Algorithmic Transparency in Machine Learning

Umang Bhatt
PhD Candidate, University of Cambridge
Enrichment Student, The Alan Turing Institute
Student Fellow, Leverhulme Center for the Future of Intelligence
Research Fellow, Harvard Center for Research on Computation and Society

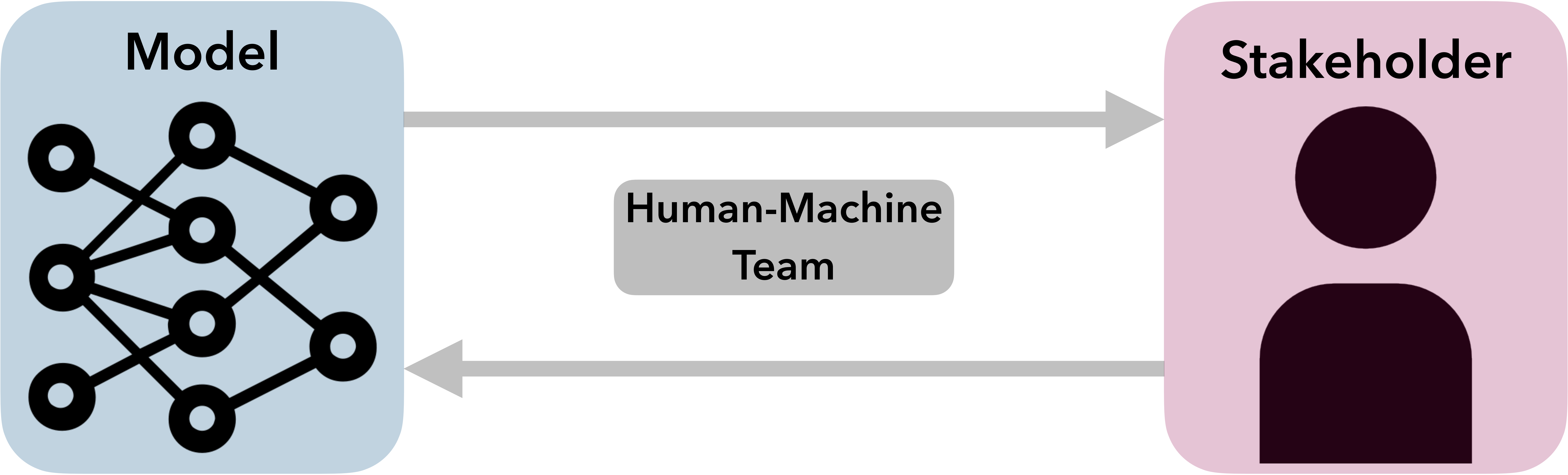
@umangsbhatt
usb20@cam.ac.uk

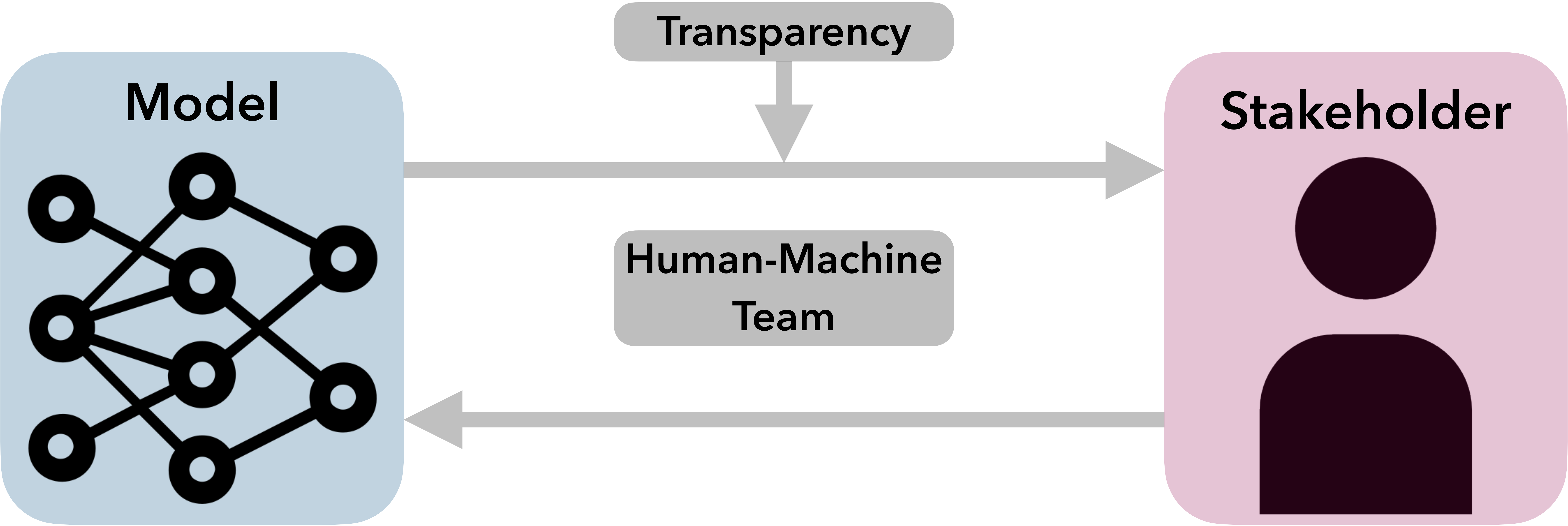
UNIVERSITY OF CAMBRIDGE

The Alan Turing Institute

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

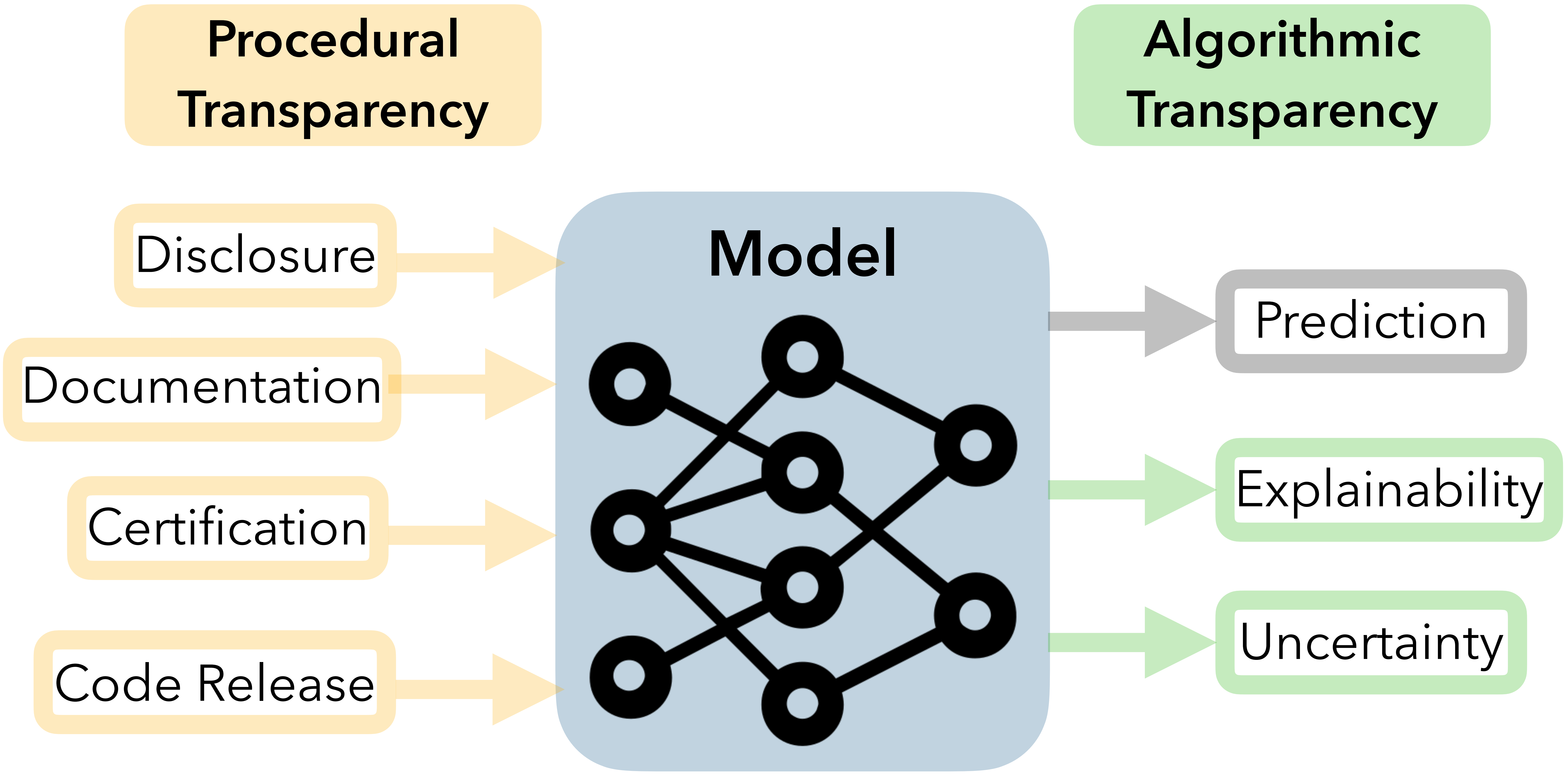
CRCS


Model
Human-Machine
Team
Stakeholder

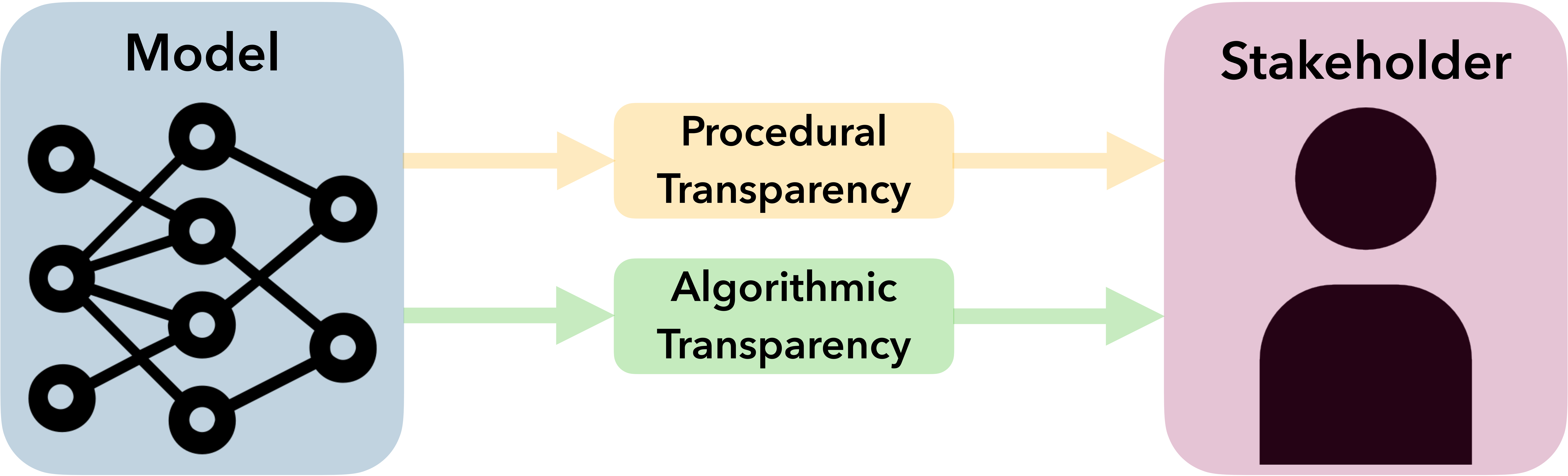


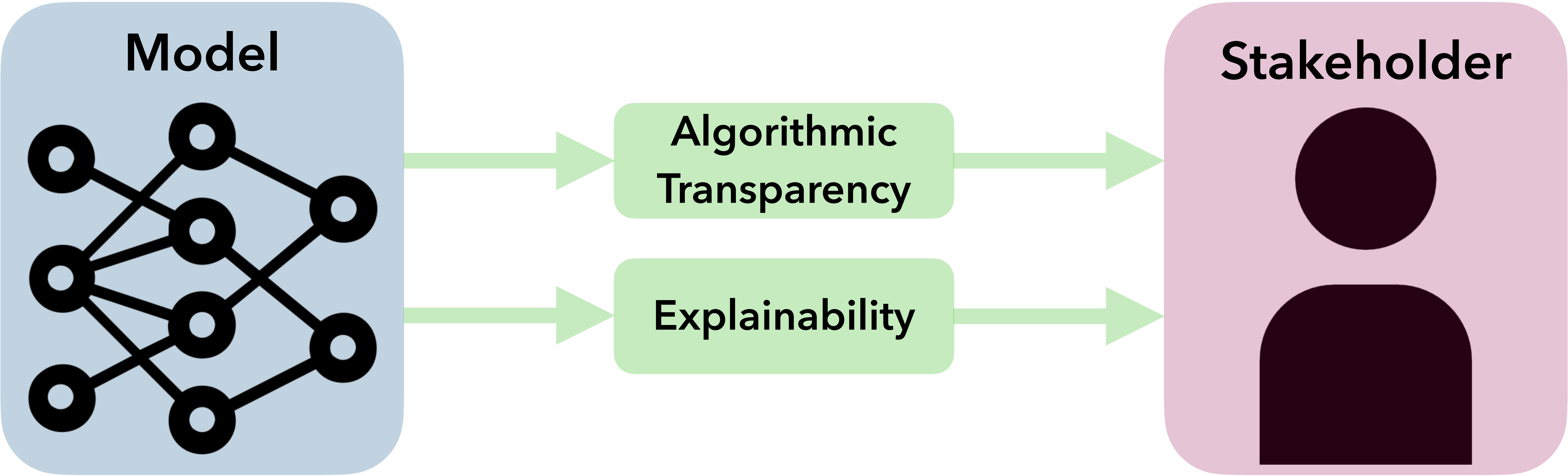
Transparency means providing stakeholders with *relevant* information about how a model works

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)

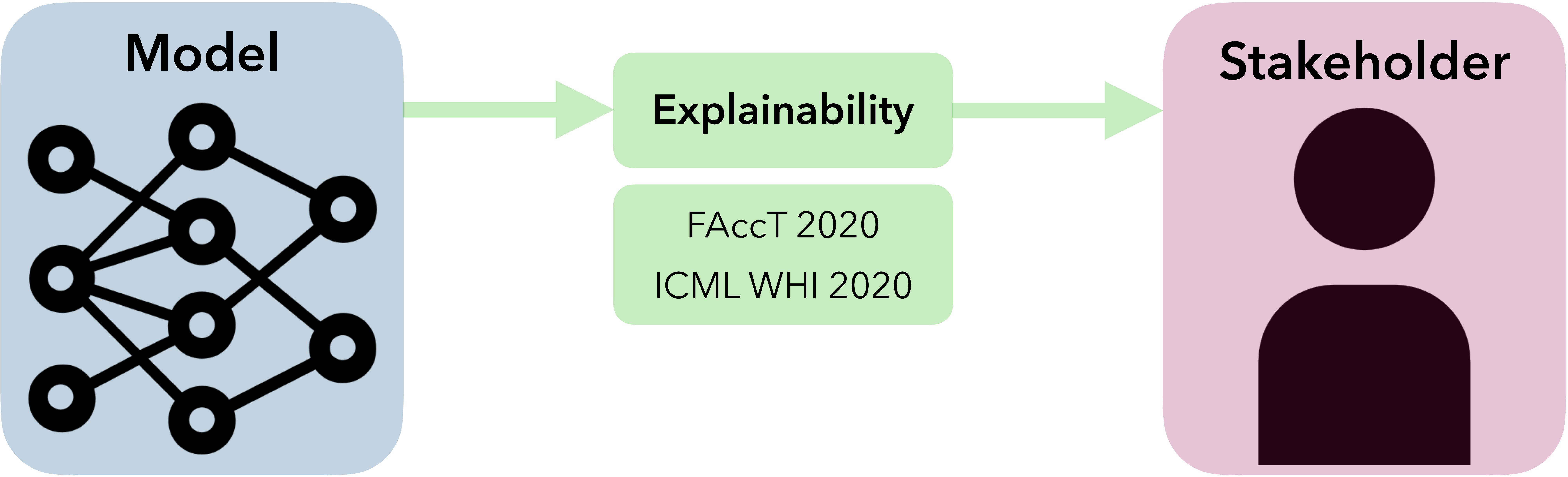
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

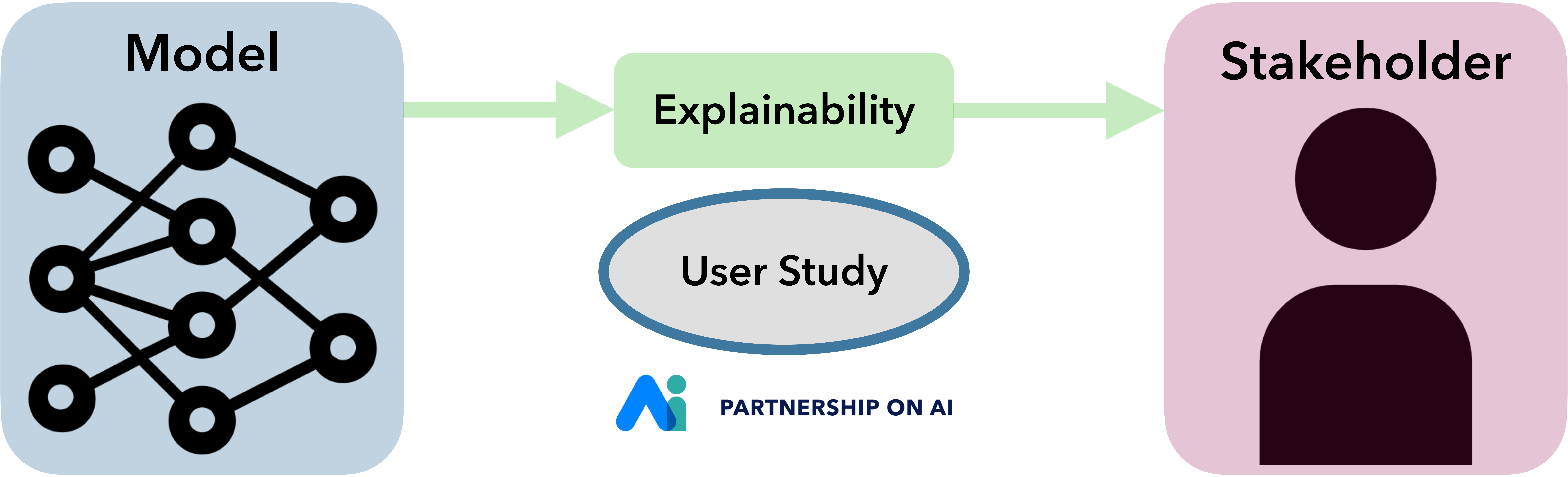
Research Style

Convenings

Methods

User Studies


Model
Explainability
FAccT 2020
ICML WHI 2020
Stakeholder

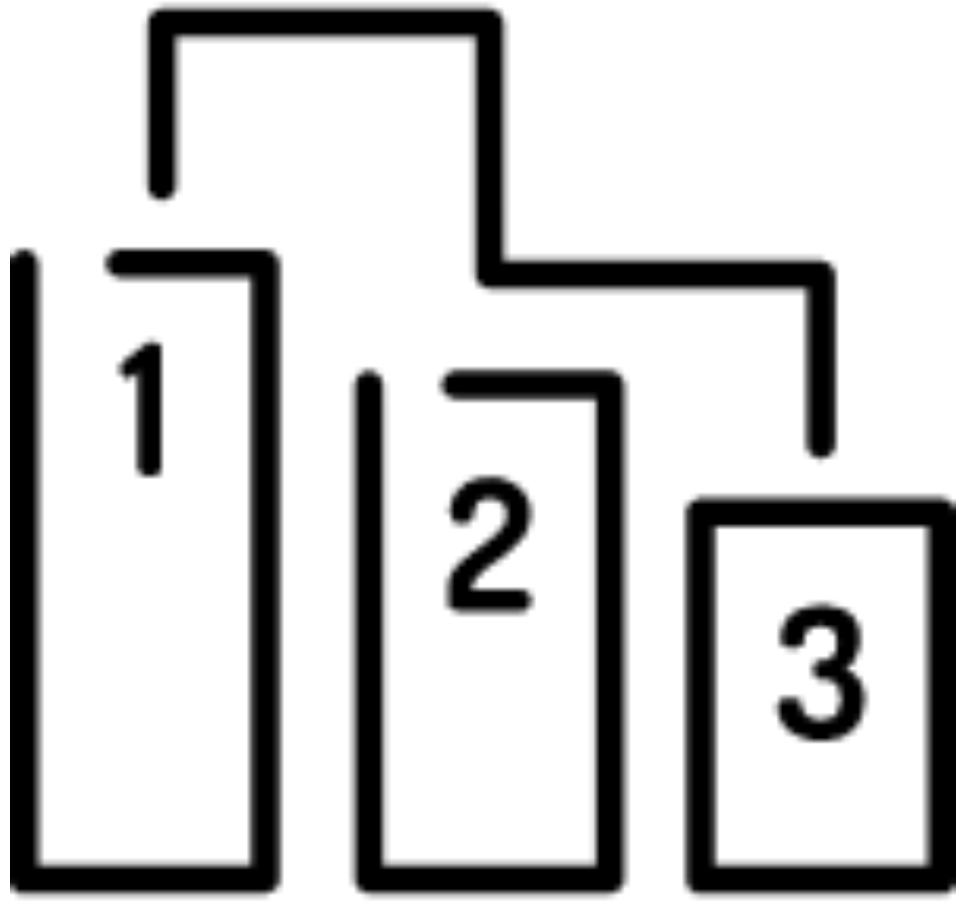


Goal: understand how explainability methods are used in *practice*

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

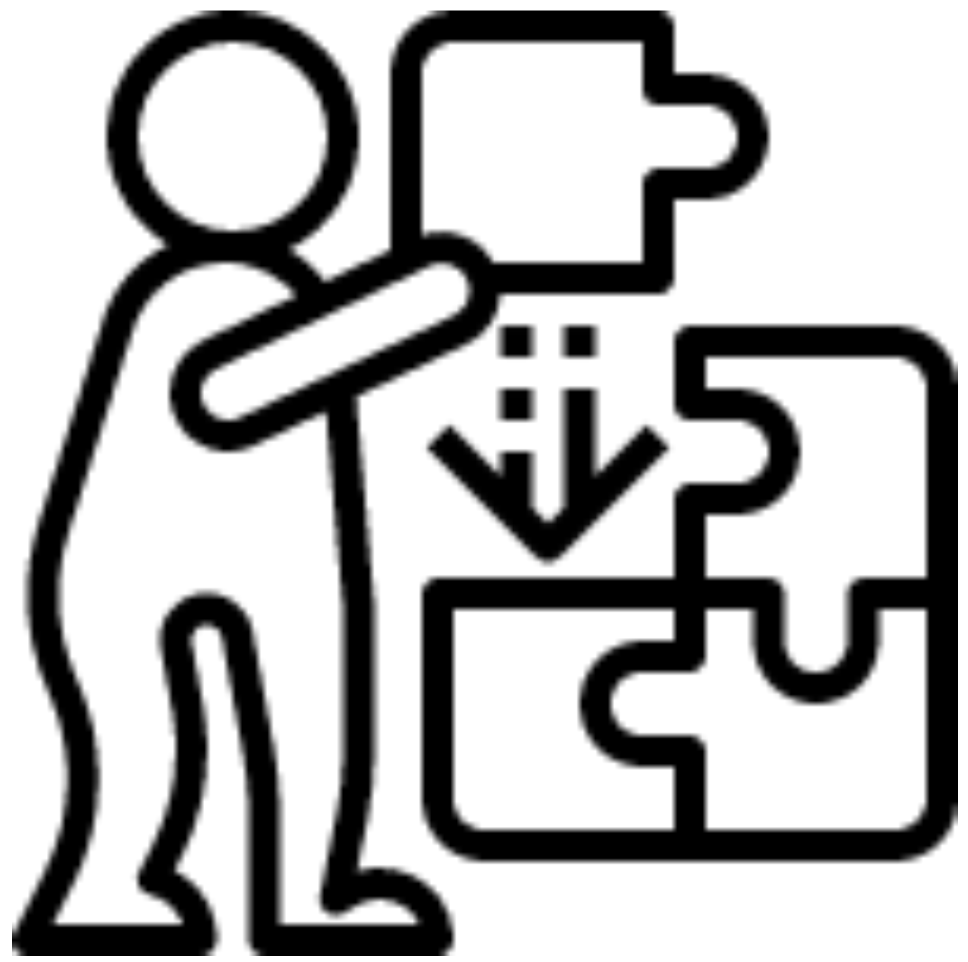
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.

Popular Explanation Styles



Feature Importance

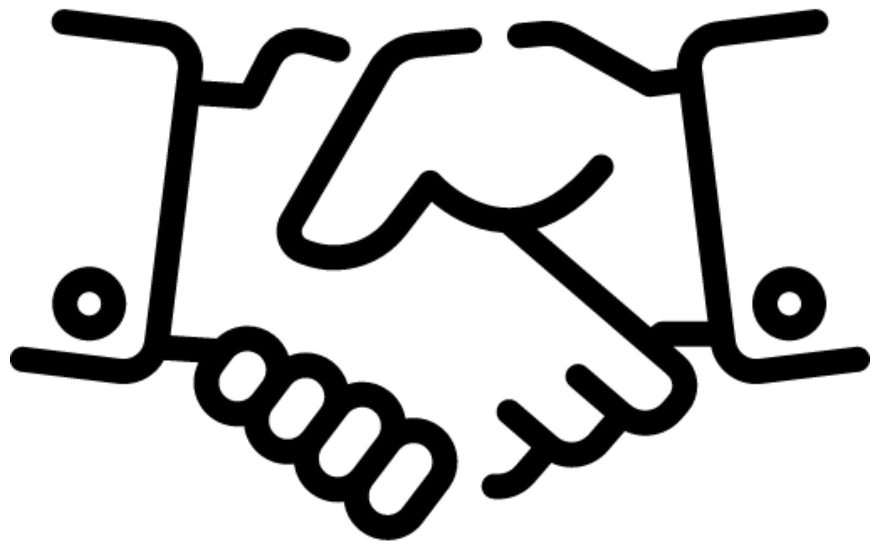
Sample Importance



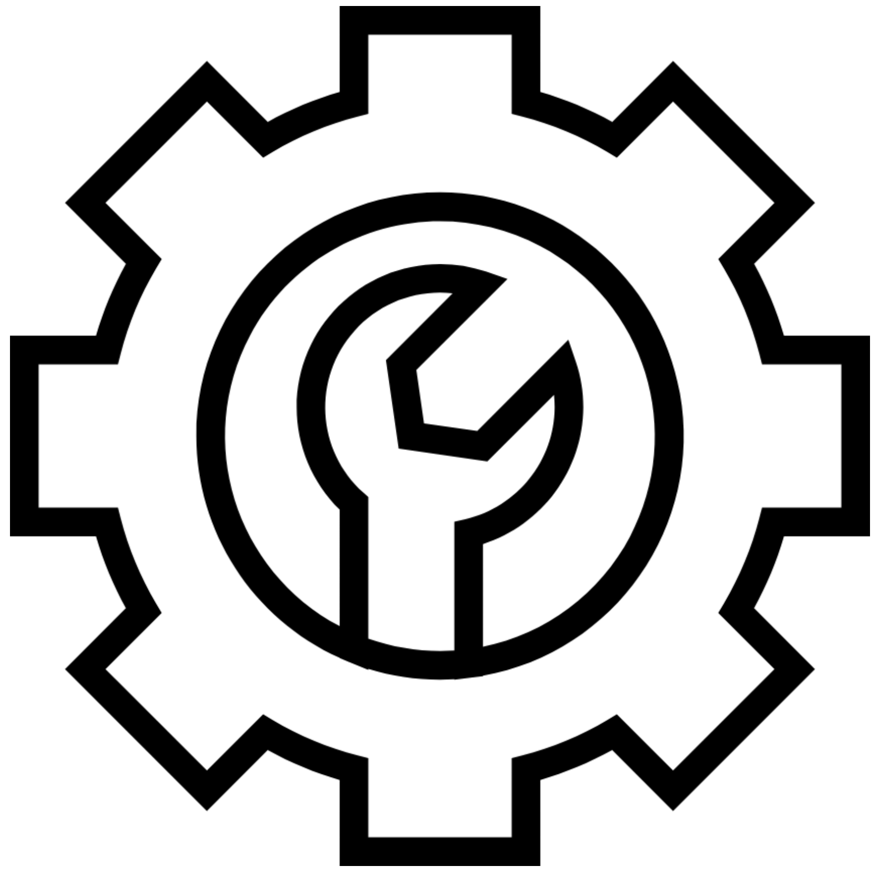
Counterfactuals

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

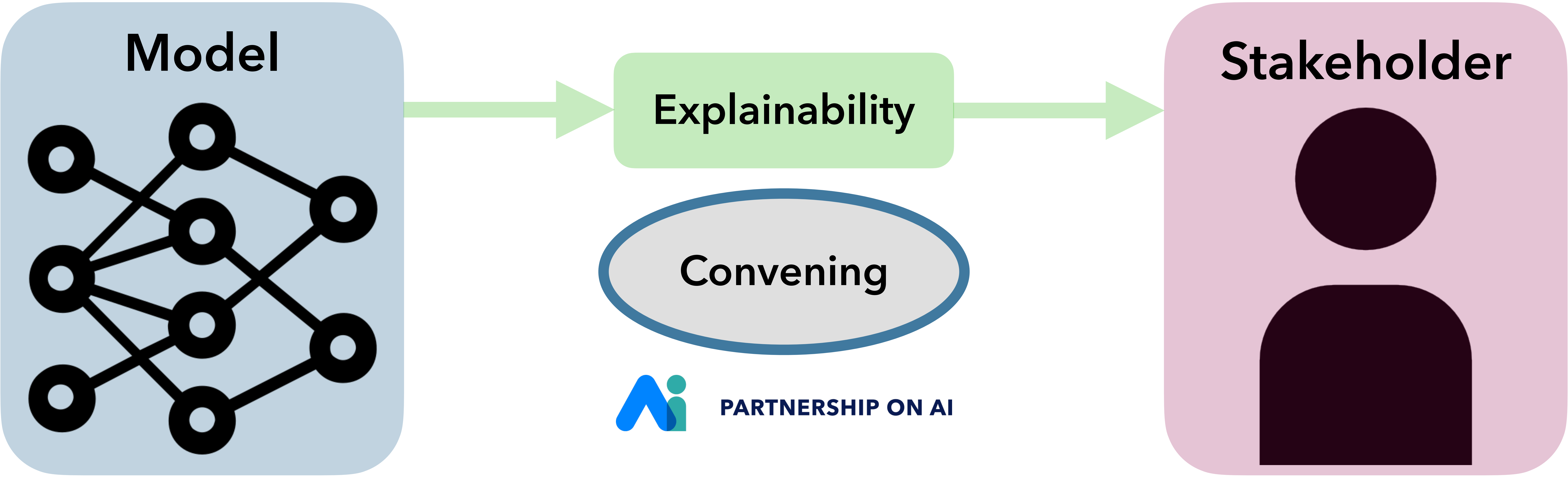
Regulators

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

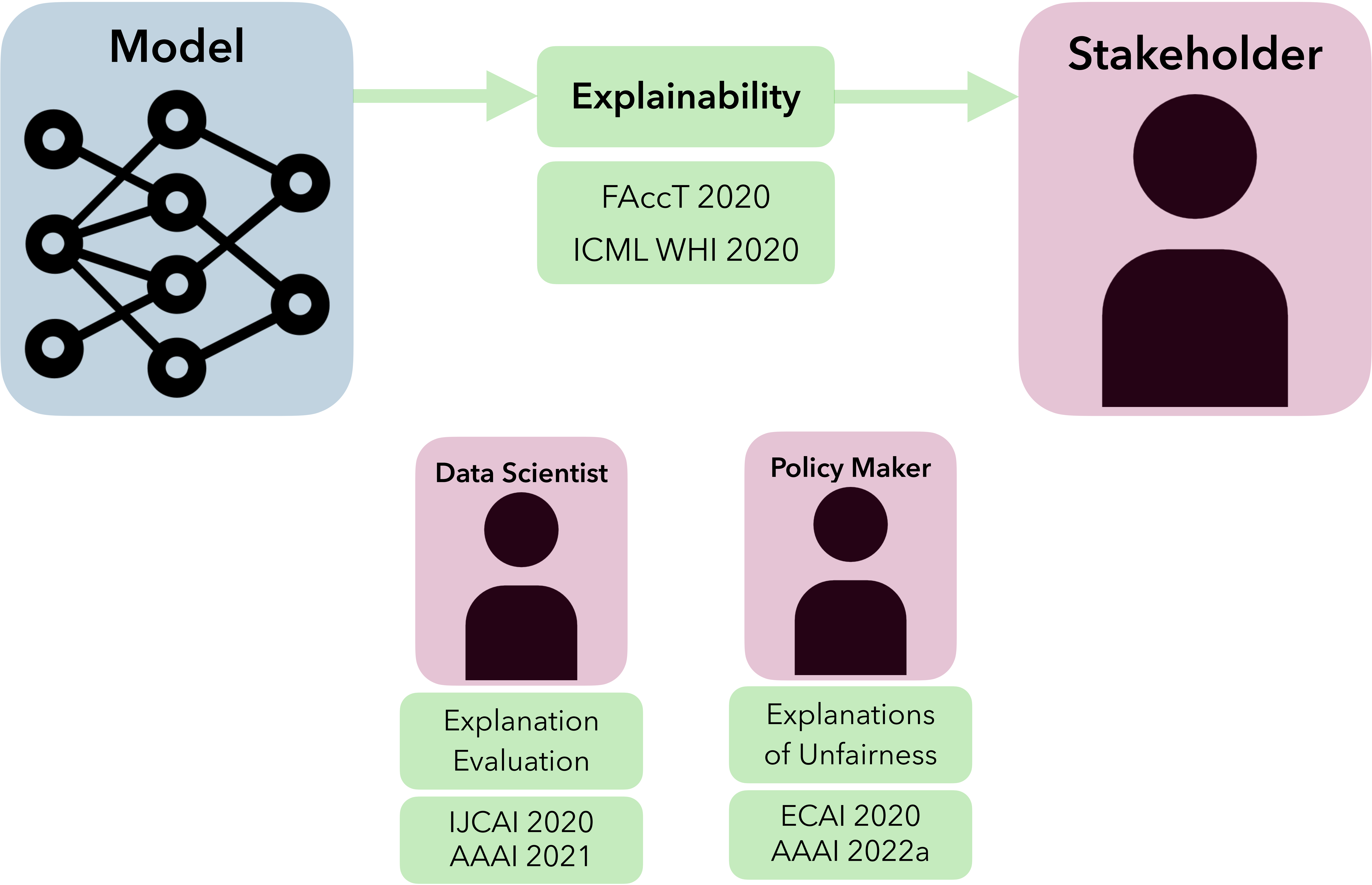
B, Xiang, Sharma, Weller, Taly, Jia, Ghosh, Puri, Moura, Eckersley. *Explainable Machine Learning in Deployment.* ACM FAccT. 2020.



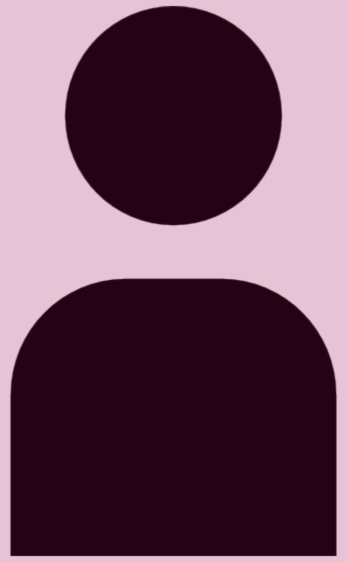
Goal: facilitate an *inter-stakeholder* conversation around explainability

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.


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

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?

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

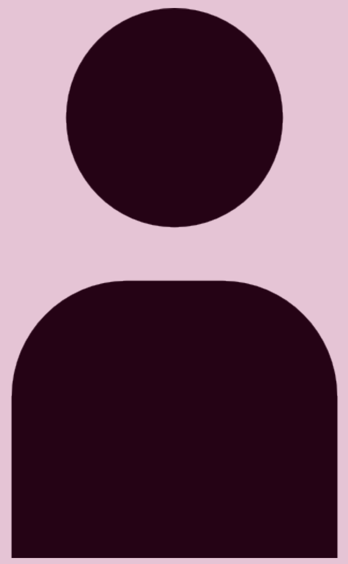
Complexity: Is the explanation digestible?

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

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

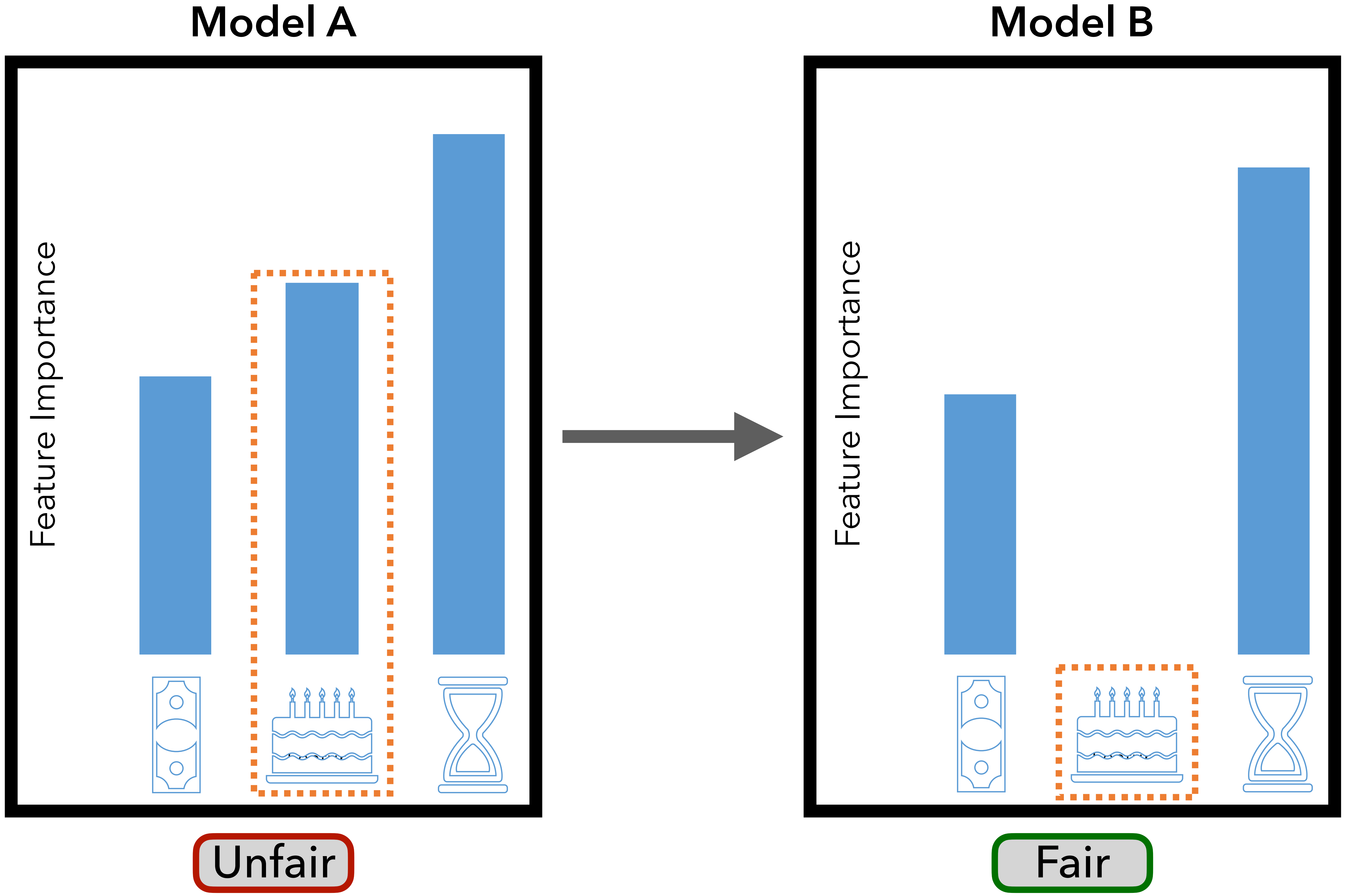


Explanations of Unfairness

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 assume model fairness via explanations

Methods

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

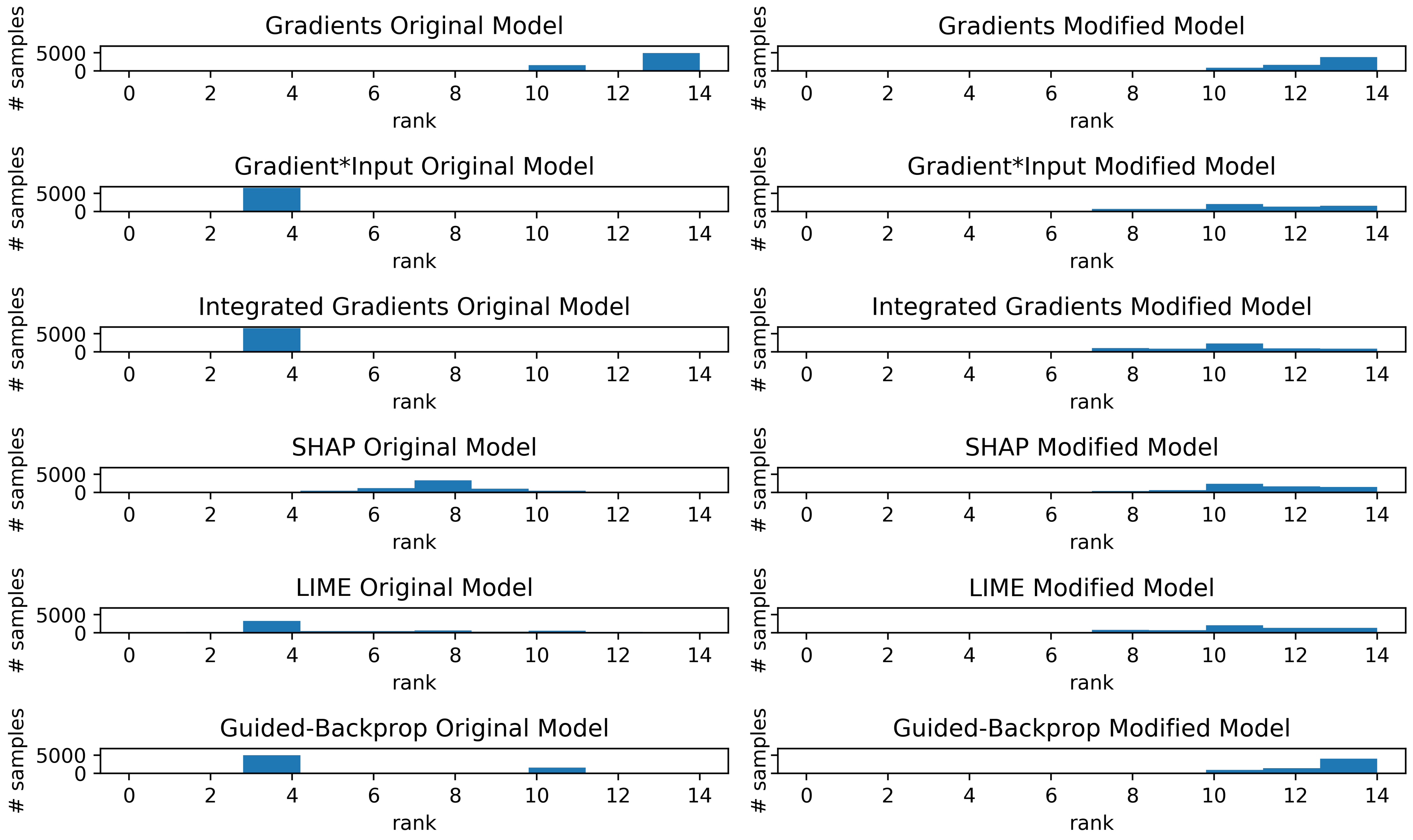
Our Goal $f_\theta \rightarrow f_{\theta+\delta}$

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

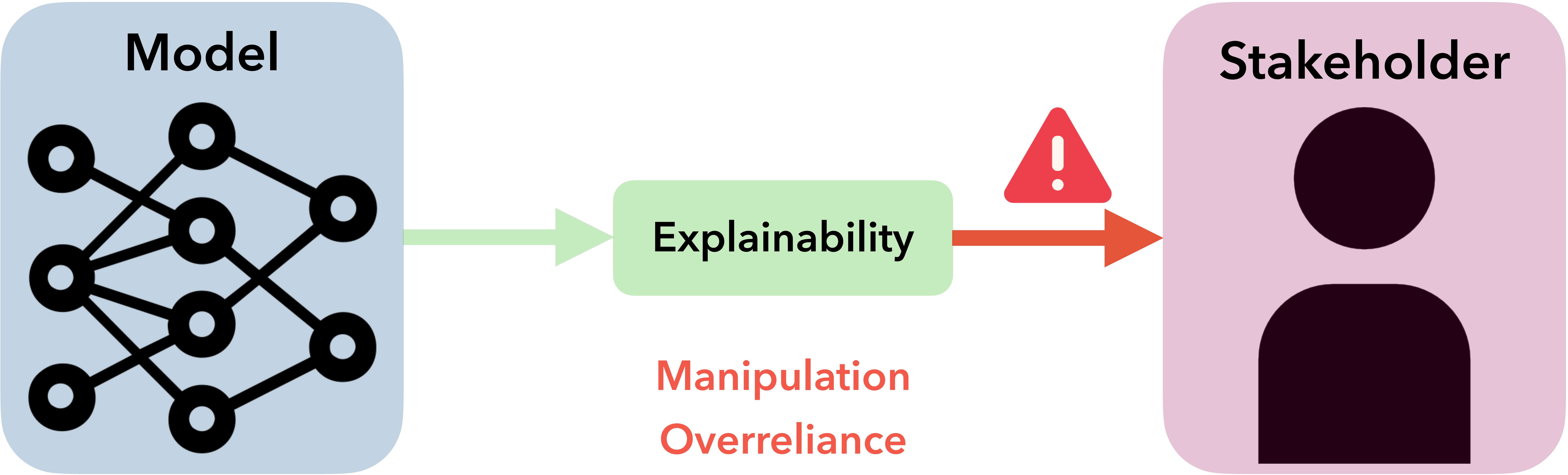
$$\text{argmin}_\delta \; L' = L(f_{\theta+\delta}, x, y) + \frac{\alpha}{n} \left|\left| \nabla_{\mathbf{X}_{:,j}} L(f_{\theta+\delta}, x, y) \right|\right|_p$$



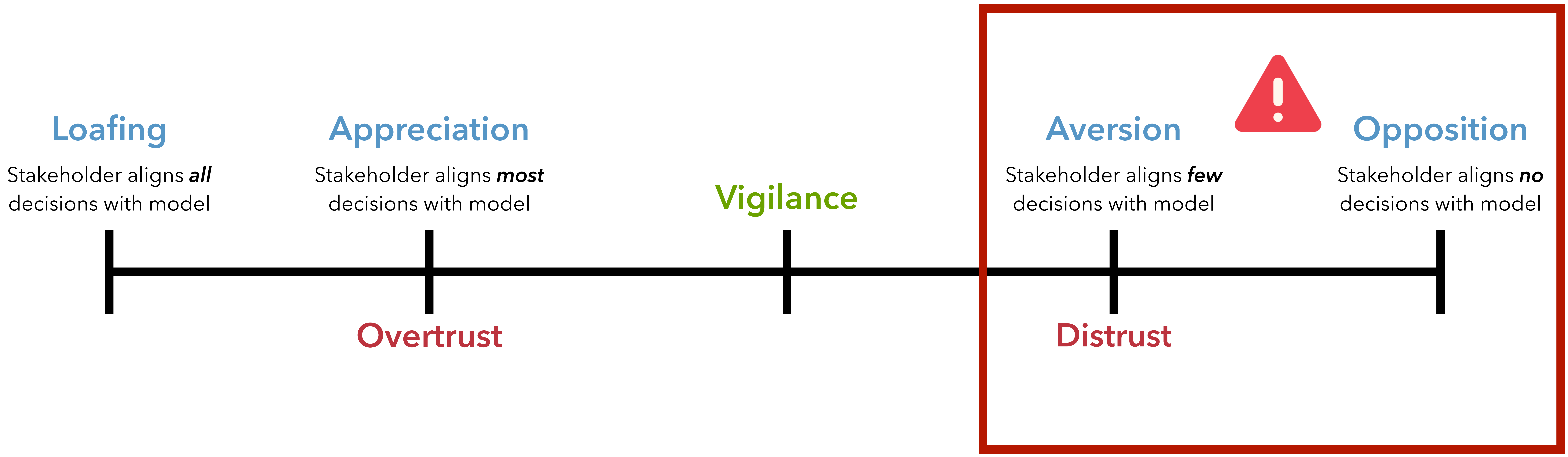
Our proposed attack:

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

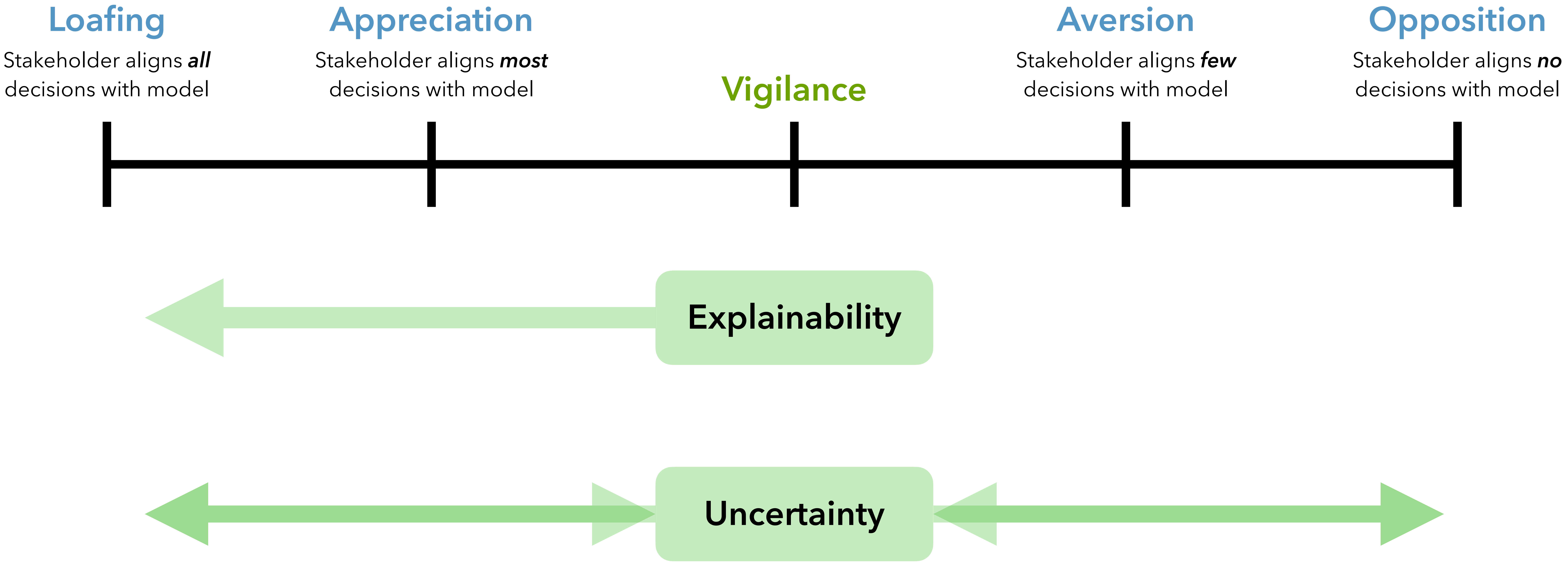
Heo, Joo, Moon. *Fooling Neural Network interpretations via adversarial model manipulation*. NeurIPS. 2019.
Dimanov, **B**, Jamnik, Weller. *You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods*. ECAI. 2020.



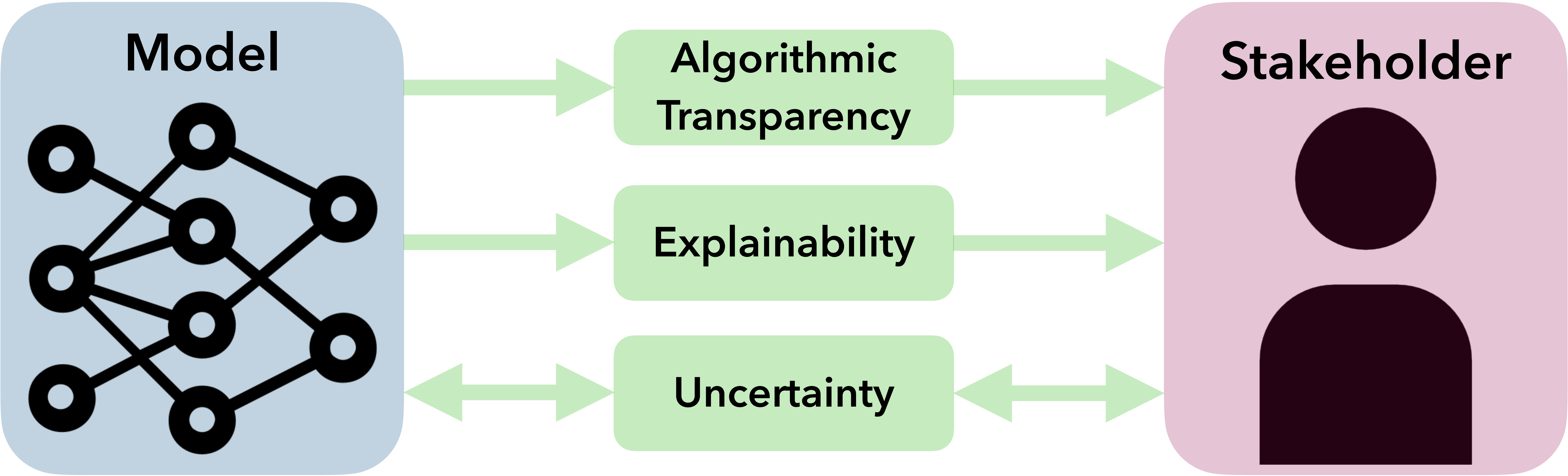
Weller. *Transparency: Motivations and Challenges*. Chapter 2 in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. 2019
Buçinca, Malaya, Gajos. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. CSCW. 2021.
Zerilli,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



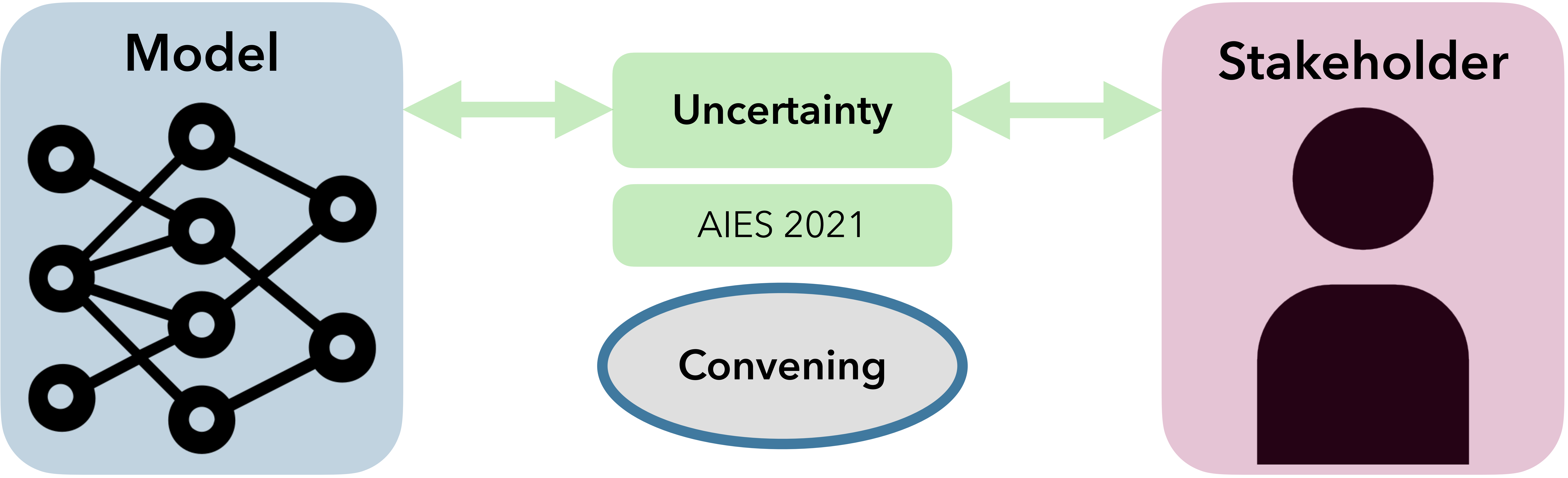
Dietvorst, Simmons, Massey. *Algorithm aversion: People Erroneously Avoid Algorithms after Seeing Them Err.* Journal of Experimental Psychology. 2015.
Logg, Minson, Moore. *Algorithm appreciation: People prefer algorithmic to human judgment.* Organizational Behavior and Human Decision Processes. 2019.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



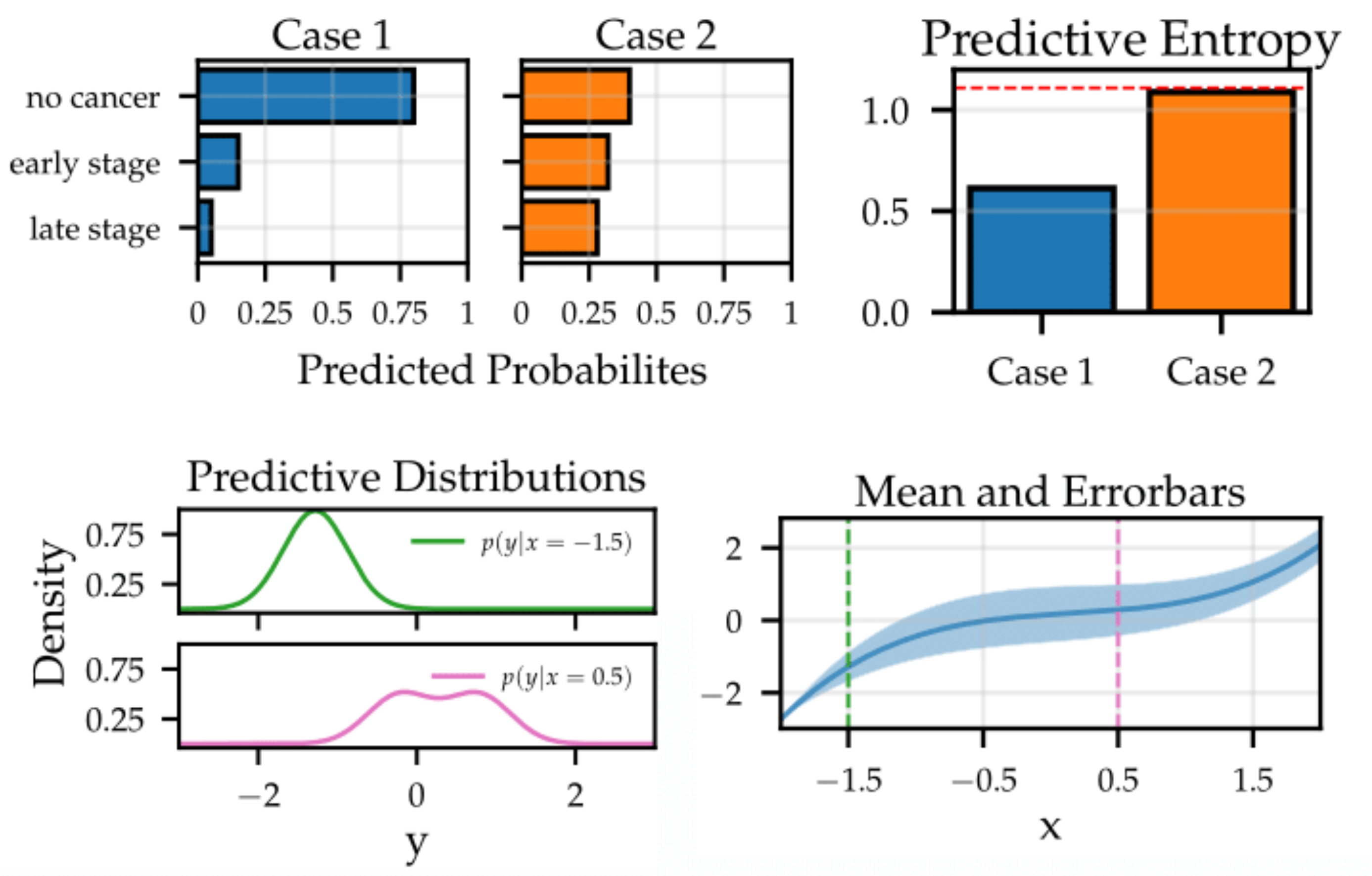
Buçinca, Malaya, Gajos. *To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making*. CSCW. 2021.
Zerilli, **B**, Weller. *How transparency modulates trust in artificial intelligence*. Patterns. 2022.



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.



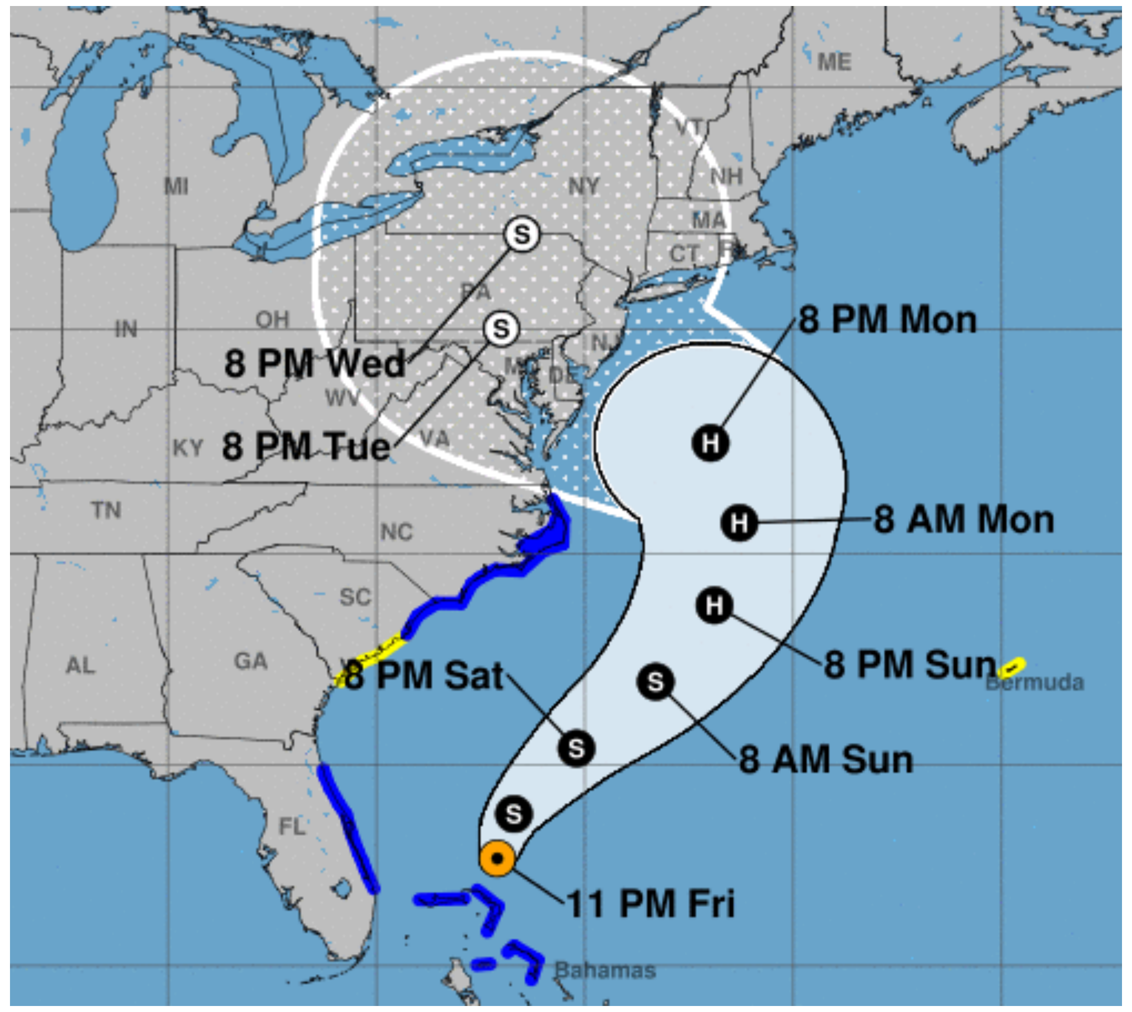
Step 1: Measuring



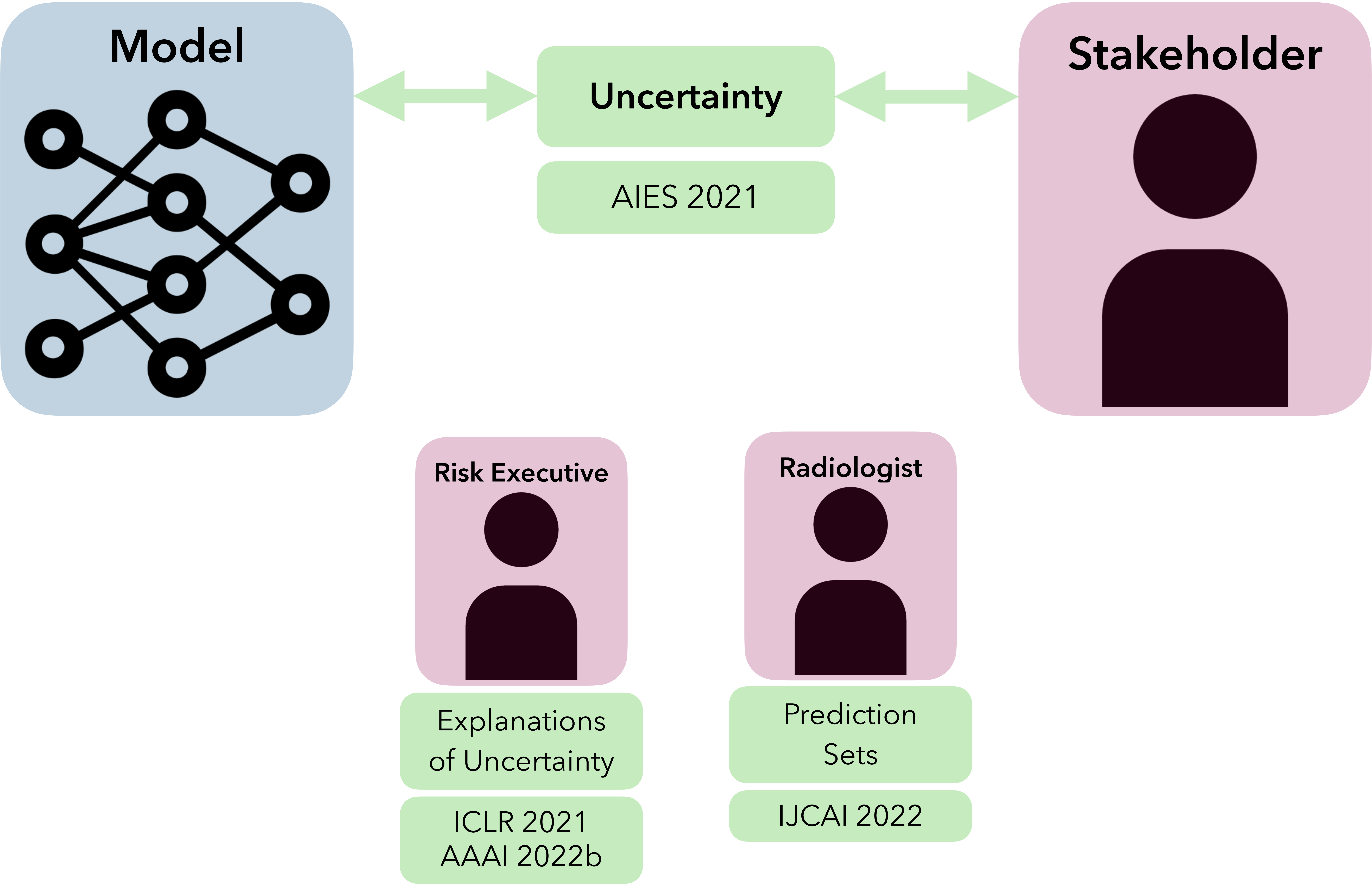
Step 2: Using

- **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

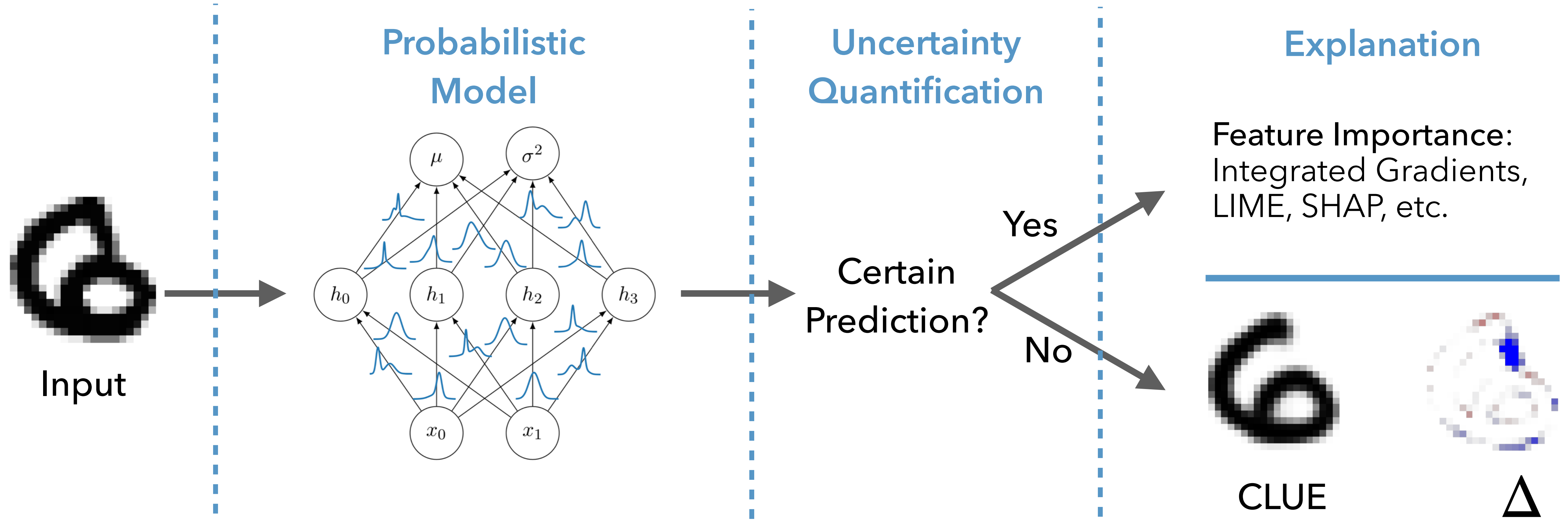
Explanations of Uncertainty

ICLR 2021
AAAI 2022b

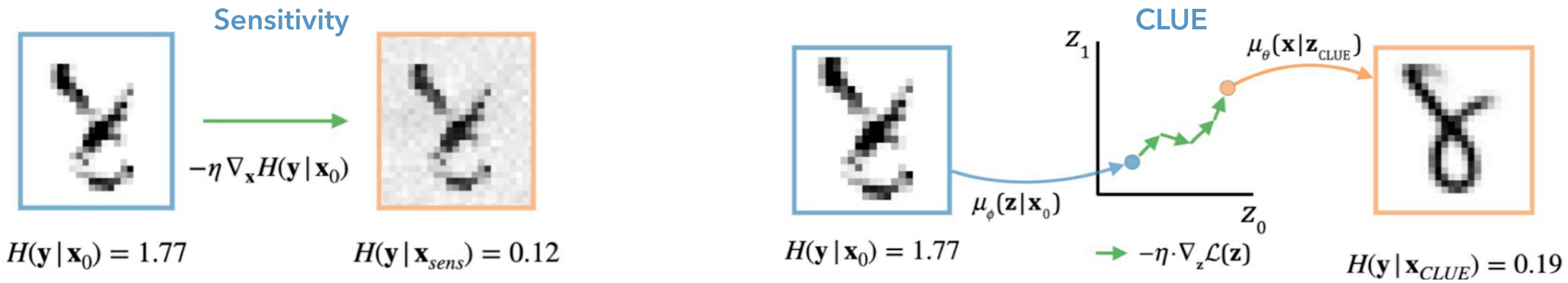
CLUE: Counterfactual Latent Uncertainty Explanations

Methods

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

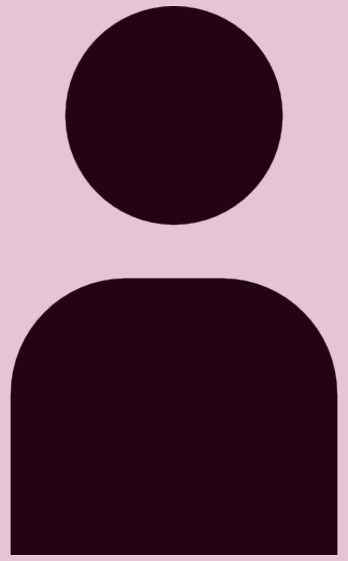


Formulation: *What is the smallest change we need to make to an input, while staying in-distribution, such that our model produces more certain predictions?*



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

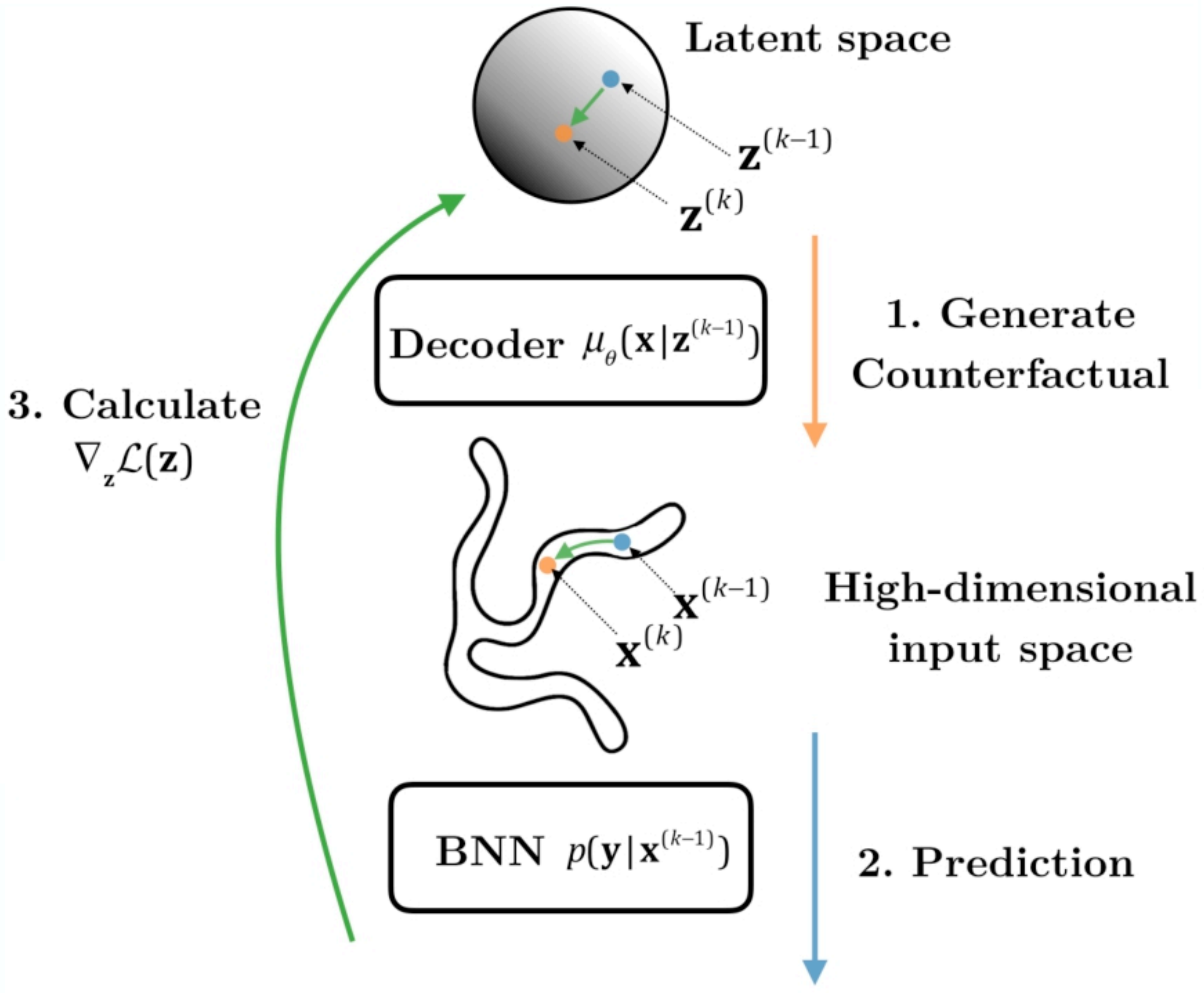


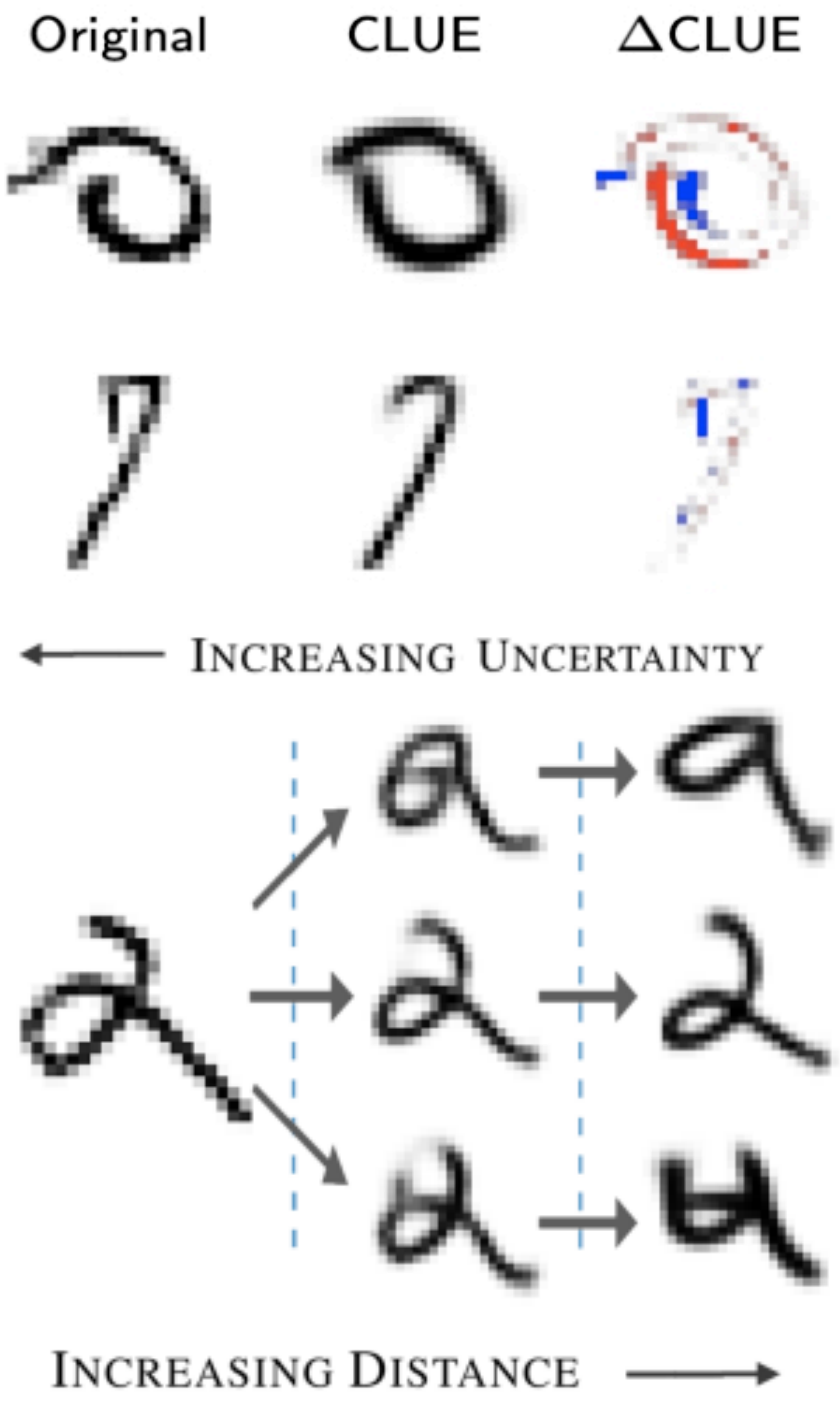
Explanations of Uncertainty

ICLR 2021
AAAI 2022b

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

ICLR 2021
AAAI 2022b

CLUE: Counterfactual Latent Uncertainty Explanations

User Studies

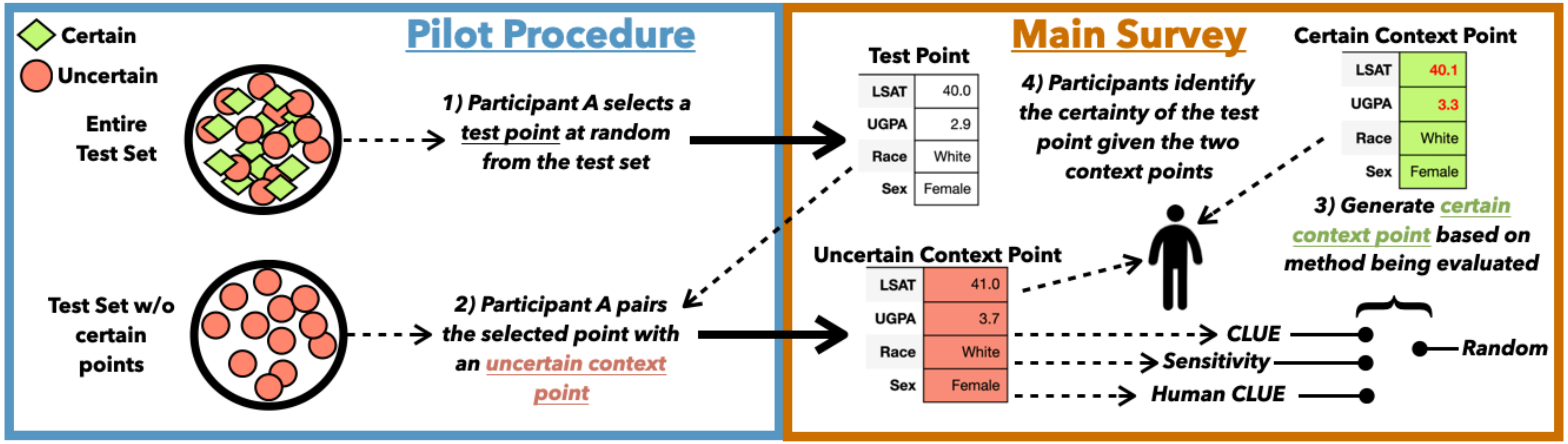
Human Simulatability: Users are shown context examples and are tasked with predicting model behavior on new datapoint.

	Uncertain	Certain	?
Age	Less than 25	Less than 25	Less than 25
Race	Caucasian	African-American	Hispanic
Sex	Male	Male	Male
Current Charge	Misdemeanour	Misdemeanour	Misdemeanour
Reoffended Before	Yes	No	No
Prior Convictions	1	0	0
Days Served	0	0	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.

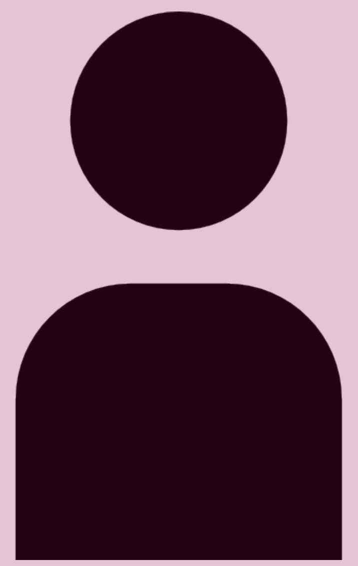
(Using Nemenyi test for average ranks across test questions)



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



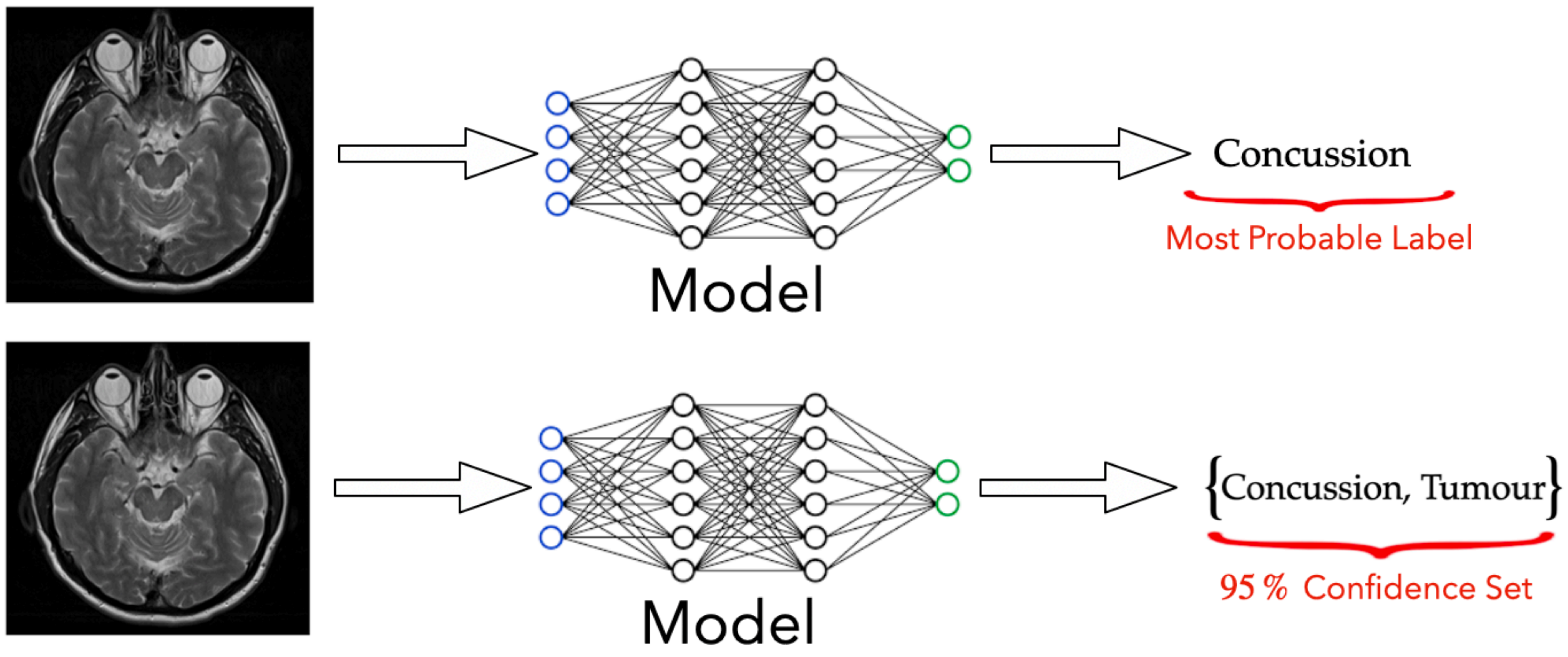
Prediction Sets

IJCAI 2022

Generate prediction sets for experts

Methods

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



Top-1 Classifier

Set Valued Classifier

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

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.

Prediction Sets

IJCAI 2022

Generate prediction sets for experts

User Studies

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

A CP Scheme!

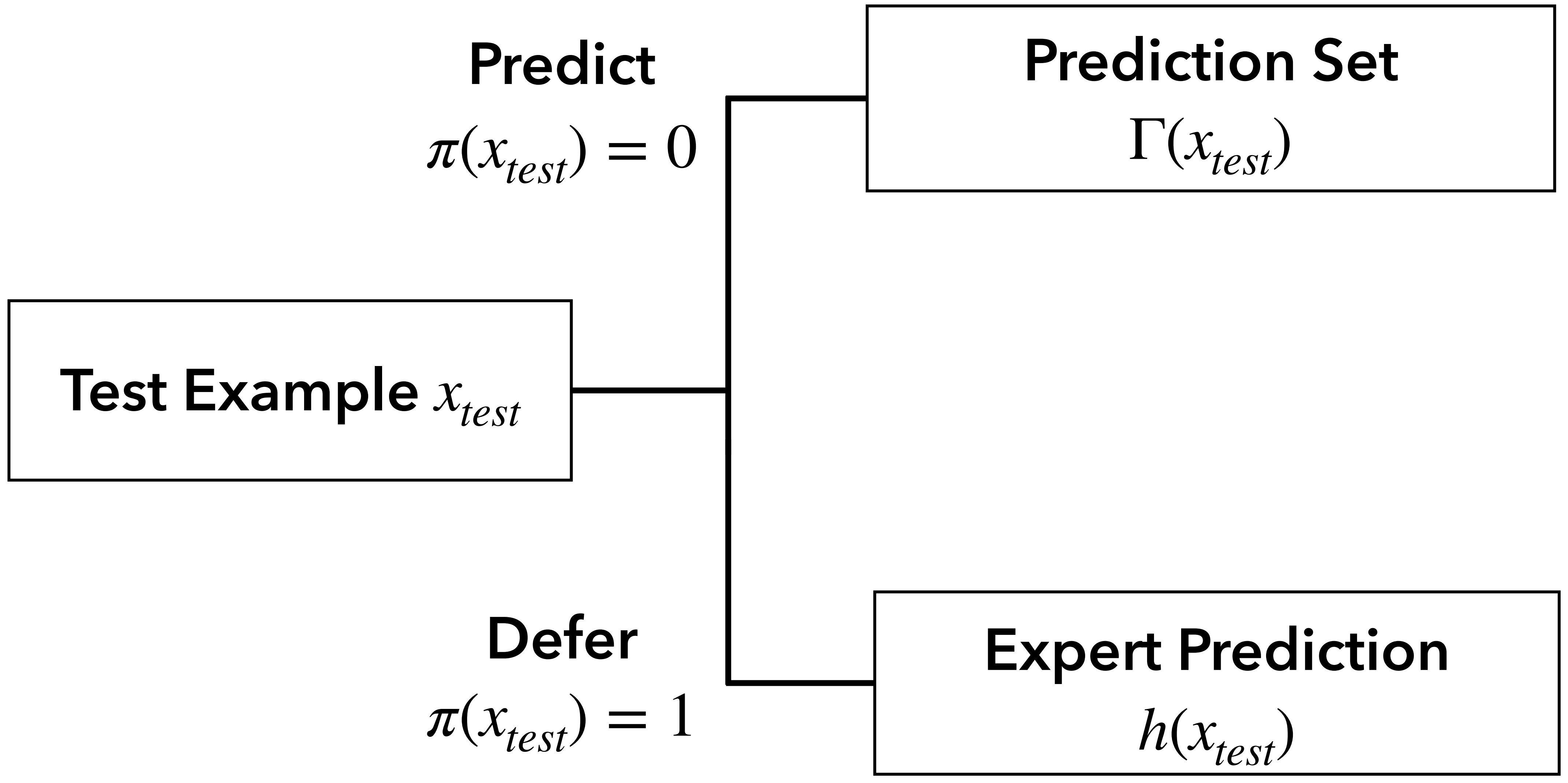
For CIFAR-100:

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	$<$ **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 $\pm$ 0.08	0.67 $\pm$ 0.05	**0.003**	0.832
Reported Utility	7.93 $\pm$ 0.39	6.32 $\pm$ 0.60	**$<$ 0.001**	1.138
Reported Confidence	7.31 $\pm$ 0.29	7.28 $\pm$ 0.29	0.862	0.046
Reported Trust in Model	8.00 $\pm$ 0.45	6.87 $\pm$ 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).

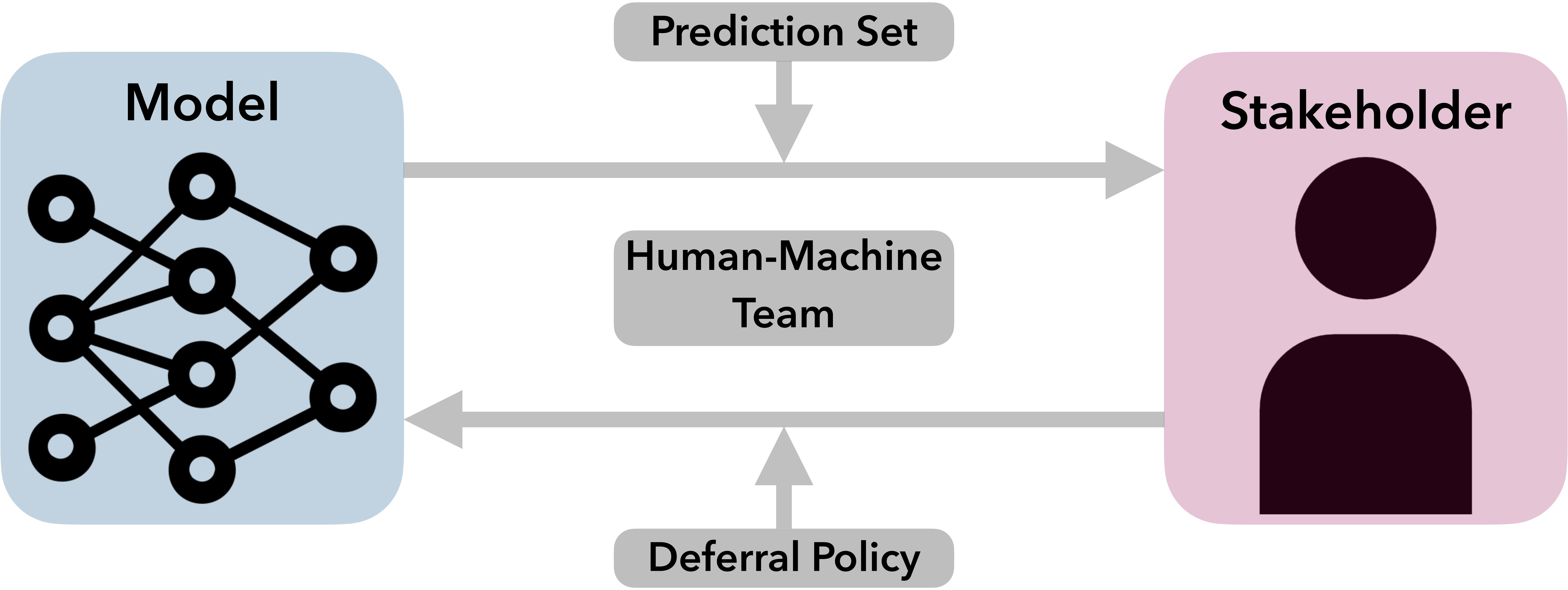
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.

Takeaways

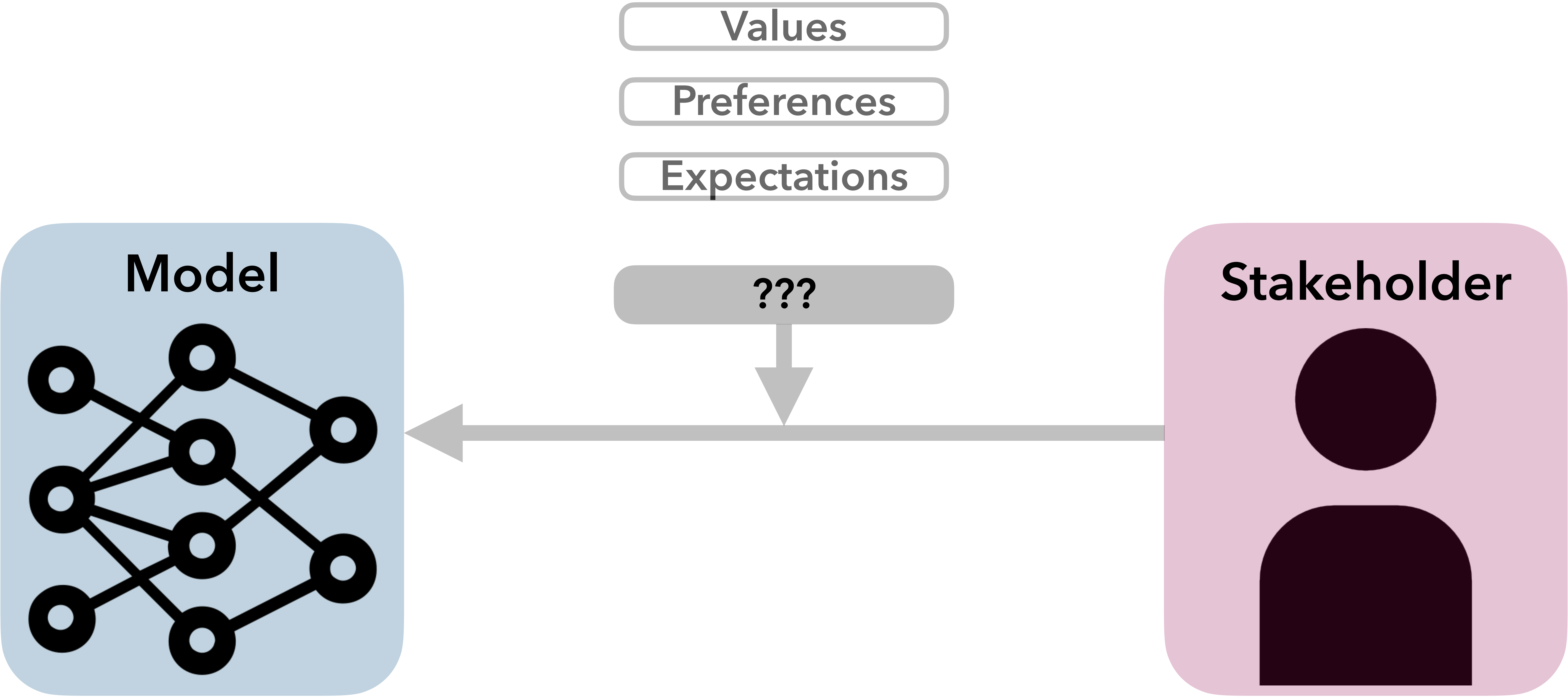
Algorithmic transparency is important but difficult

- 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

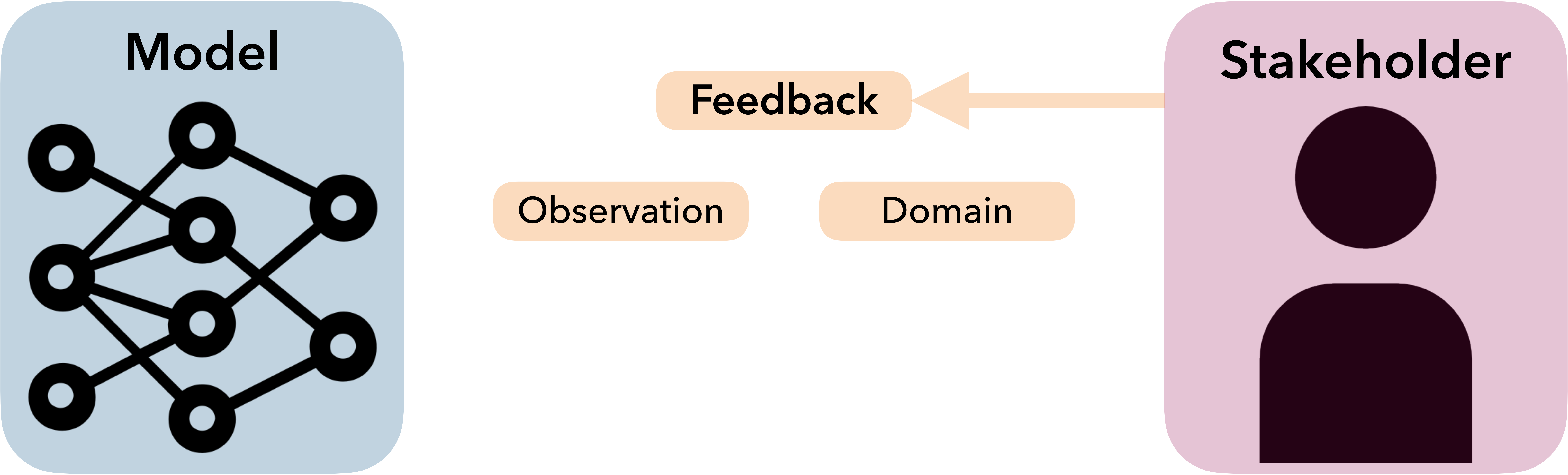
Convening is powerful tool to motivate technical and socio-technical research



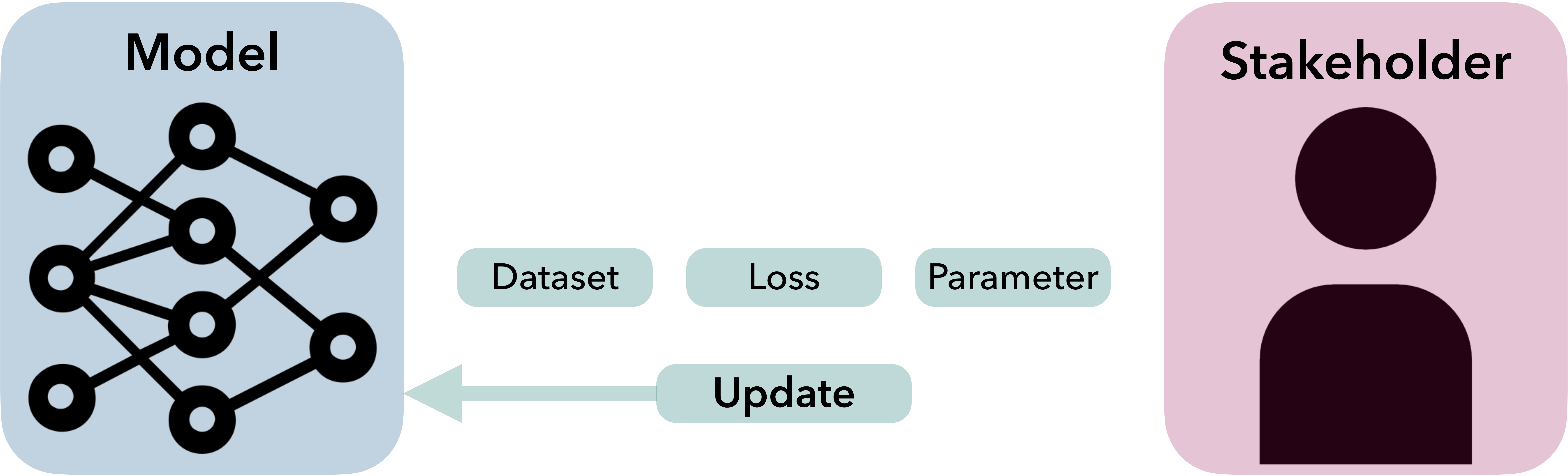
Babbar, **B**, Weller. *On the Utility of Prediction Sets in Human-AI Teams*. IJCAI. 2022.



Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* ICML Workshop on Updatable ML. 2022.

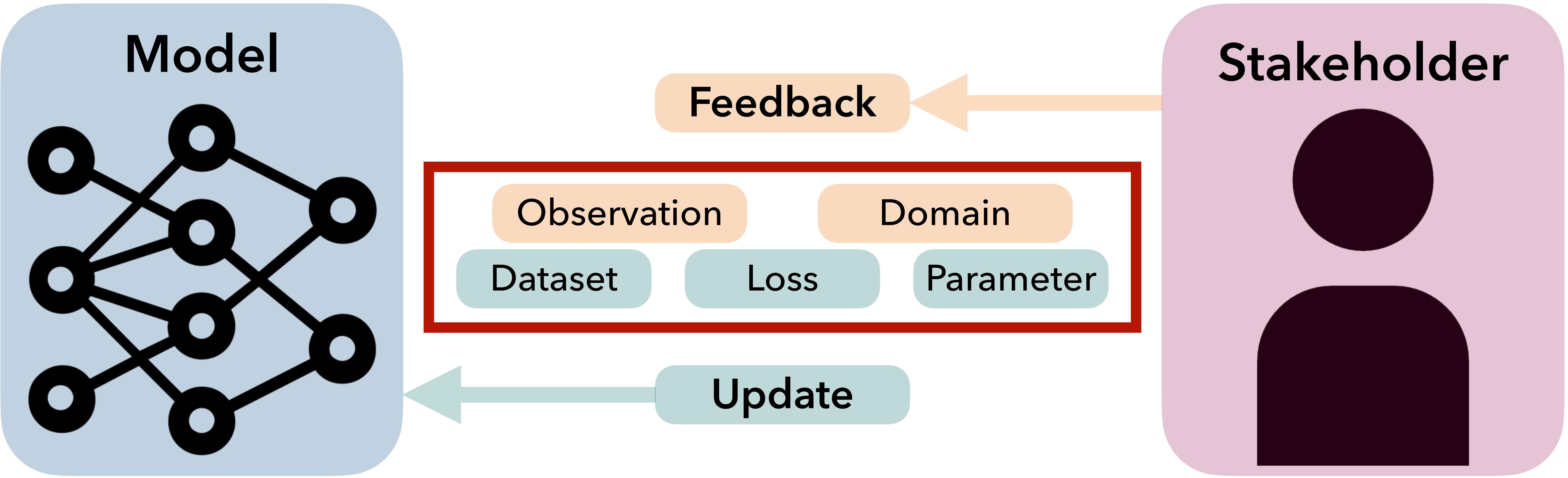


Hertwig, Erev. *The description–experience gap in risky choice*. Trends in Cognitive Science. 2009.
Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* ICML Workshop on Updatable ML. 2022.



Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* ICML Workshop on Updatable ML. 2022.

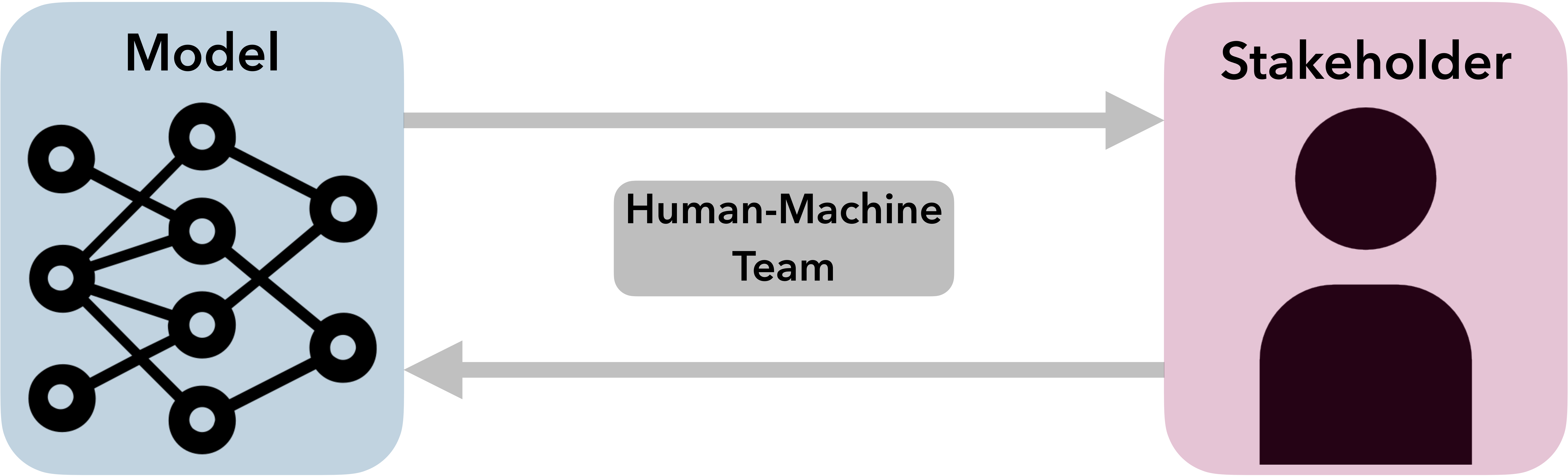
Feedback-Update Taxonomy

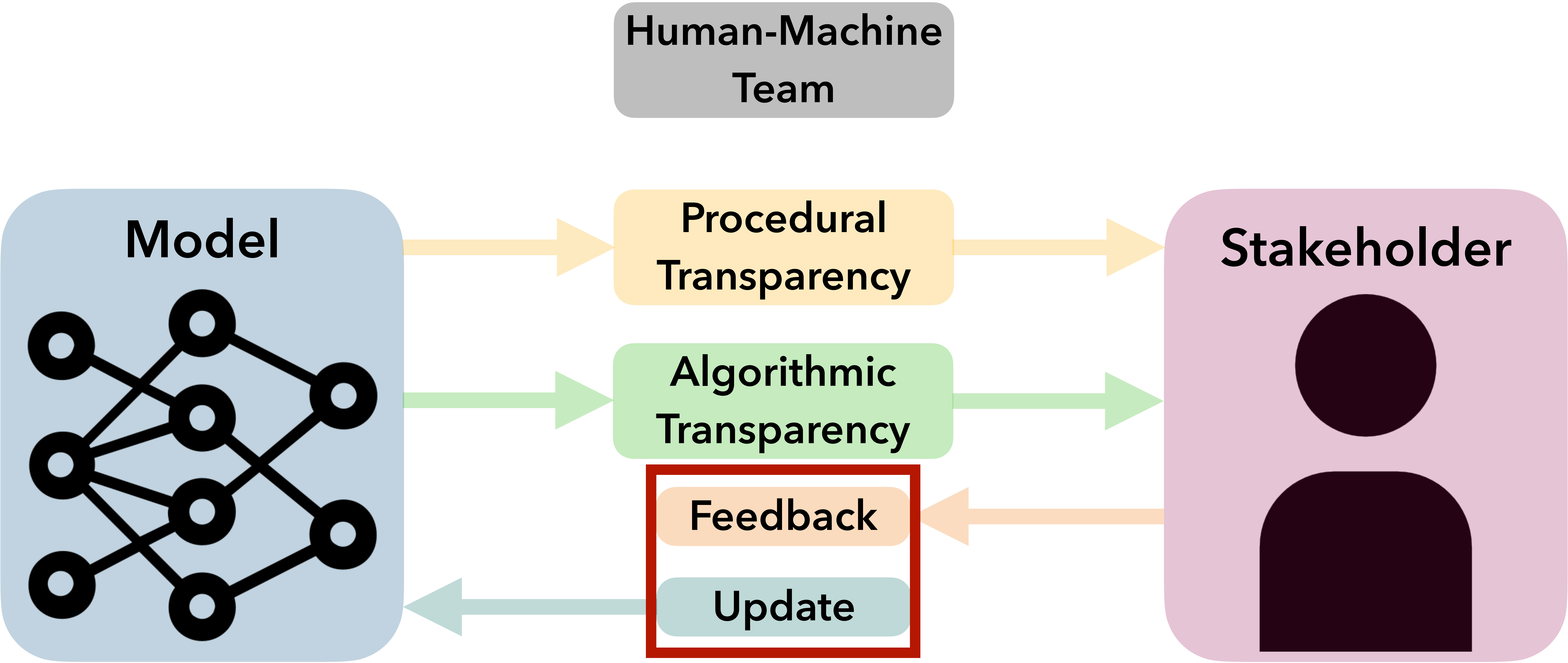


Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates*. ICML Workshop on Updatable ML. 2022.

Future Directions

- Open technical questions around algorithmic transparency can be addressed with new methods and well-designed user studies
- Study the socio-technical nature and societal implications of providing transparency in specific contexts
- Conduct general research into human-machine teams


Model
Human-Machine Team
Stakeholder



Chen*, **B***, Heidari, Weller, Talwalkar. *Perspectives on Incorporating Expert Feedback into Model Updates.* ICML Workshop on Updatable ML. 2022.

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

Computer Science


Miri Zilka
Cambridge


Weiyang Liu
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


B. Schölkopf
MPI


Isabel Valera
MPI


J. von Kügelgen
MPI


Botty Dimanov
Cambridge


Varun Babbar
Cambridge


Isabel Chien
Cambridge


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
Oxford


Alice Xiang
Sony AI


Var Shankar
RAII

Design


Emma Kallina
Cambridge


Becca Ricks
Mozilla


Dorian Peters
Imperial

Philosophy


Rune Nyrup
Cambridge


Stephen Cave
Cambridge

Algorithmic Transparency in Machine Learning

Thank you for listening! Questions?

@umangsbhatt
usb20@cam.ac.uk