excess loss over optimal loss

Algorithm	Invariant	Strictly Better	Varying
H-ONLY	0.00 ± 0.01	0.09 ± 0.08	0.50 ± 0.06
H-MODEL	0.00 ± 0.01	0.22 ± 0.19	0.35 ± 0.05
H-CONSENSUS	0.00 ± 0.01	0.23 ± 0.13	0.27 ± 0.08
Population	0.00 ± 0.02	0.18 ± 0.08	0.15 ± 0.03
THREAD-LinUCB	0.00 ± 0.01	0.17 ± 0.05	0.19 ± 0.05
THREAD-KNN	0.00 ± 0.01	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.08 \pm 0.02}$

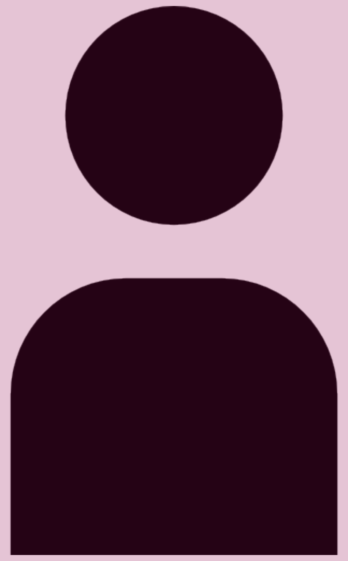
MMLU

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.

Decision Maker

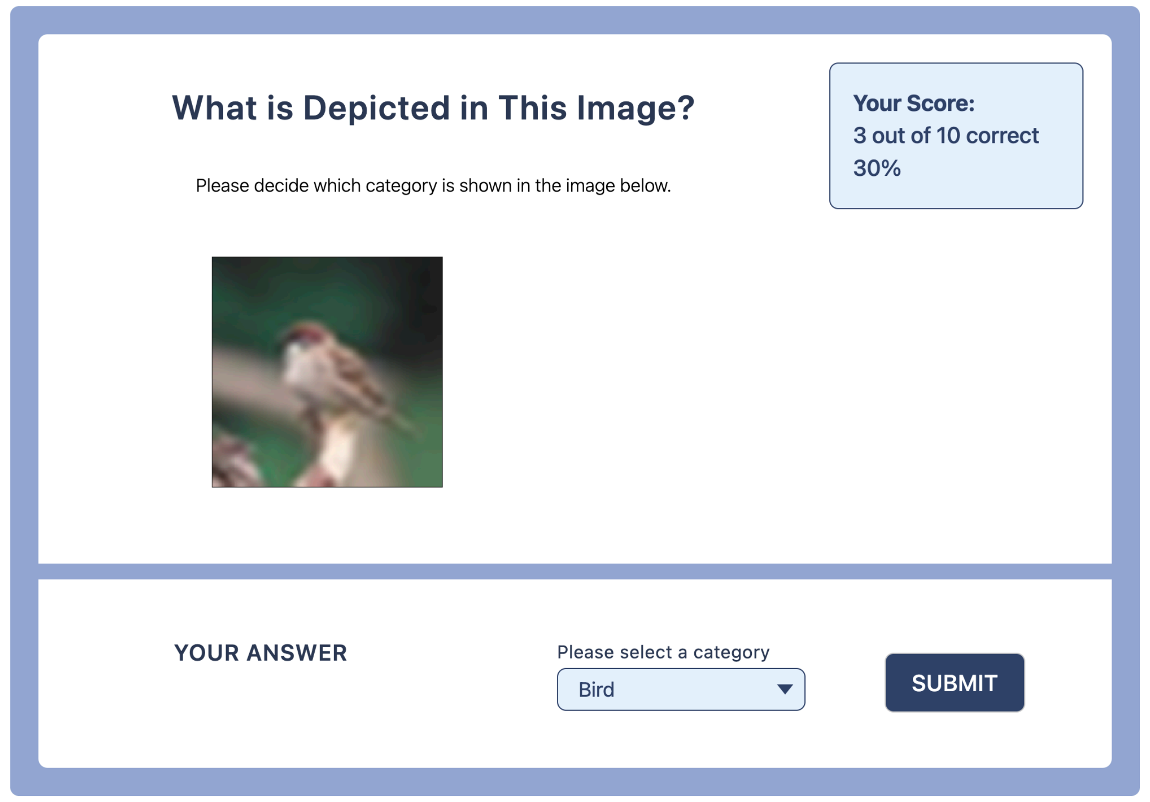


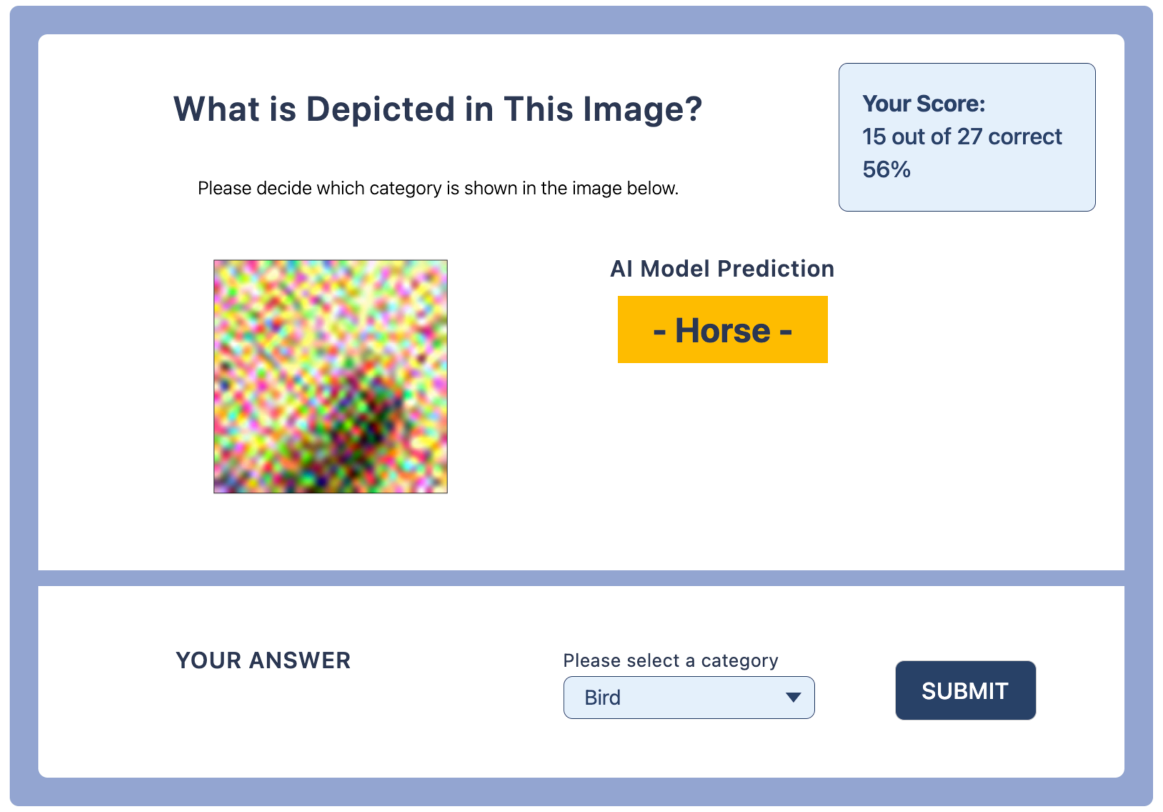
Personalize Access

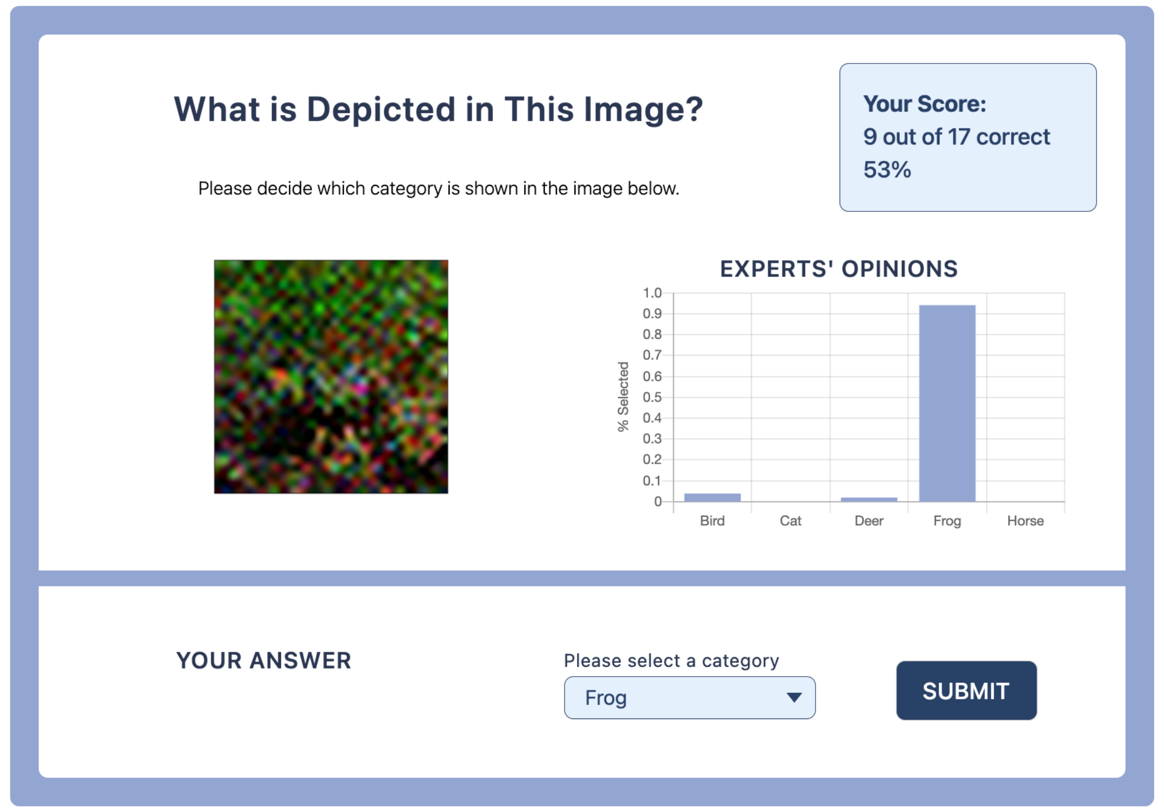
Learning Personalized Decision Support Policies

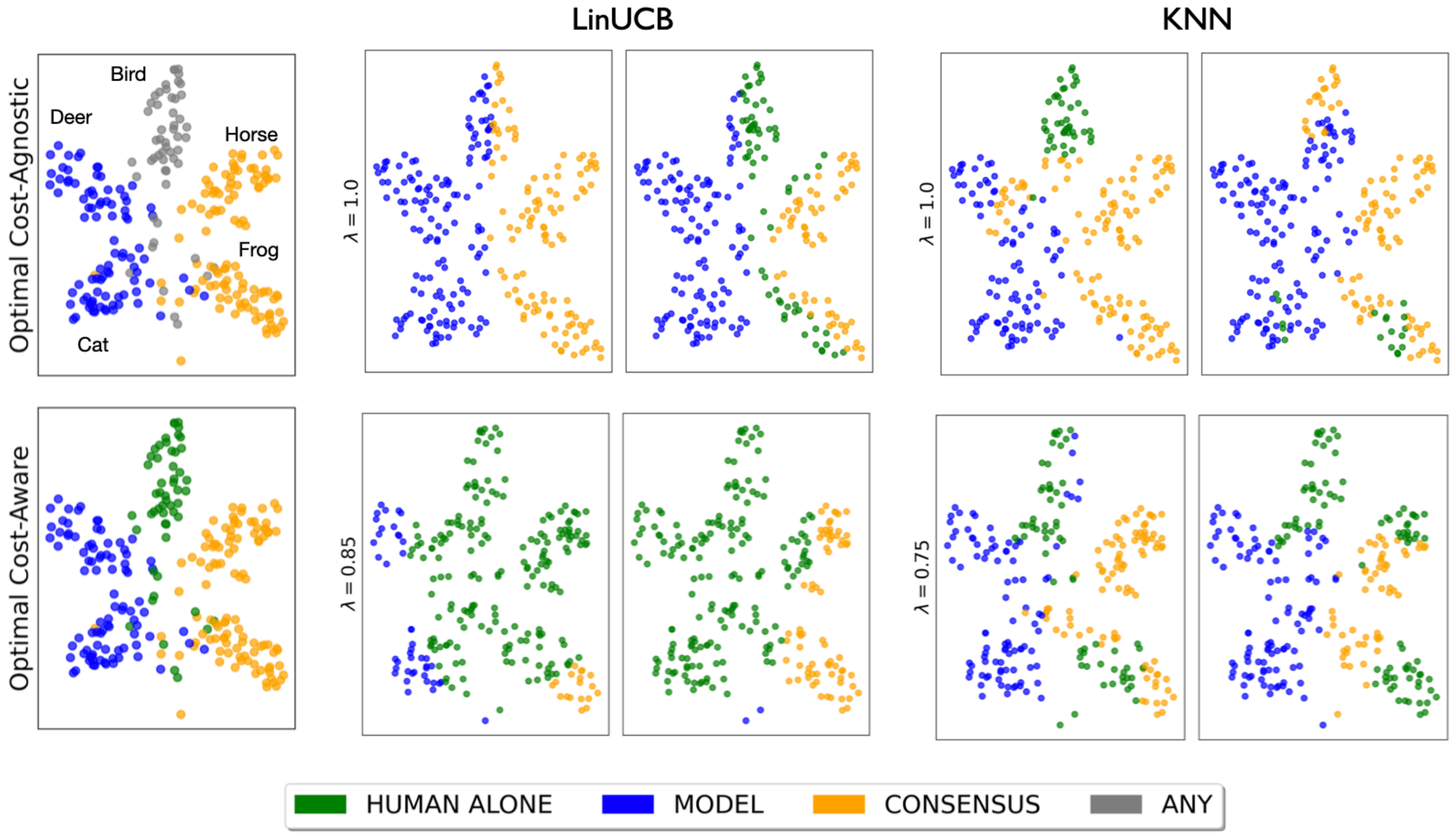
User Studies

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.

Decision Maker

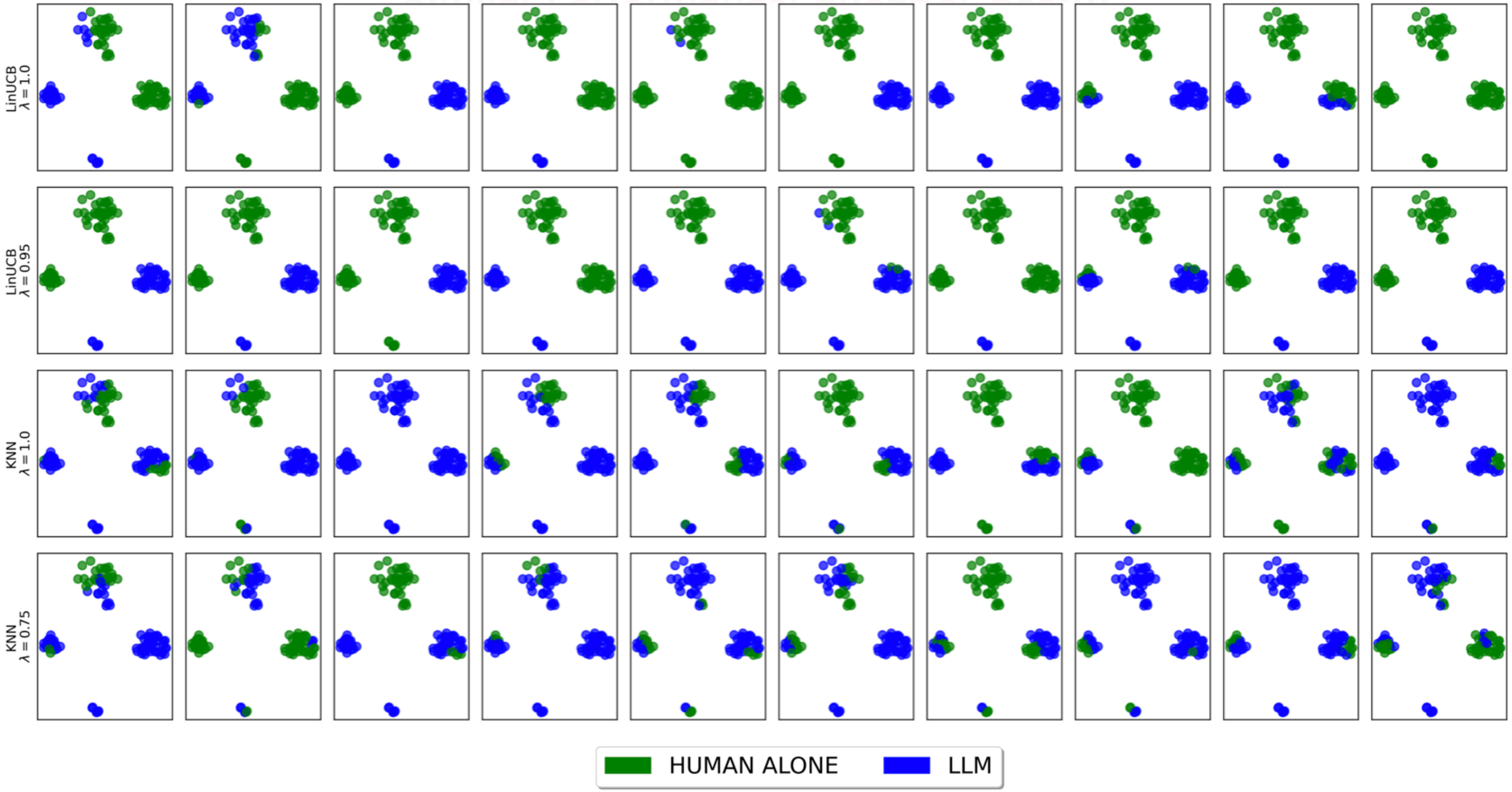
Personalize Access

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!!!

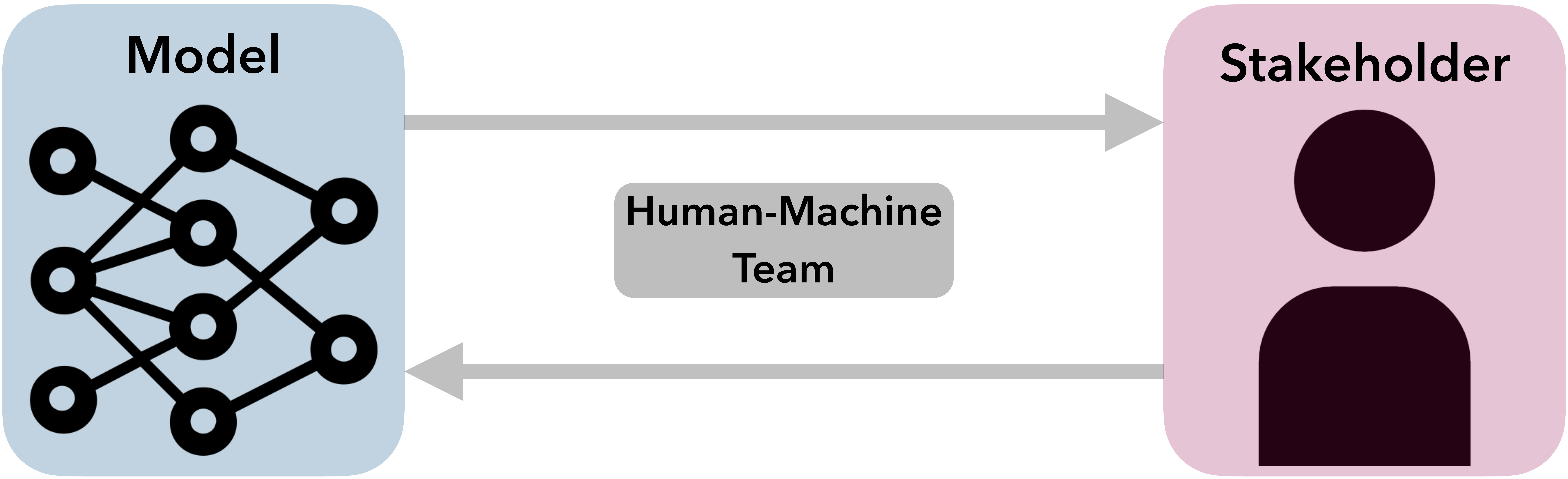


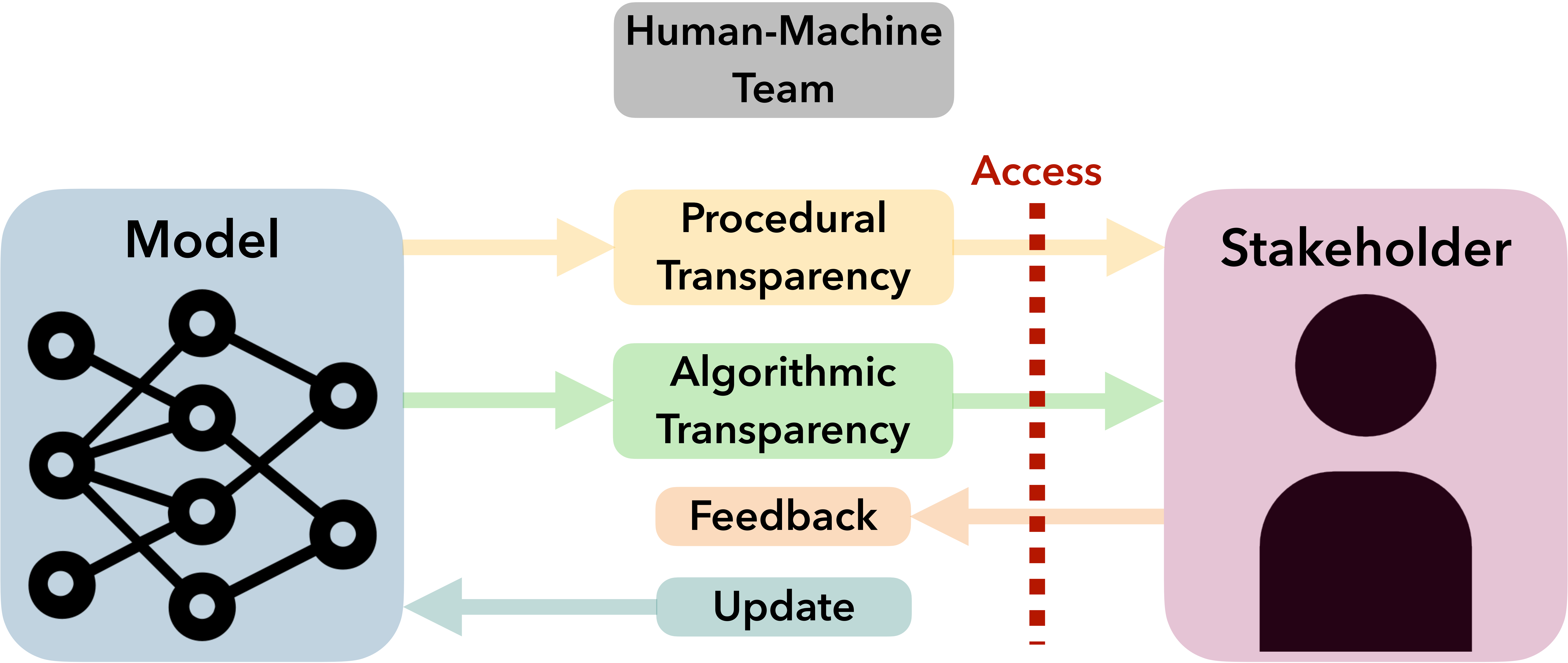
B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies*. Under Review. 2023.

Additional Takeaways

Personalized access to decision support (e.g., ML models) can be learned and improve decision-maker performance

- Forms of decision support may be offline (e.g., expert consensus)
- Selectivity is just one way to operationalize stakeholder-model interaction and to preempt aversive behavior
- Testbeds (a la Modiste) can validate online learning algorithms in practice


Model
Human-Machine
Team
Stakeholder



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

Future Directions

- Show selective access to models helps in deployed settings: this may mean selective transparency based on stakeholder expertise
- Study the socio-technical nature and societal implications of providing model predictions and subsequent transparency in specific contexts
- Leverage stronger priors in learning when decision-makers should be and want to be supported

Thank you to all my collaborators, mentors, and students!

Computer Science


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
Amazon


Ruchir Puri
IBM


Yunfeng Zhang
Twitter


Vera Liao
Microsoft


Ankur Taly
Google

Psychology


Bradley Love
UCL


Simone Schnall
Cambridge

Law


John Zerilli
Edinburgh


Alice Xiang
Sony AI


Madhu Srikumar
PAI

Design


Emma Kallina
Cambridge


Becca Ricks
Mozilla


Dorian Peters
Imperial

Philosophy


Rune Nyrup
Aarhus


Stephen Cave
Cambridge

Trustworthy Machine Learning

From Algorithmic Transparency to Decision Support

Thank you for listening! Questions?

@umangsbhatt

umangbhatt@nyu.edu