Regulating AI-Assisted Decision-Making

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

(Self)-Regulating AI-Assisted Decision-Making

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

When Should Algorithms Resign?

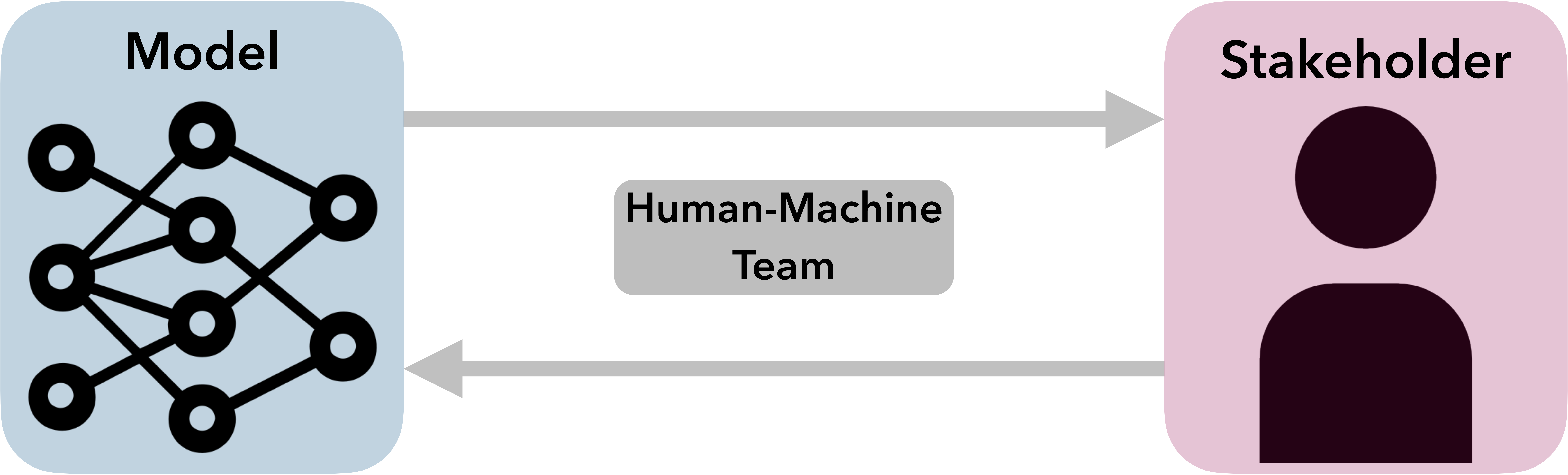
Umang Bhatt
Assistant Professor/Faculty Fellow, New York University
Research Associate, The Alan Turing Institute
Associate Fellow, Leverhulme Center for the Future of Intelligence

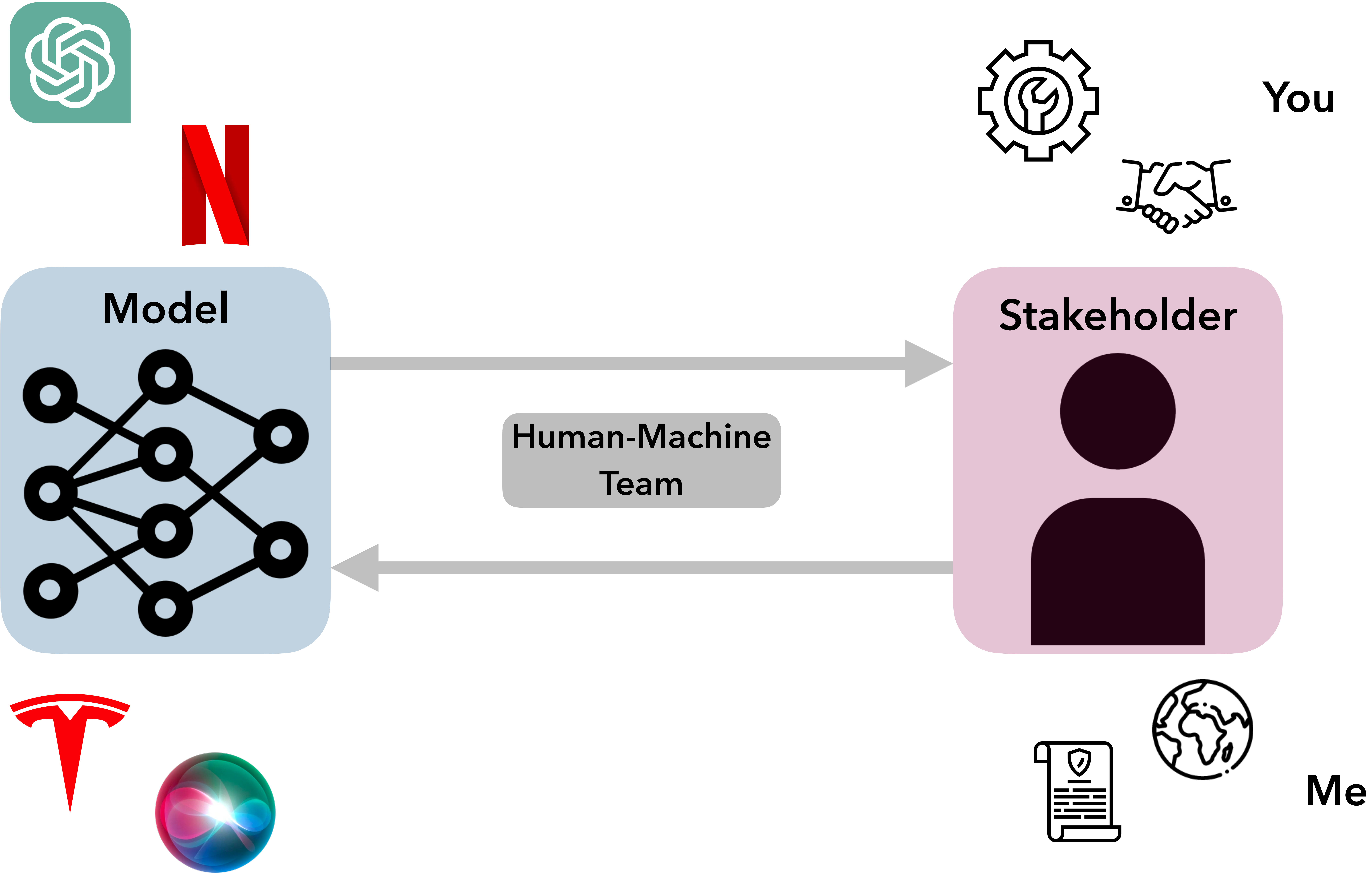
@umangsbhatt
umangbhatt@nyu.edu

NYU

The
Alan Turing
Institute

CFI


Model
Human-Machine Team
Stakeholder



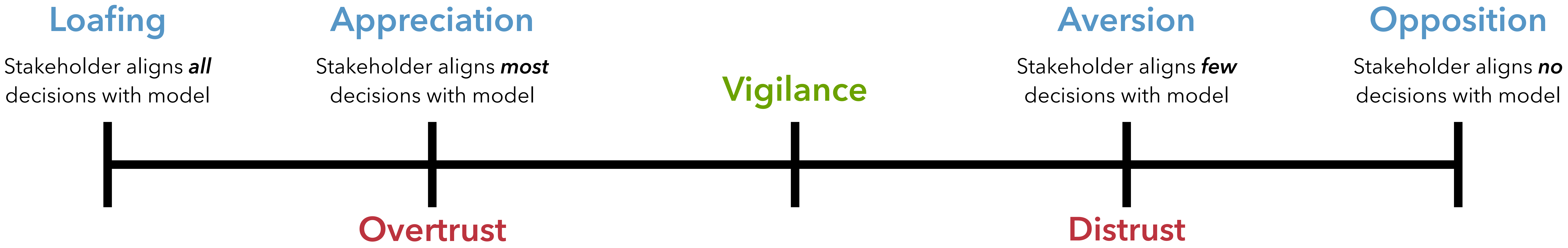
Model

Human-Machine Team

Stakeholder

You

Me



Loafing

Stakeholder aligns ***all*** decisions with model

Appreciation

Stakeholder aligns ***most*** decisions with model

Vigilance

Aversion

Stakeholder aligns ***few*** decisions with model

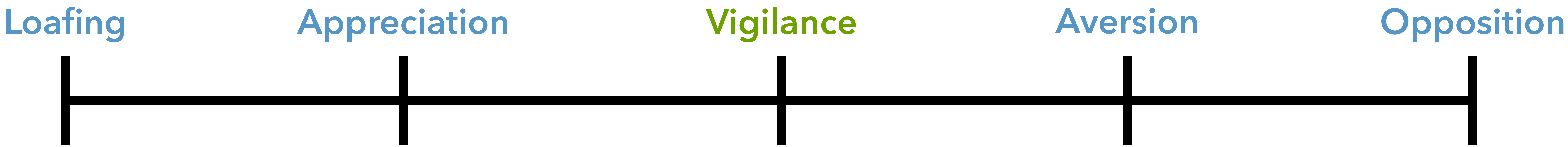
Opposition

Stakeholder aligns ***no*** decisions with model

Overtrust

Distrust

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.



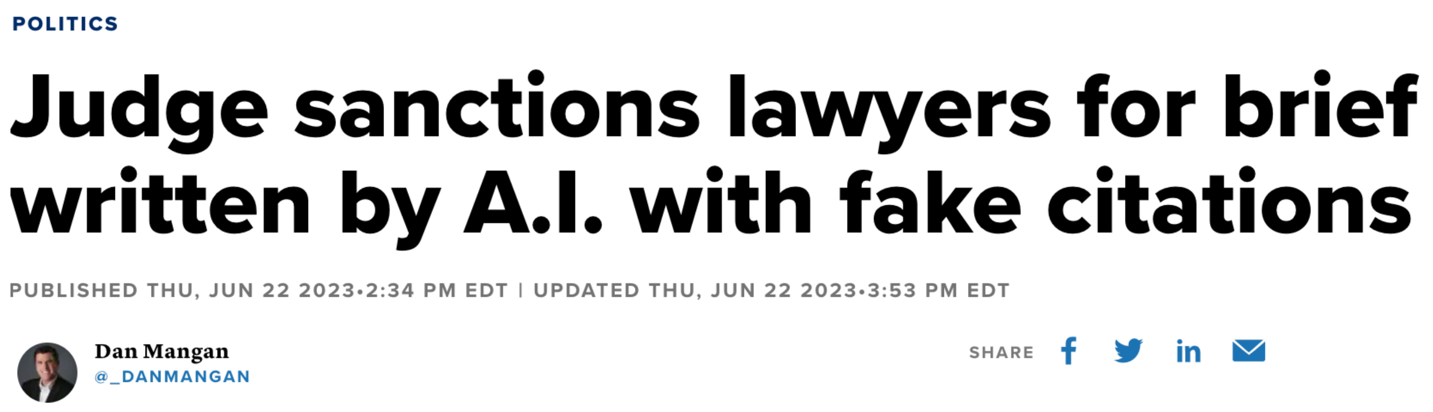
Loafing

Appreciation

Vigilance

Aversion

Opposition



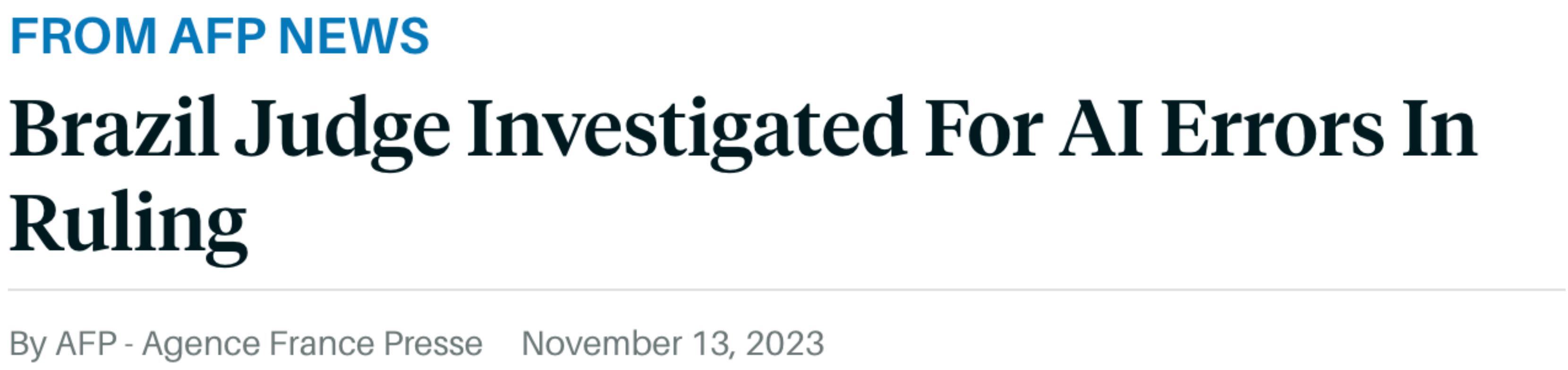
POLITICS

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

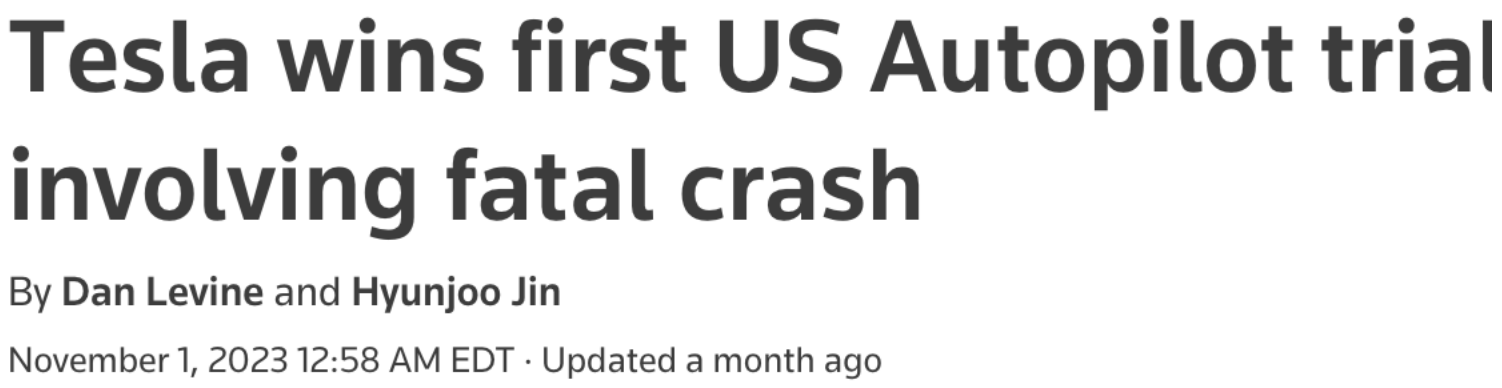
SHARE



FROM AFP NEWS

Brazil Judge Investigated For AI Errors In Ruling

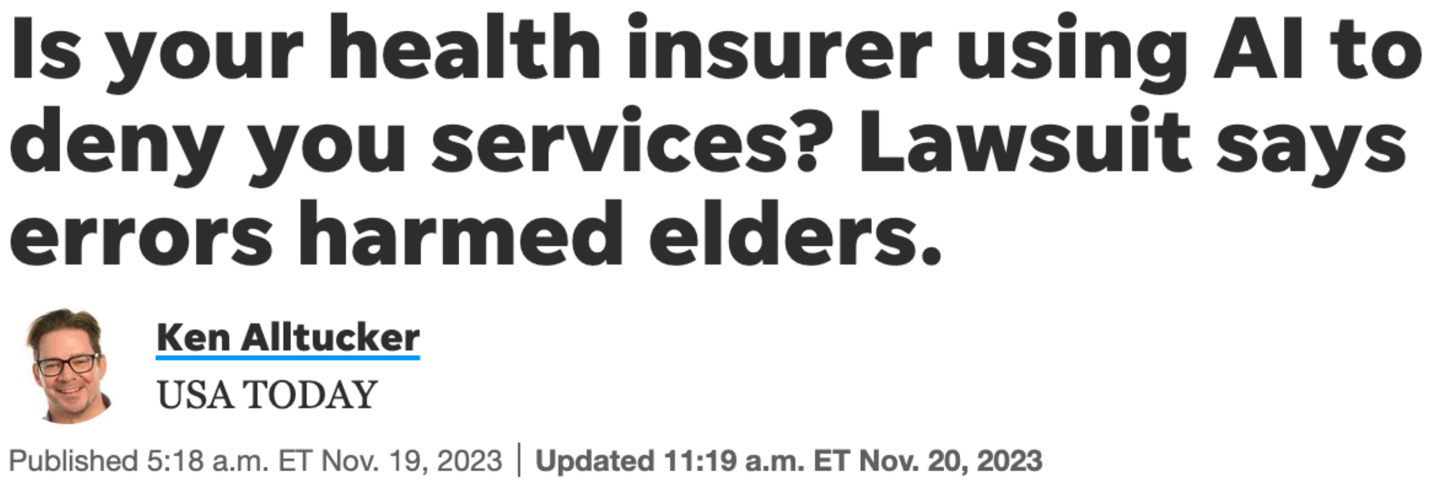
By AFP - Agence France Presse November 13, 2023



Tesla wins first US Autopilot trial involving fatal crash

By Dan Levine and Hyunjoo Jin

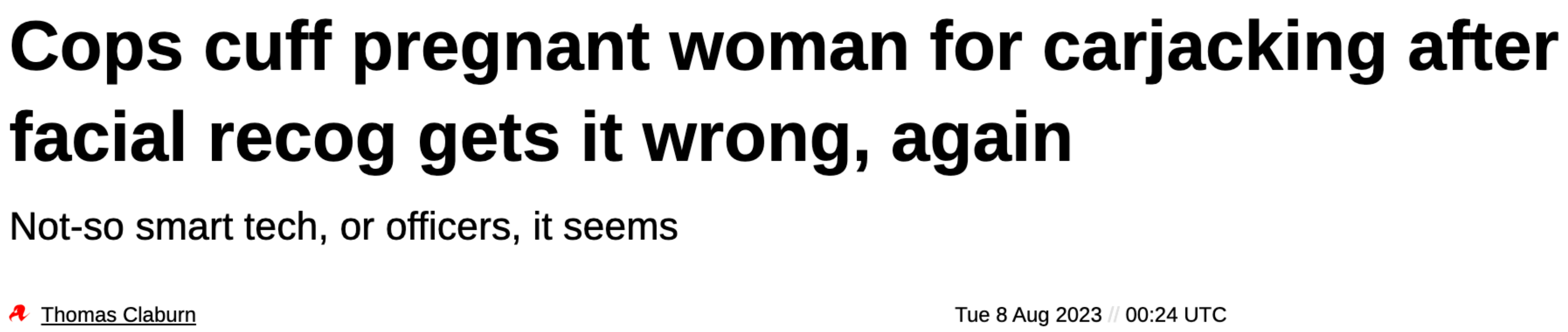
November 1, 2023 12:58 AM EDT · Updated a month ago



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

Ken Alltucker
USA TODAY

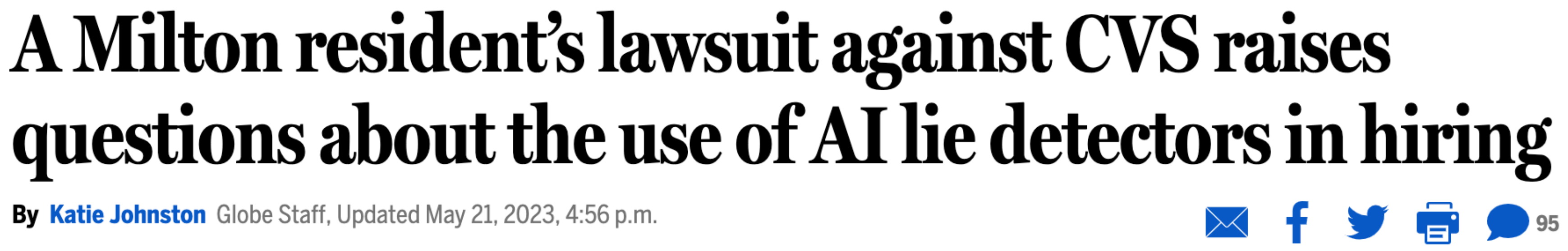
Published 5:18 a.m. ET Nov. 19, 2023 | **Updated 11:19 a.m. ET Nov. 20, 2023**



Cops cuff pregnant woman for carjacking after facial recog gets it wrong, again

Not-so smart tech, or officers, it seems

Thomas Claburn Tue 8 Aug 2023 00:24 UTC

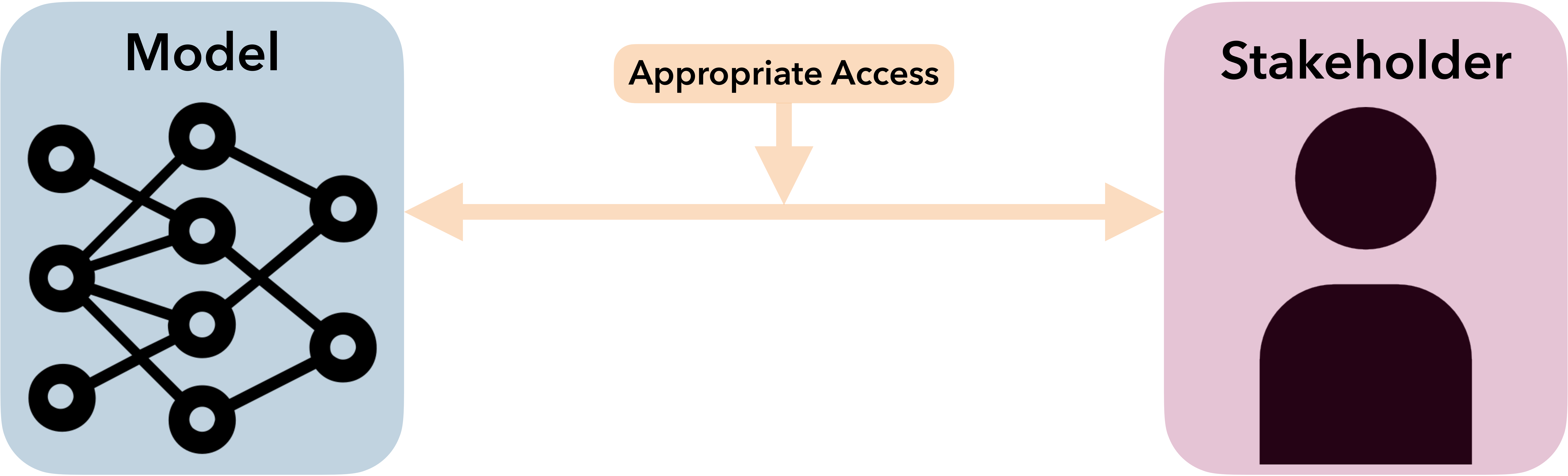


A Milton resident's lawsuit against CVS raises questions about the use of AI lie detectors in hiring

By **Katie Johnston** Globe Staff, Updated May 21, 2023, 4:56 p.m.

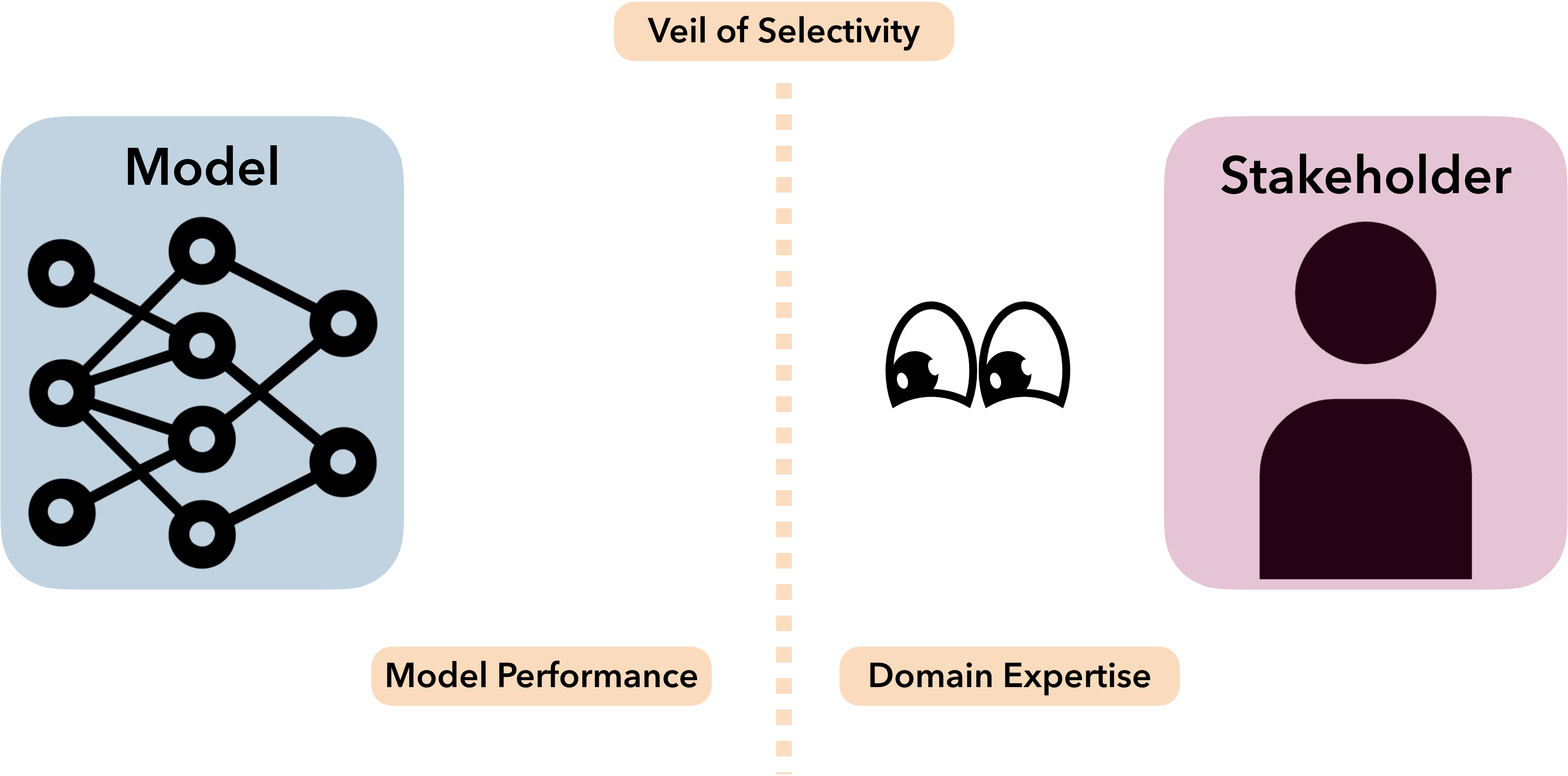
95

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



B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

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



Veil of Selectivity

Model

Stakeholder

Model Performance

Domain Expertise

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

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

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Outline

I. What is *Algorithmic Resignation*?

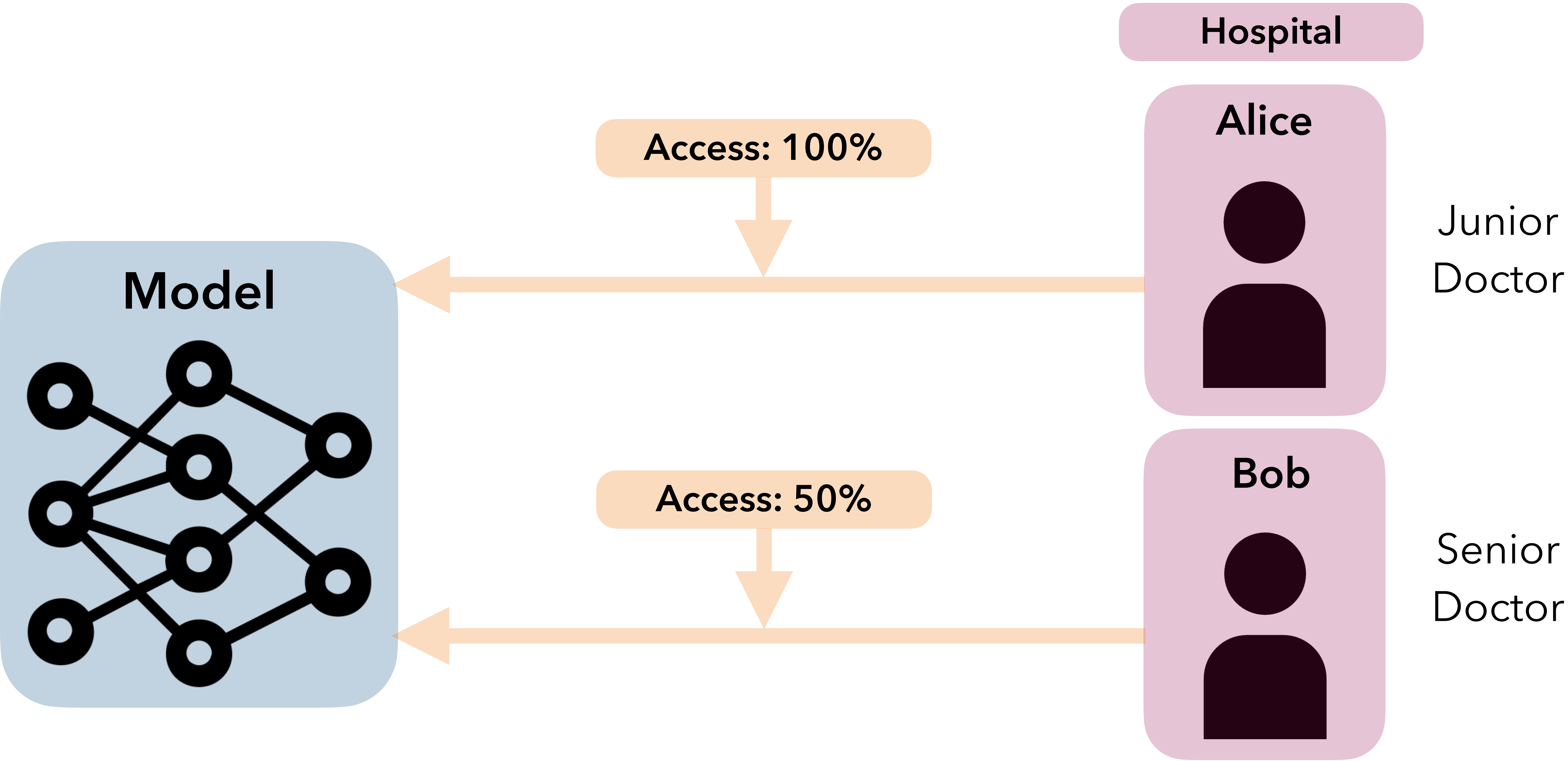
II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

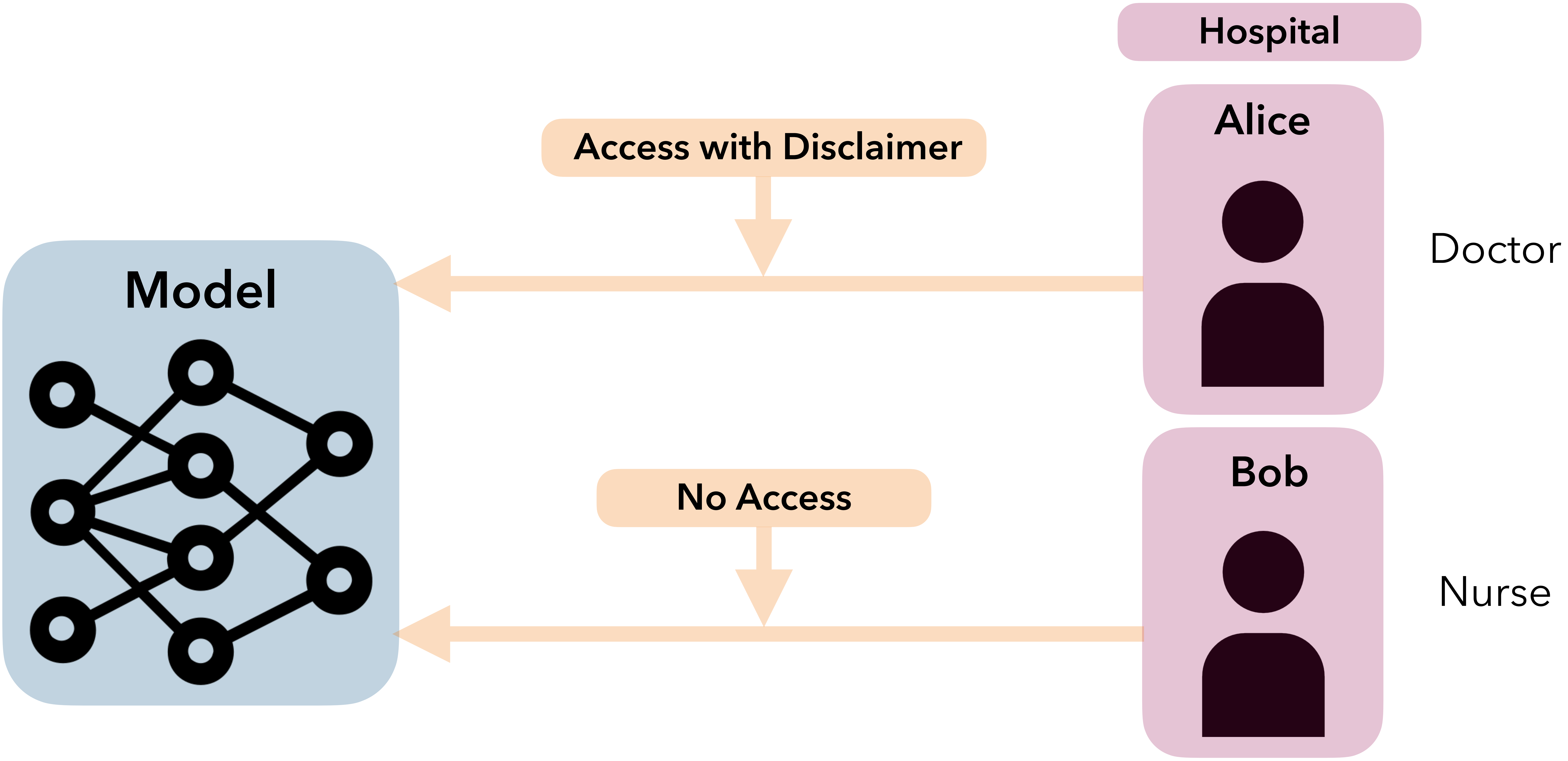
Algorithmic resignation is the *deliberate* and *informed* disengagement from AI assistance in certain scenarios.

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.



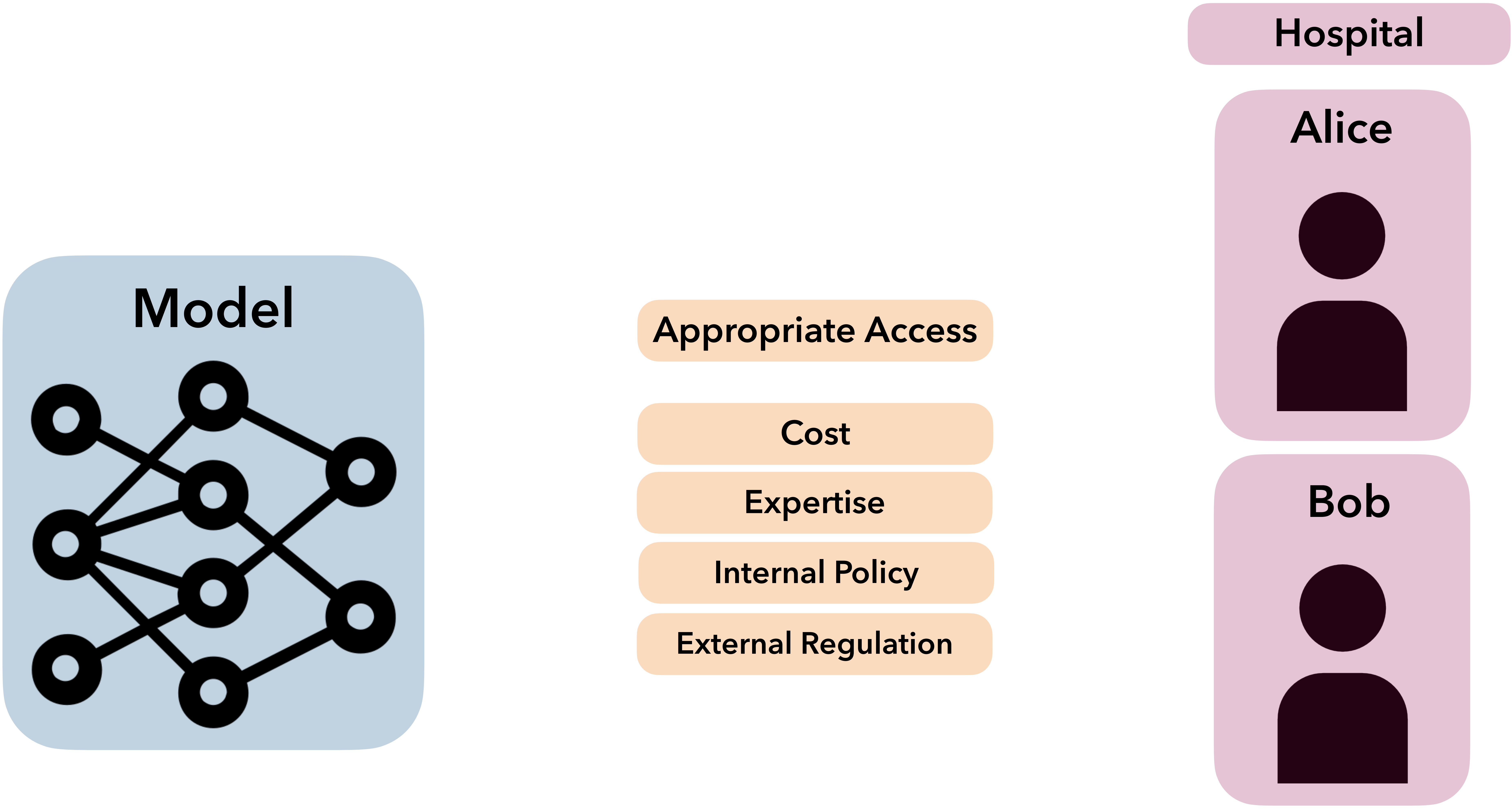
B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

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



B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

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



Model

Appropriate Access

Cost

Expertise

Internal Policy

External Regulation

Hospital

Alice

Bob

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

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

Algorithmic resignation goes 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?* Preprint. 2023.

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

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Benefits of Algorithmic Resignation

Economic Efficiency

Reputational Gain

Legal Compliance

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Considerations for Algorithmic Resignation

Directionality of Selectivity

Stakeholder Incentives

Level of Engagement

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Outline

I. What is *Algorithmic Resignation*?

II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

Why am I discussing this with a room full of lawyers?

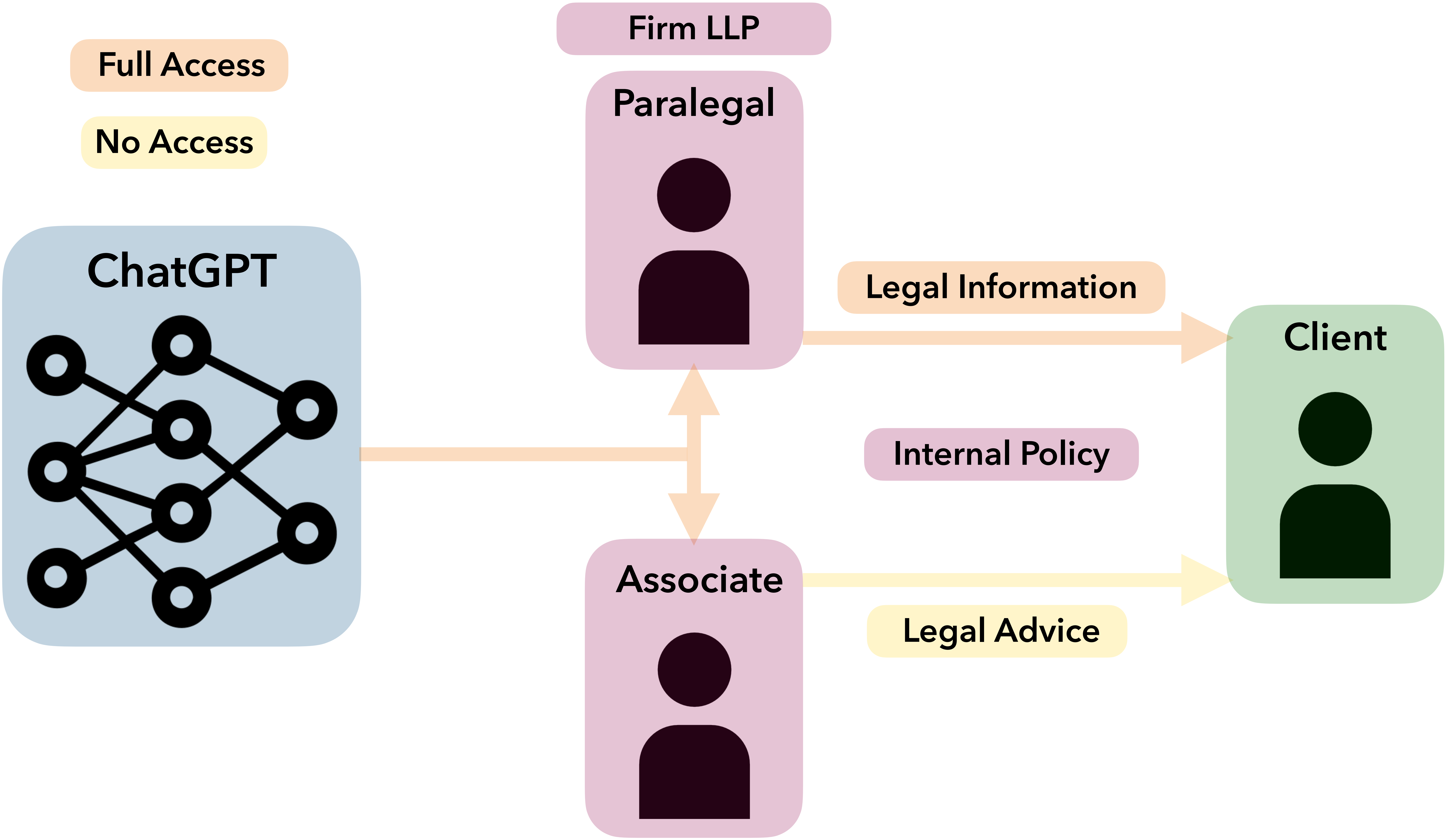
Algorithmic Resignation…

1. Enables a new mechanism for self-regulating within organizations (e.g., corporate compliance can establish policies to restrict use of AI)

2. Orchestrates human-machine collaboration to improve outcomes and processes (e.g., AI-powered content moderation tools may only escalate content to human moderators as and when needed)

3. Warrants clever interpretation of regulation like GDPR's "automated processing" since AI may now be invoked selectively (e.g., counsel can argue that AI was not used since it resigned in favor of human judgement)

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.



Full Access

No Access

Firm LLP

ChatGPT

Paralegal

Associate

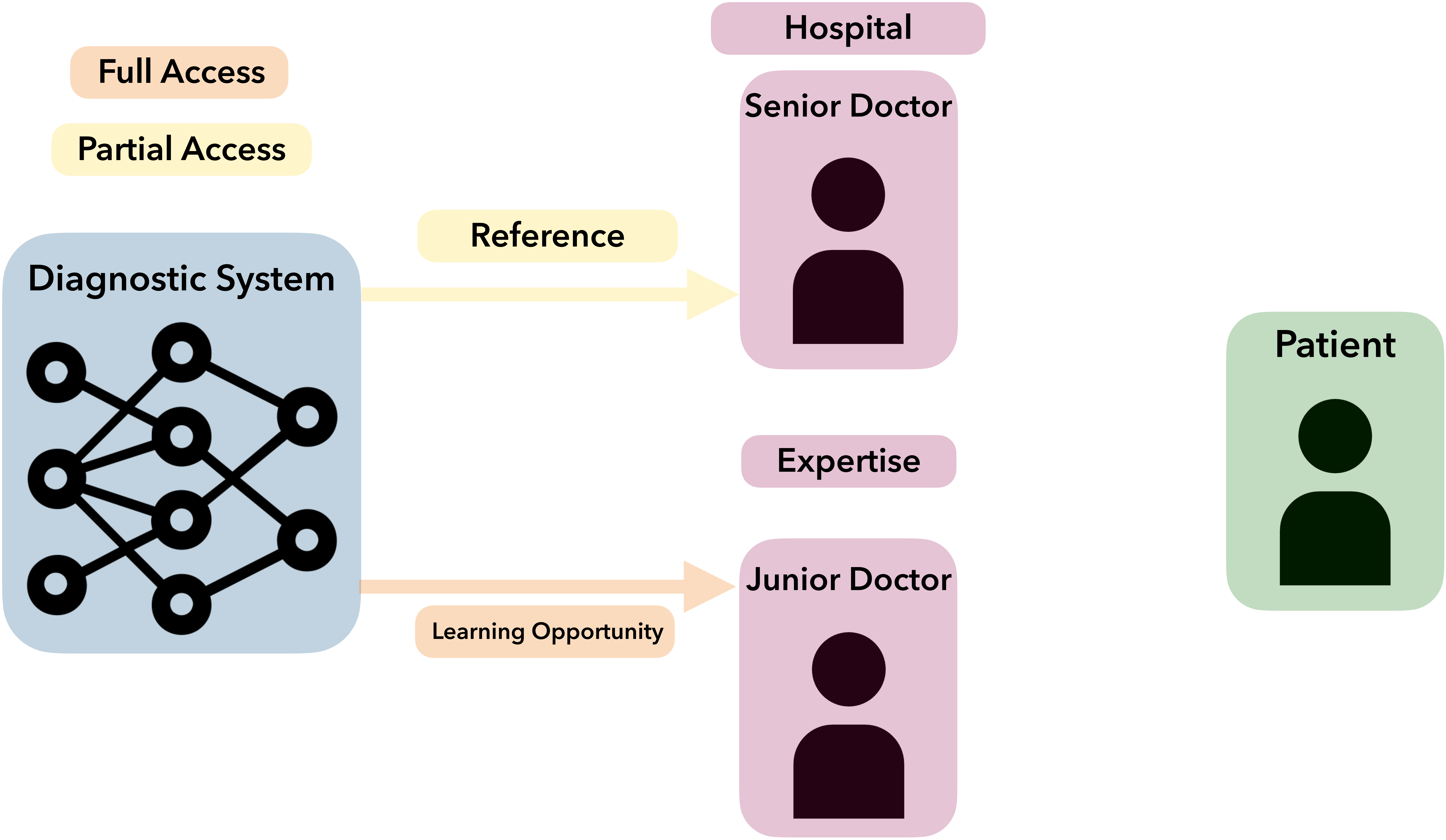
Legal Information

Internal Policy

Legal Advice

Client

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.



Full Access

Partial Access

Diagnostic System

Reference

Learning Opportunity

Hospital

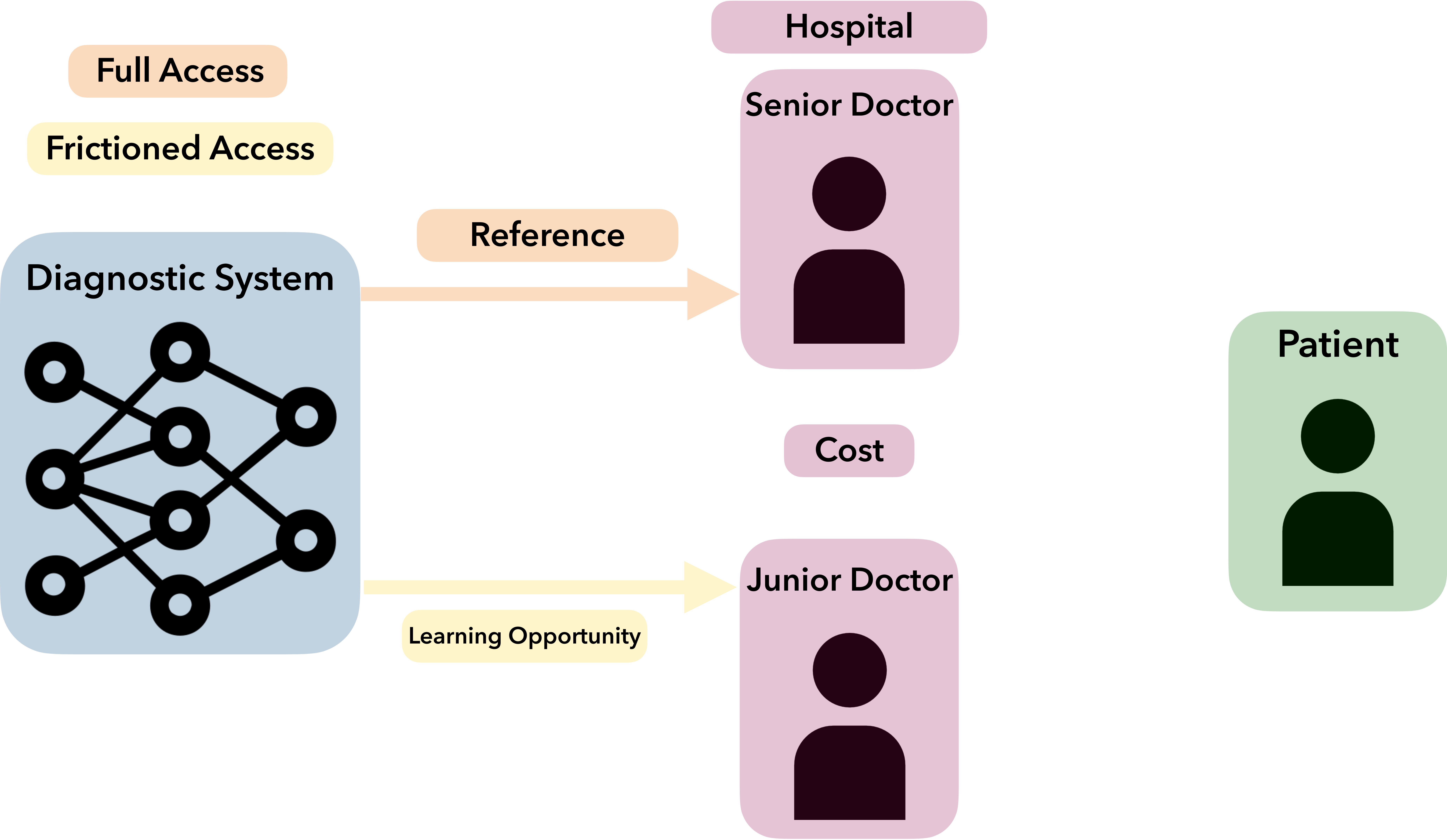
Senior Doctor

Expertise

Junior Doctor

Patient

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.



Full Access

Frictioned Access

Diagnostic System

Reference

Learning Opportunity

Hospital

Senior Doctor

Cost

Junior Doctor

Patient

B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

When Should Algorithms Resign?

Thank you to my collaborators!

John Zerilli
Edinburgh

P. Kamalaruban
Turing

Emma Kallina
Cambridge

Katie Collins
Cambridge

Adrian Weller
Cambridge

Holli Sargeant
Berkman Klein

Valerie Chen
CMU

Ameet Talwalkar
CMU

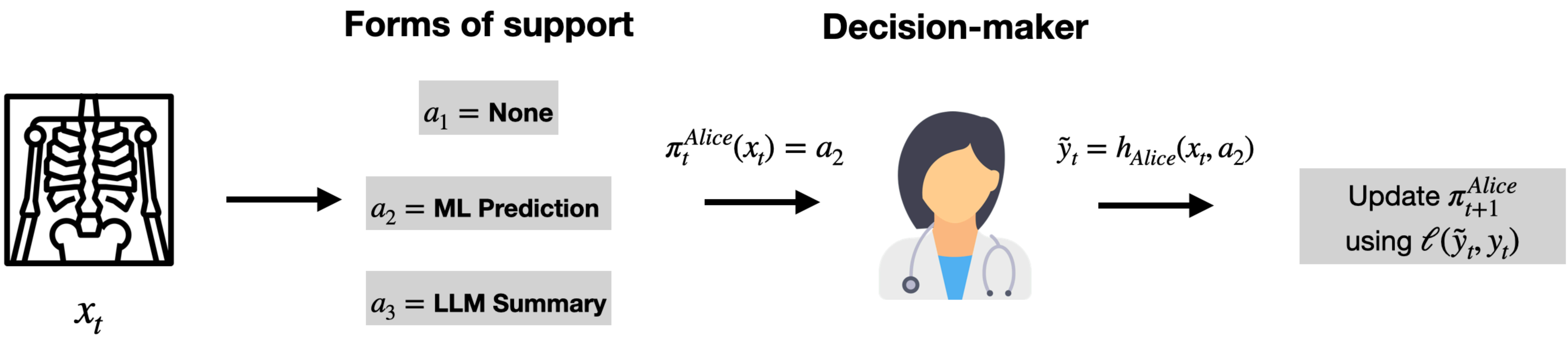
@umangsbhatt

umangbhatt@nyu.edu

Appendix

Learning Personalized Decision Support Policies

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**

a_3 = **LLM Summary**

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

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

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

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.

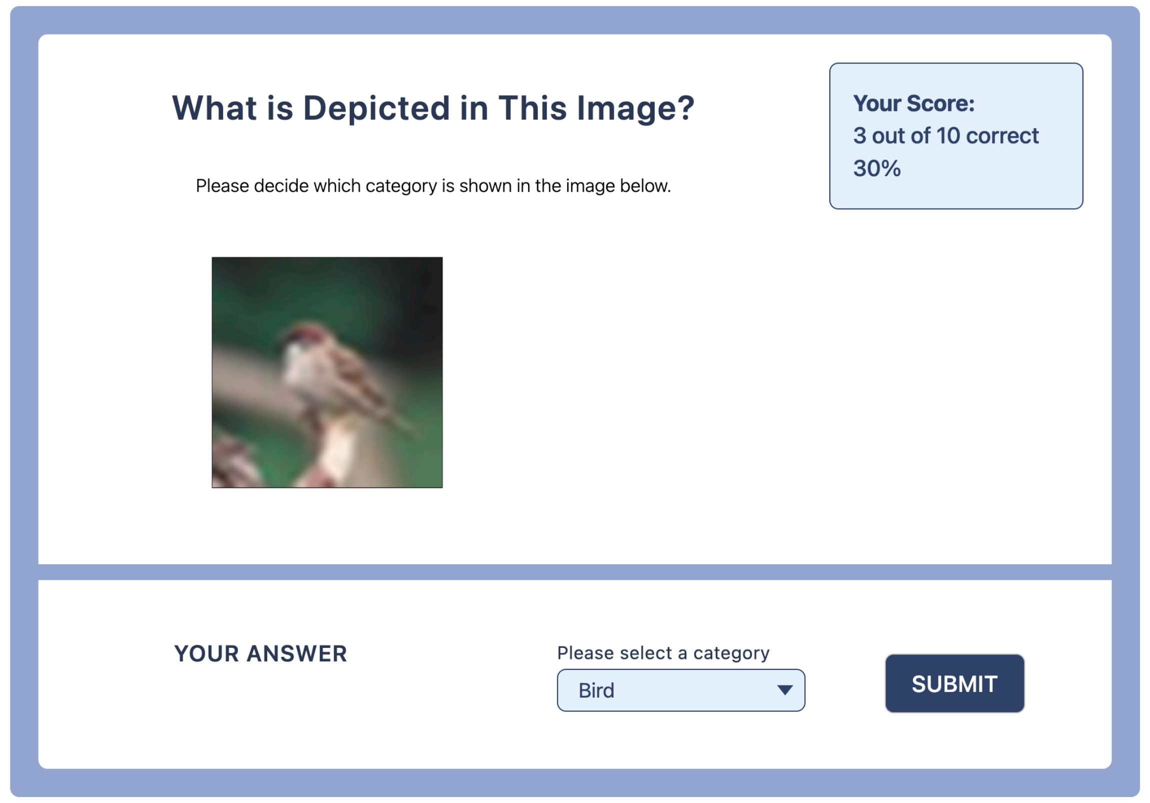
Decision Maker

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.



What is Depicted in This Image?

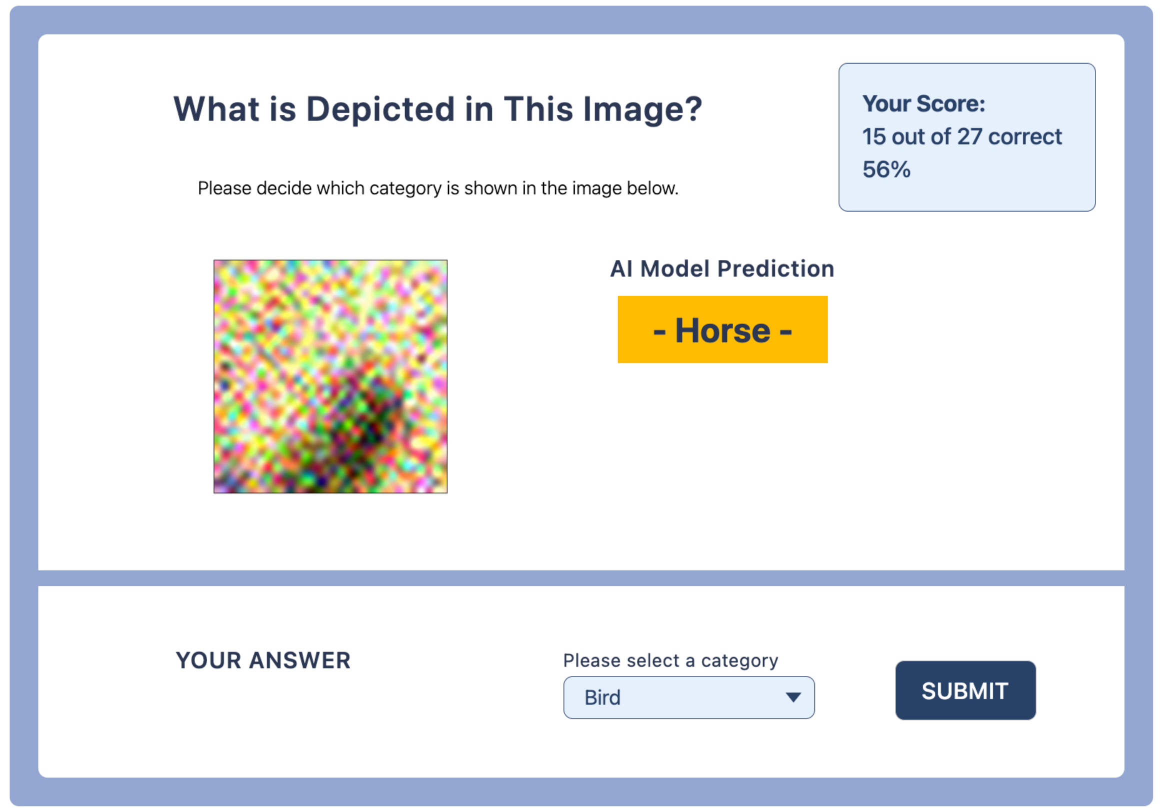
Please decide which category is shown in the image below.

Your Score:
3 out of 10 correct
30%

YOUR ANSWER

Please select a category

Bird



What is Depicted in This Image?

Please decide which category is shown in the image below.

Your Score:
15 out of 27 correct
56%

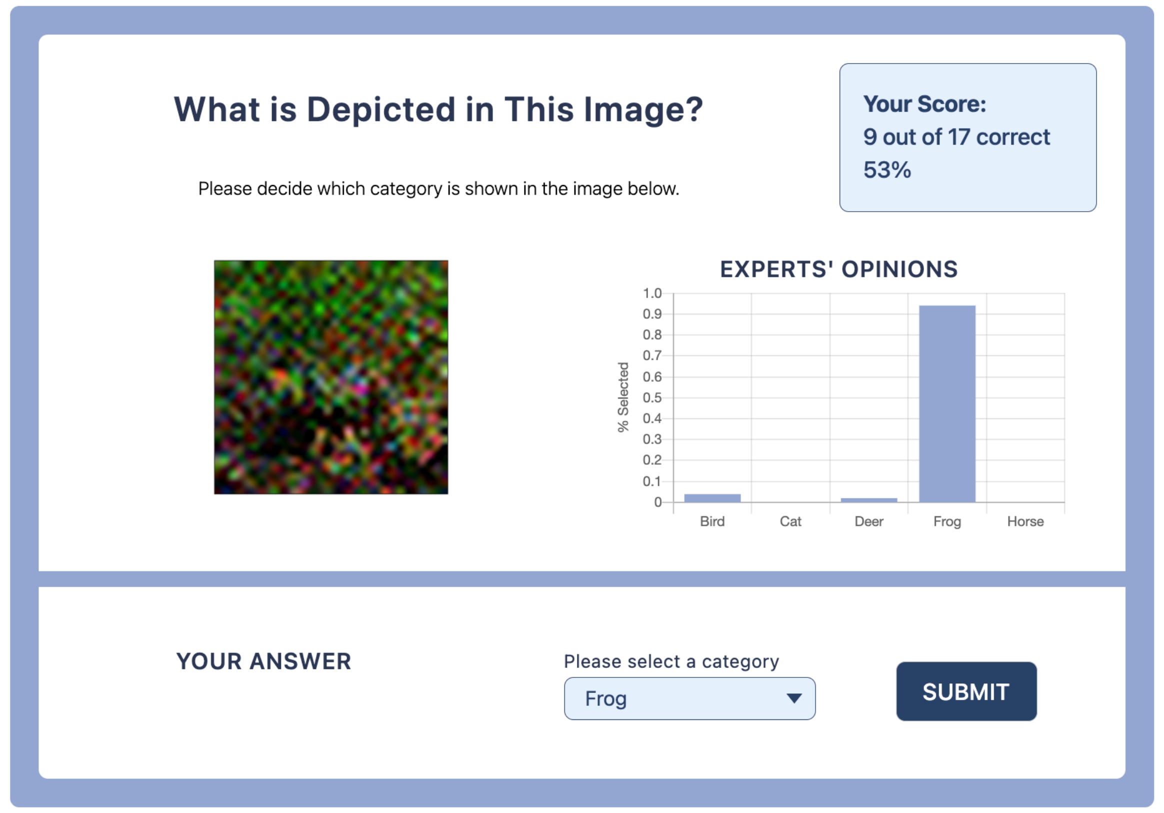
AI Model Prediction

- Horse -

YOUR ANSWER

Please select a category

Bird



What is Depicted in This Image?

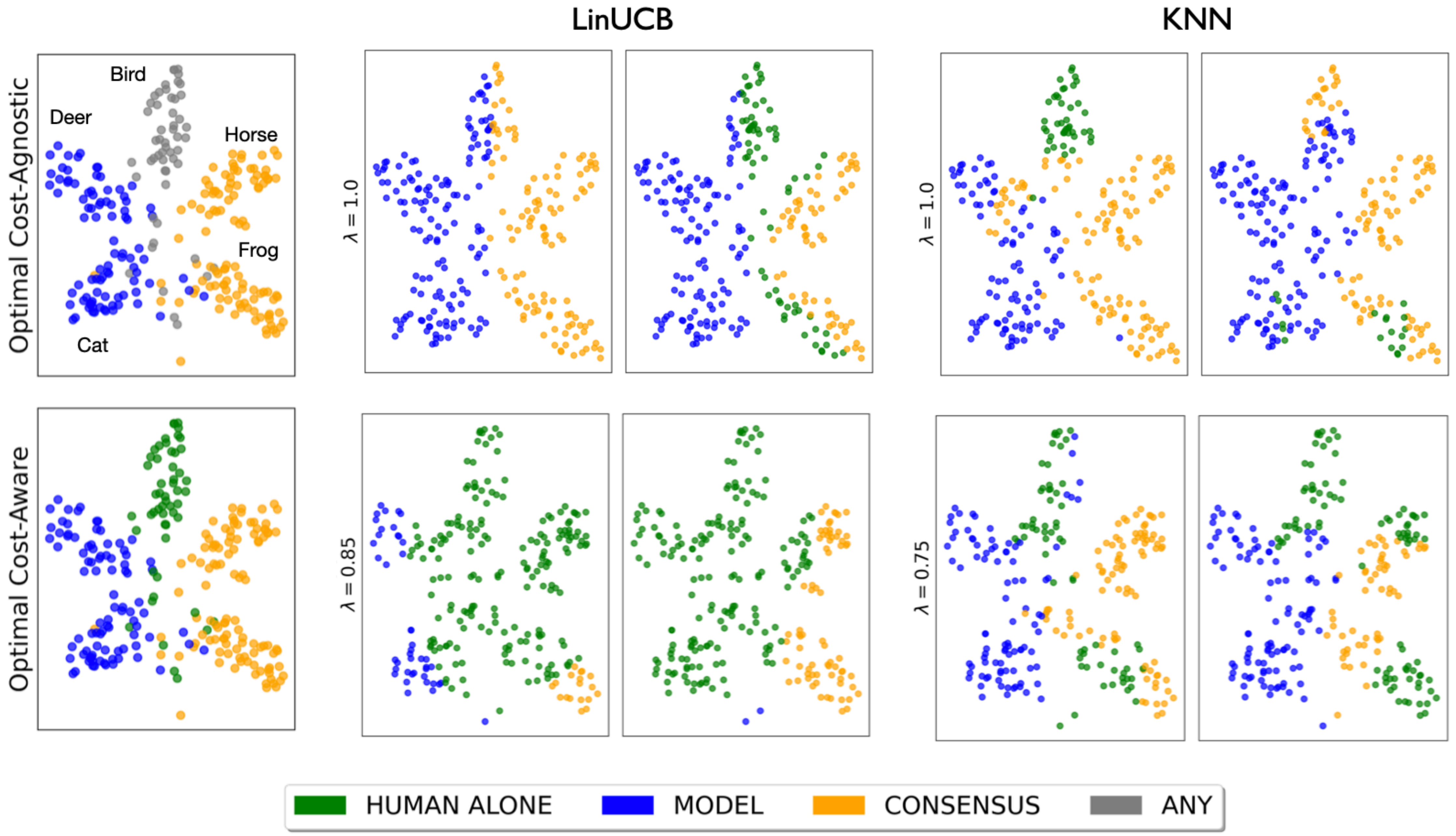
Please decide which category is shown in the image below.

Your Score:
9 out of 17 correct
53%

YOUR ANSWER

Please select a category

Frog



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

Decision Maker

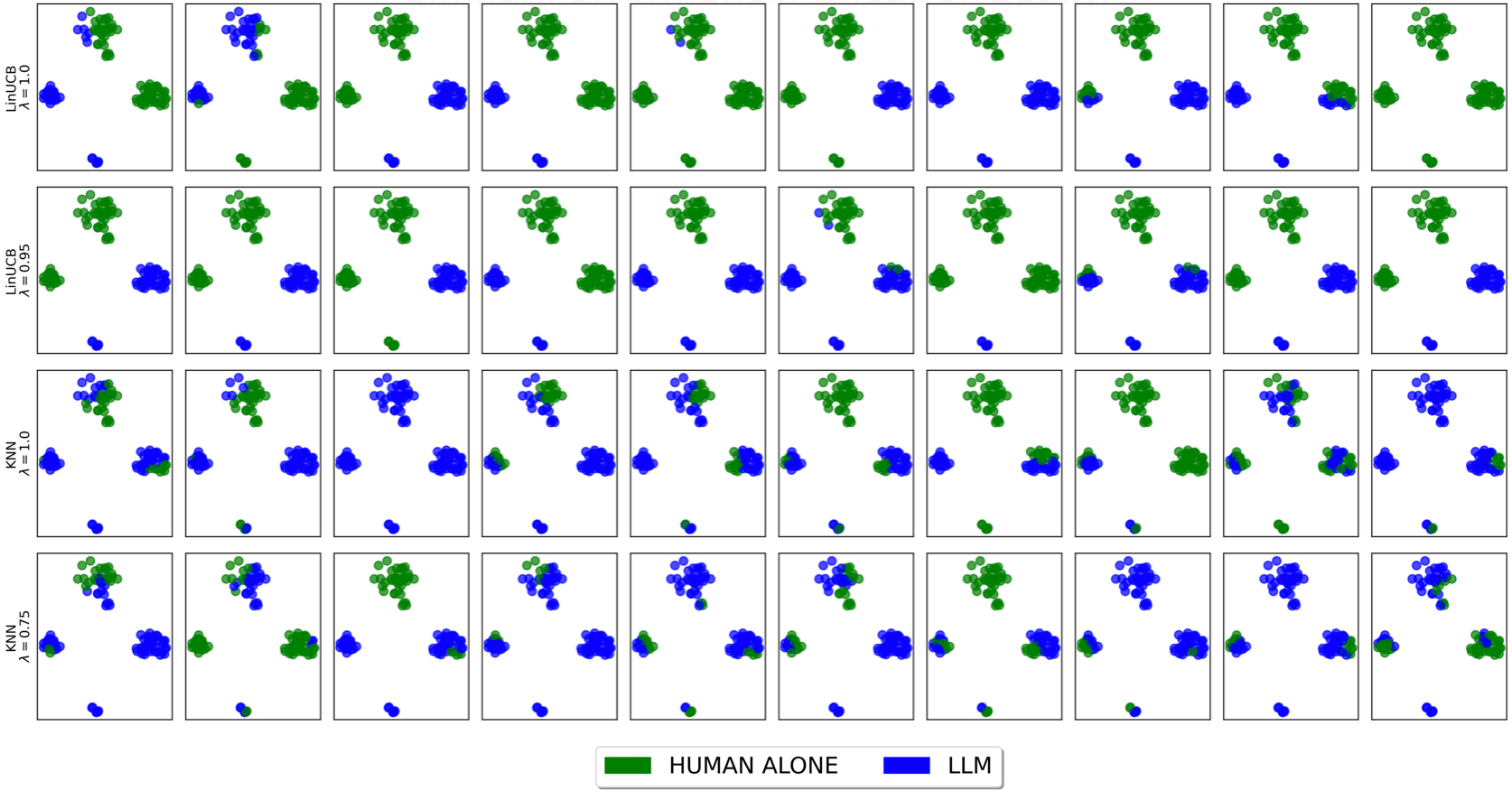
Personalize Access

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.

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