Algorithmic Resignation

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

The Alan Turing Institute

CFI

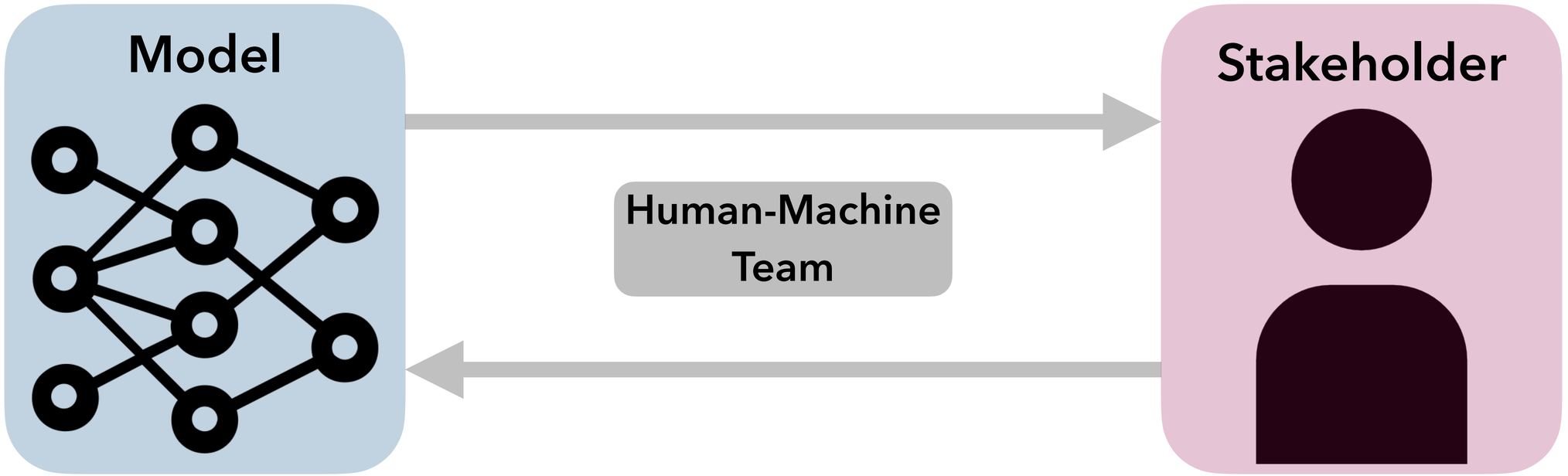
When Should Algorithms Resign?

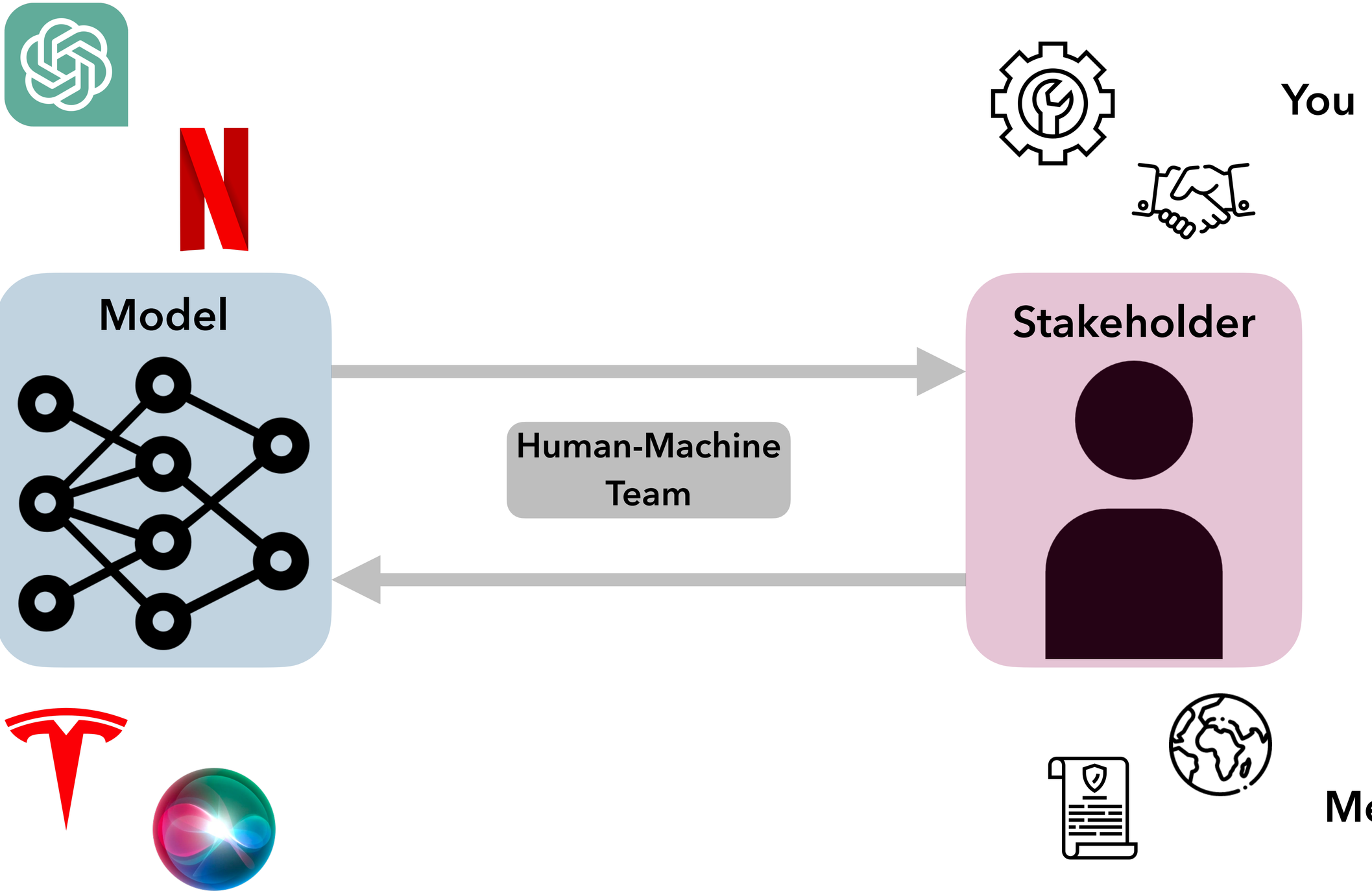
Umang Bhatt
Assistant Professor/Faculty Fellow, New York University
Research Associate, The Alan Turing Institute
Associate Fellow, Leverhulme Center for the Future of Intelligence

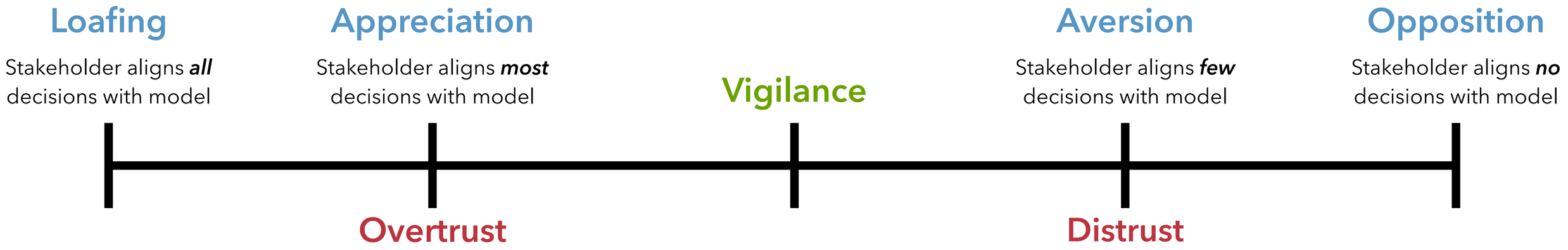
@umangsbhatt
umangbhatt@nyu.edu

The Alan Turing Institute

CFI


Model
Human-Machine
Team
Stakeholder


You
Model
Stakeholder
Human-Machine
Team
Me



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

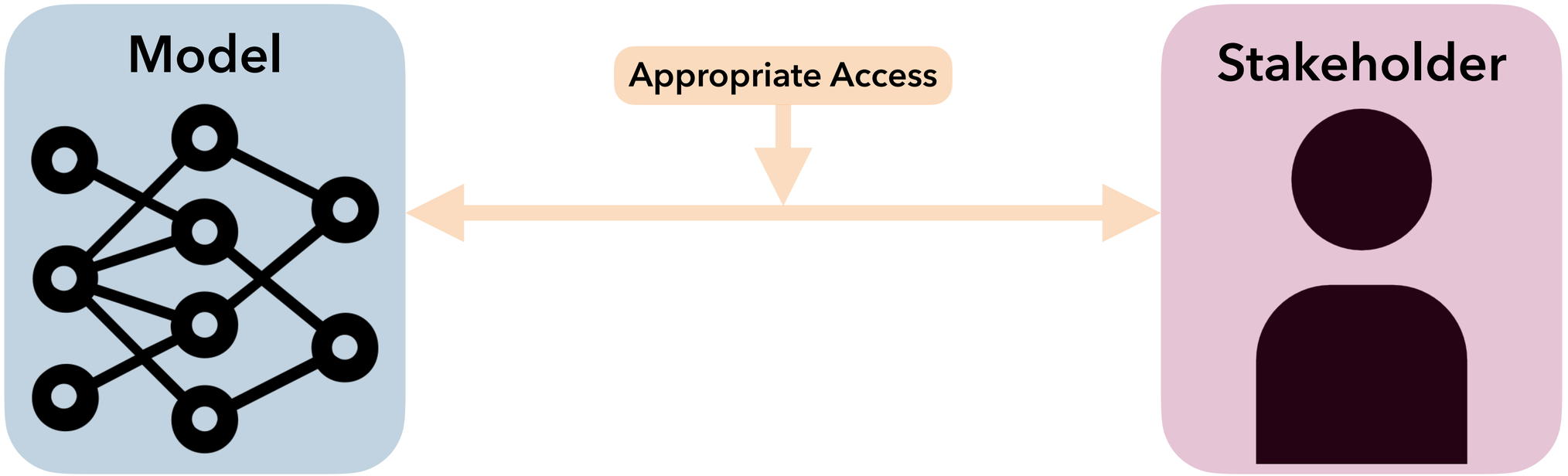
Thomas Claburn Tue 8 Aug 2023 | 00:24 UTC

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.

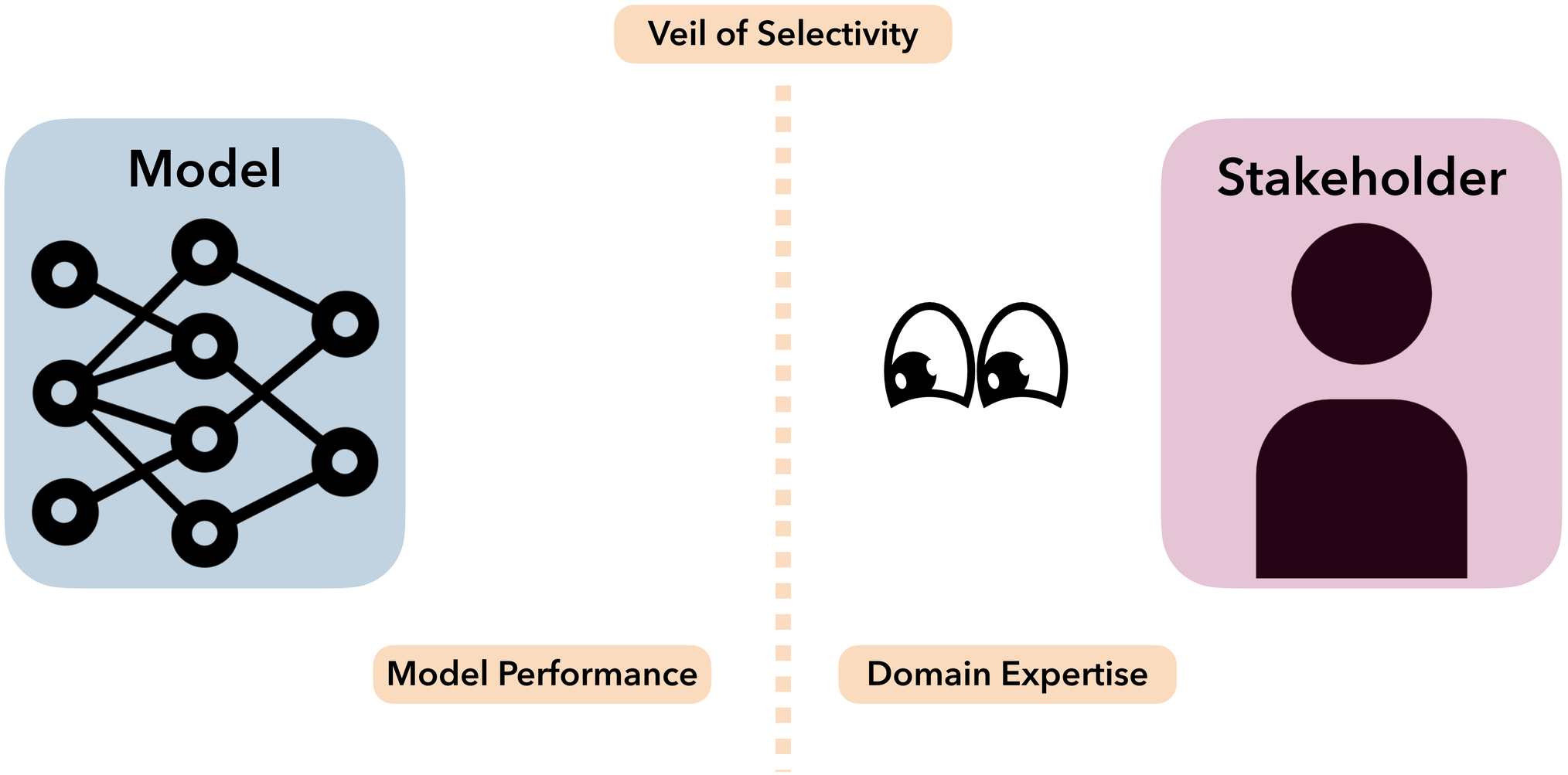
95

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.



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*?

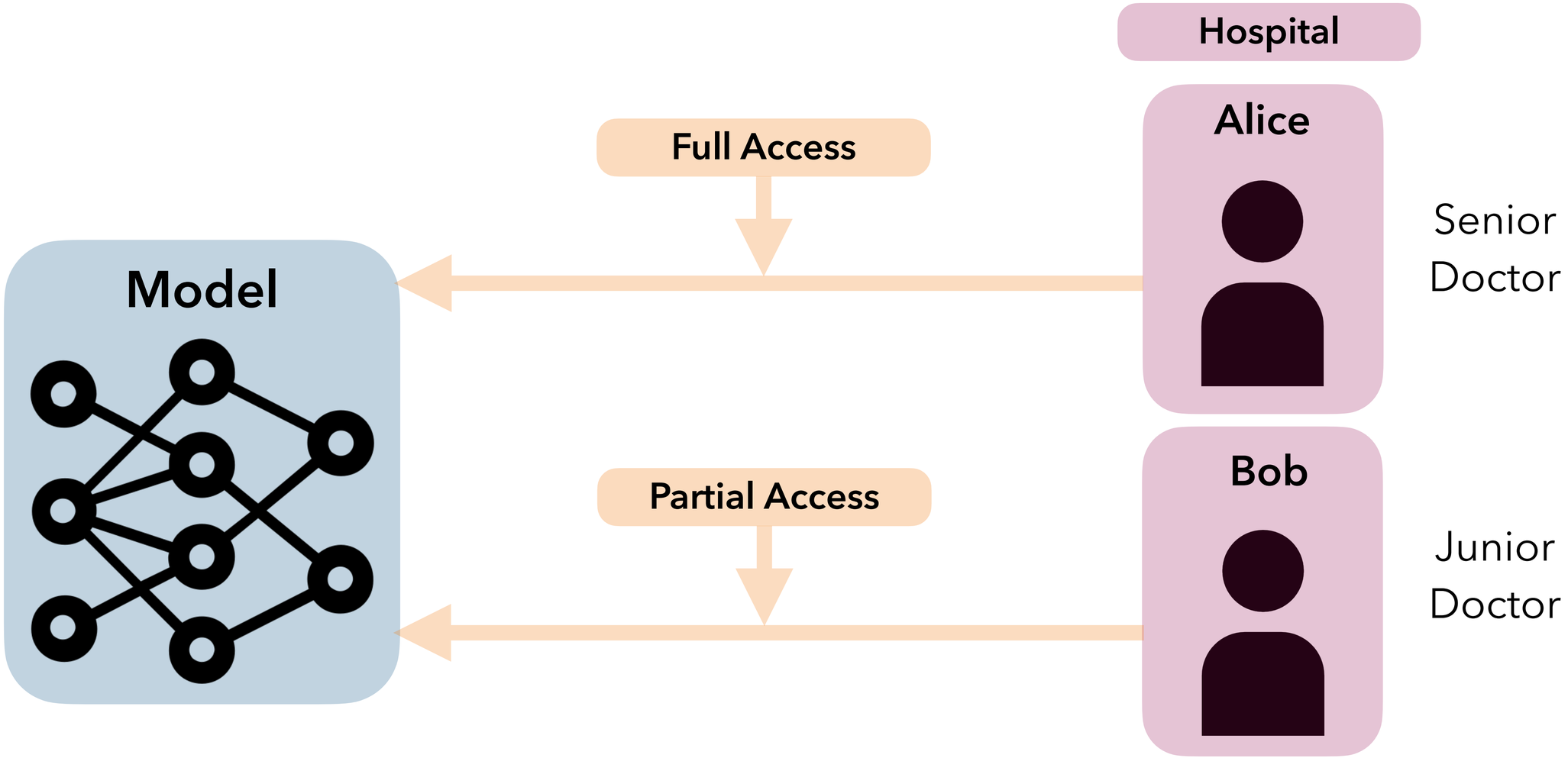
II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

IV. *Algorithmic Resignation* in Practice

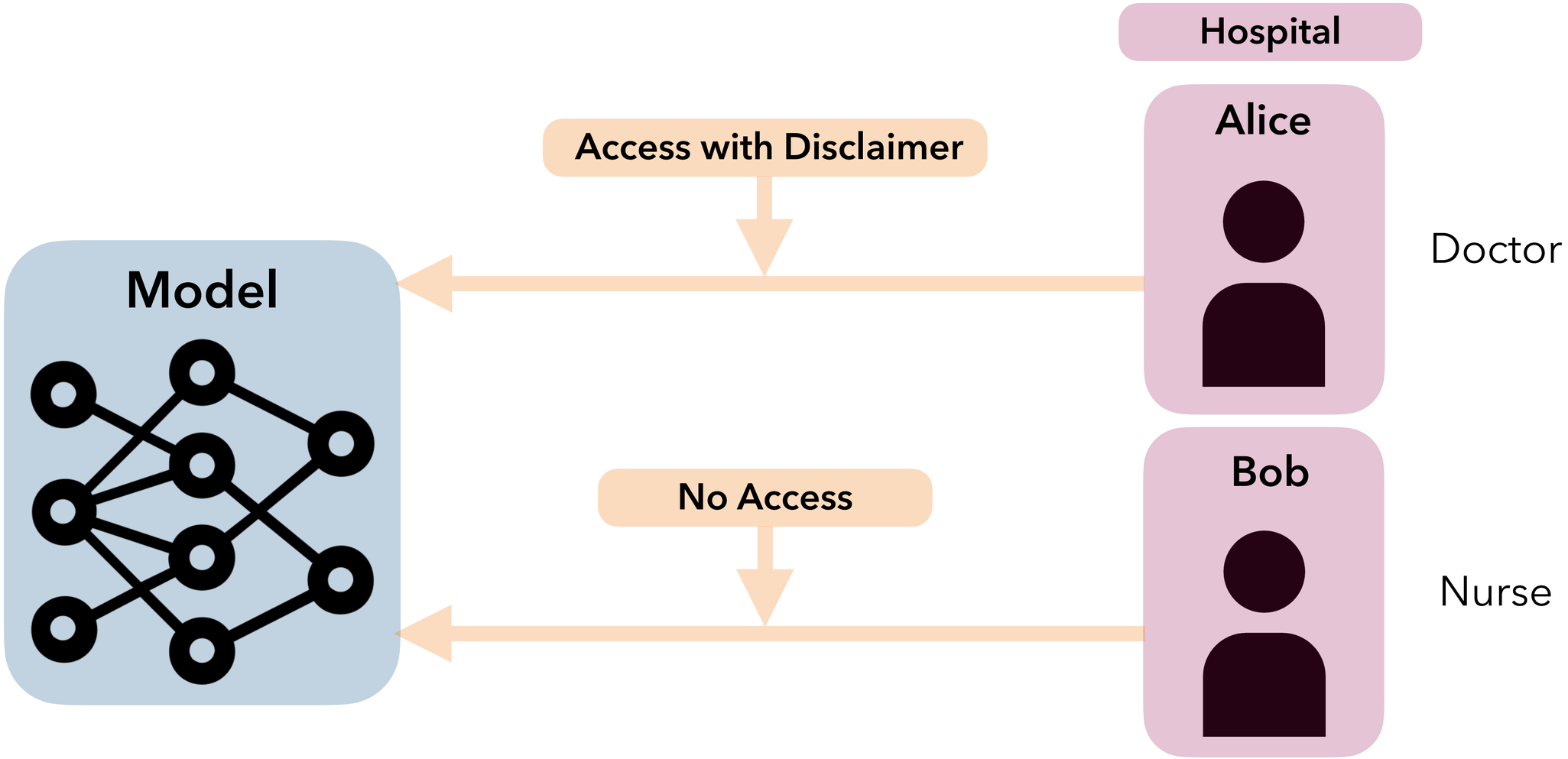
Algorithmic resignation is the *deliberate* and *informed* disengagement from AI assistance in certain scenarios.

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.

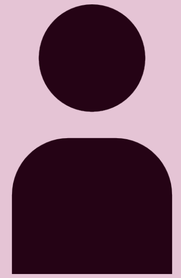


B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.

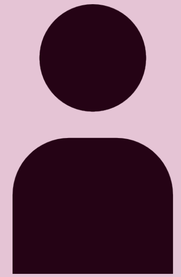
B*, Chen*, Collins, P. Kamalaruban, Kallina, Weller, Talwalkar. *Learning Personalized Decision Support Policies.* Under Review. 2023.

Hospital

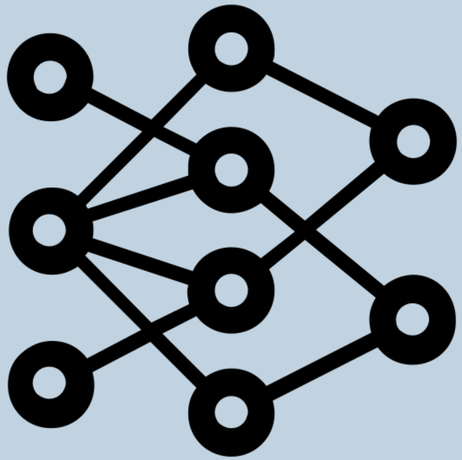
Alice



Bob



Model



Appropriate Access

Cost

Expertise

Internal Policy

External Regulation

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*?

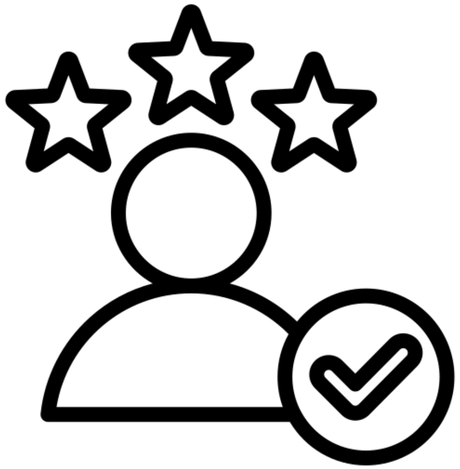
II. Benefits of *Algorithmic Resignation*

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*?

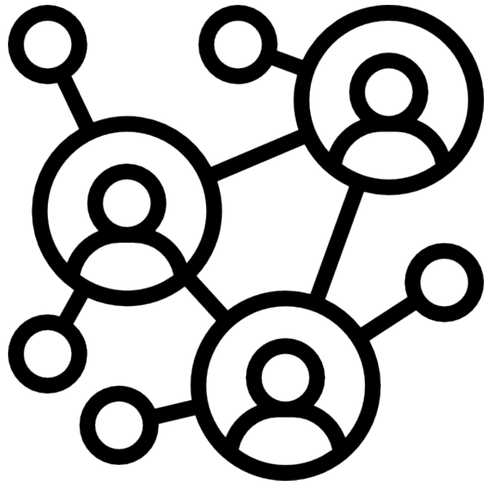
II. Benefits of *Algorithmic Resignation*

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

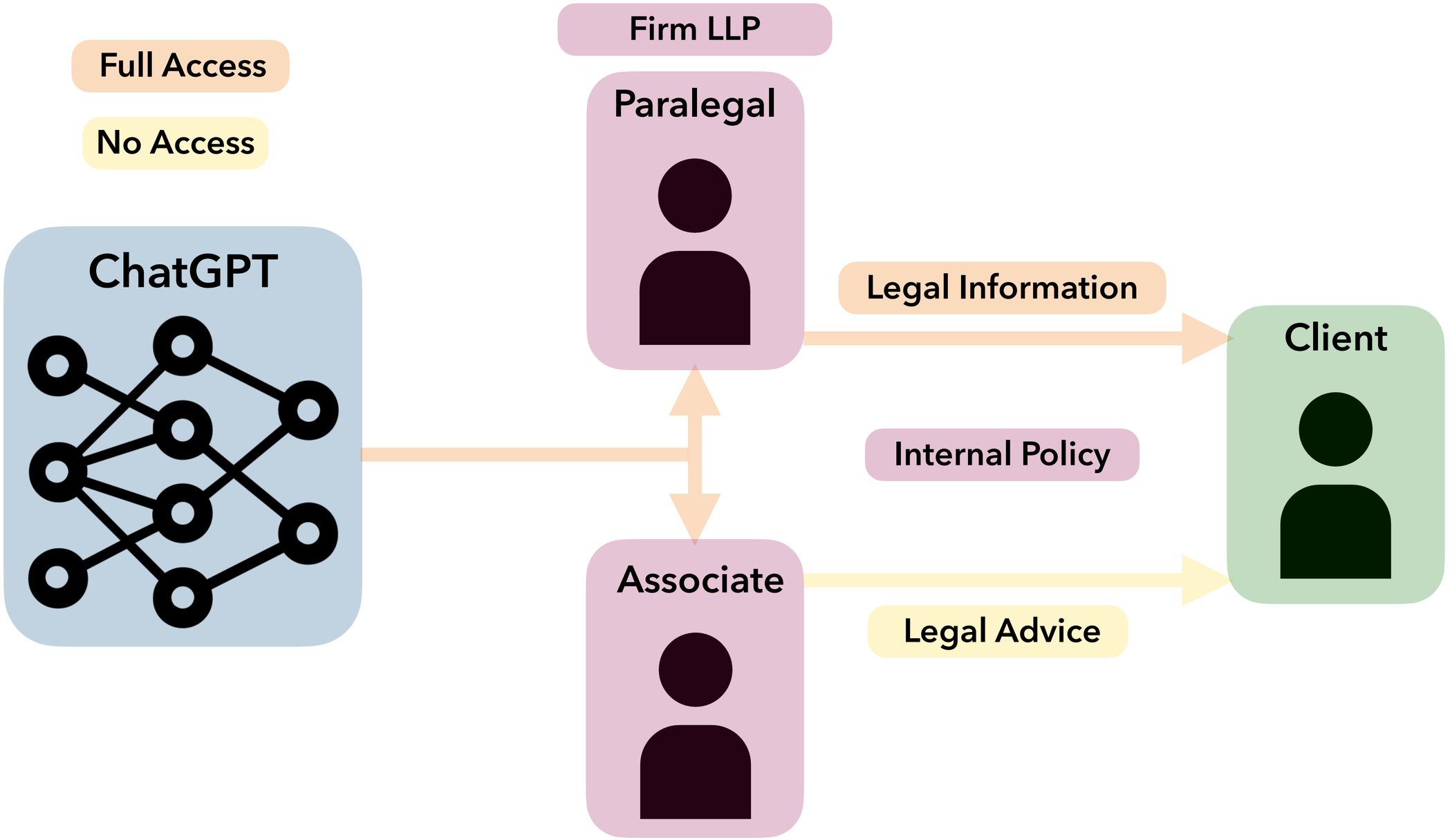
Outline

I. What is *Algorithmic Resignation*?

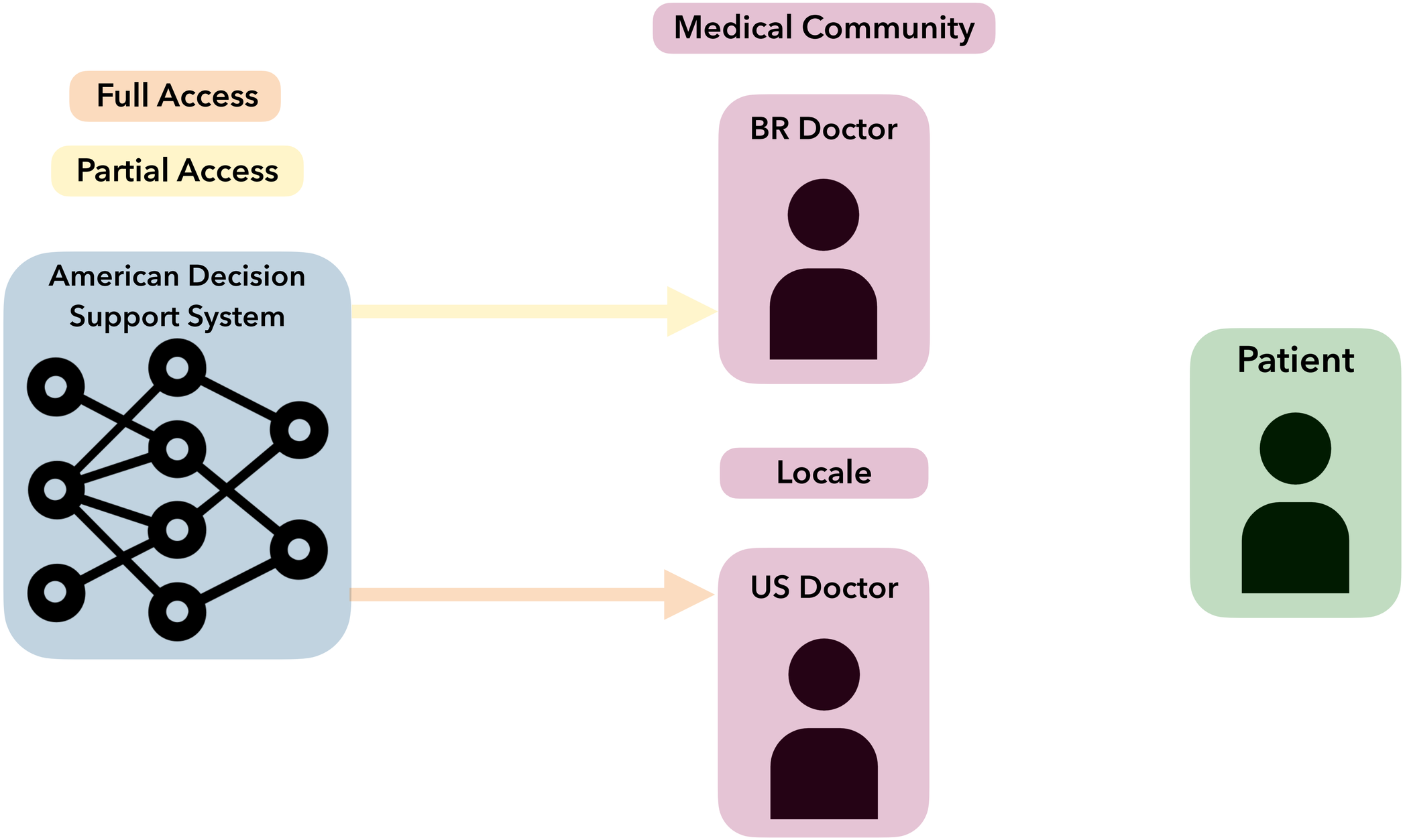
II. Benefits of *Algorithmic Resignation*

III. Considerations for *Algorithmic Resignation*

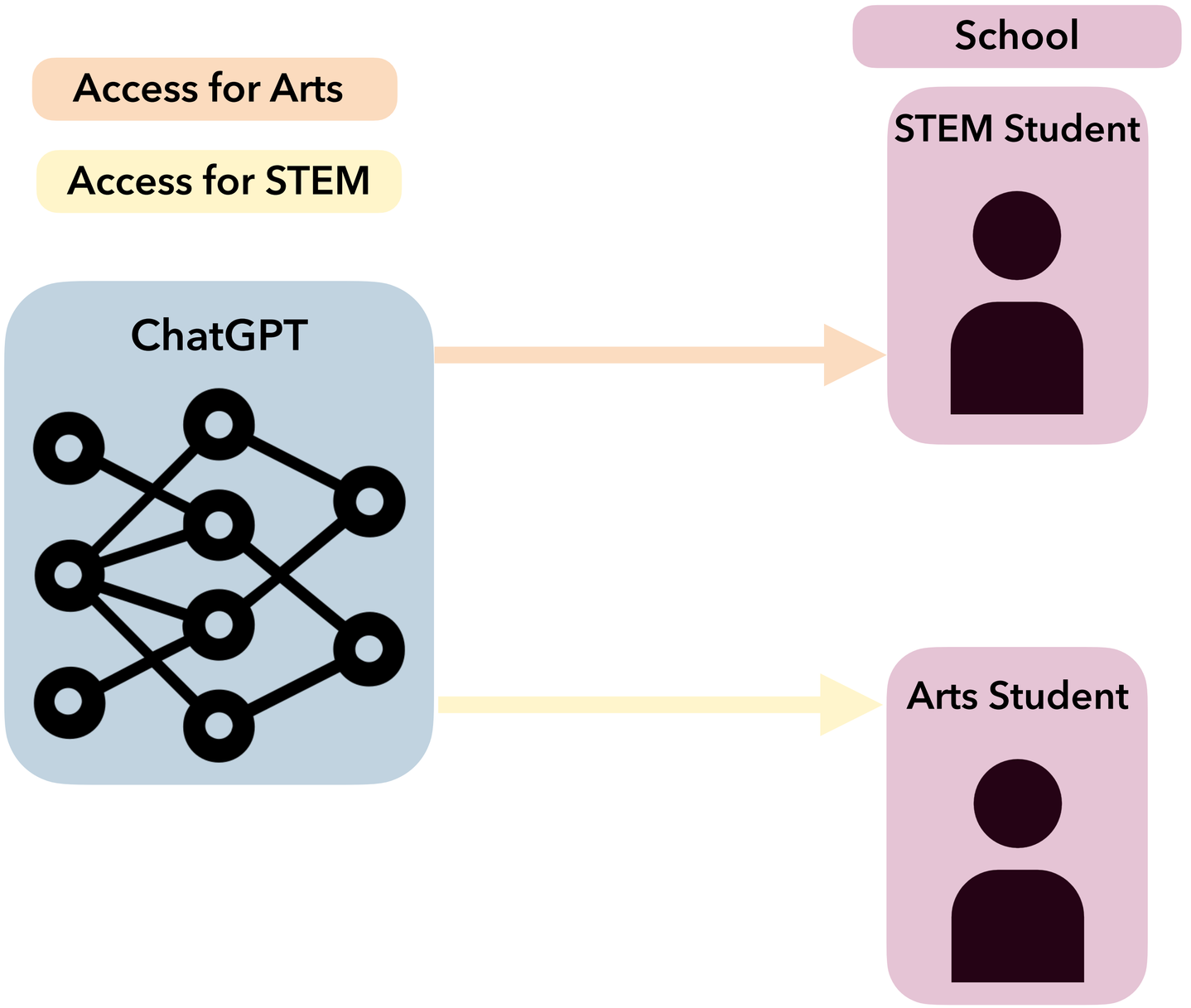
IV. *Algorithmic Resignation* in Practice



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



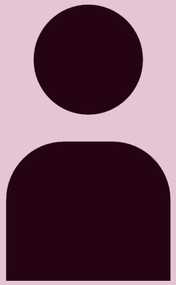
B*, Sargeant*. *When Should Algorithms Resign?* Preprint. 2023.



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?* Preprint. 2023.

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?"*

x_t

Forms of support

a_1 = **None**

a_2 = **ML Prediction**

a_3 = **LLM Summary**

Decision-maker

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

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

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: $r_{A_i}(x; h) = \mathbb{E}_{y|x}[\ell(y, h(x, A_i))]$

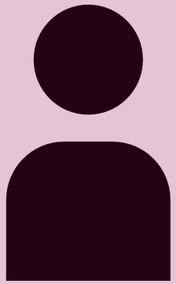
Core Idea of THREAD

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.

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

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

Excess loss over optimal loss

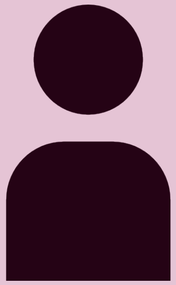
MMLU

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.

Student



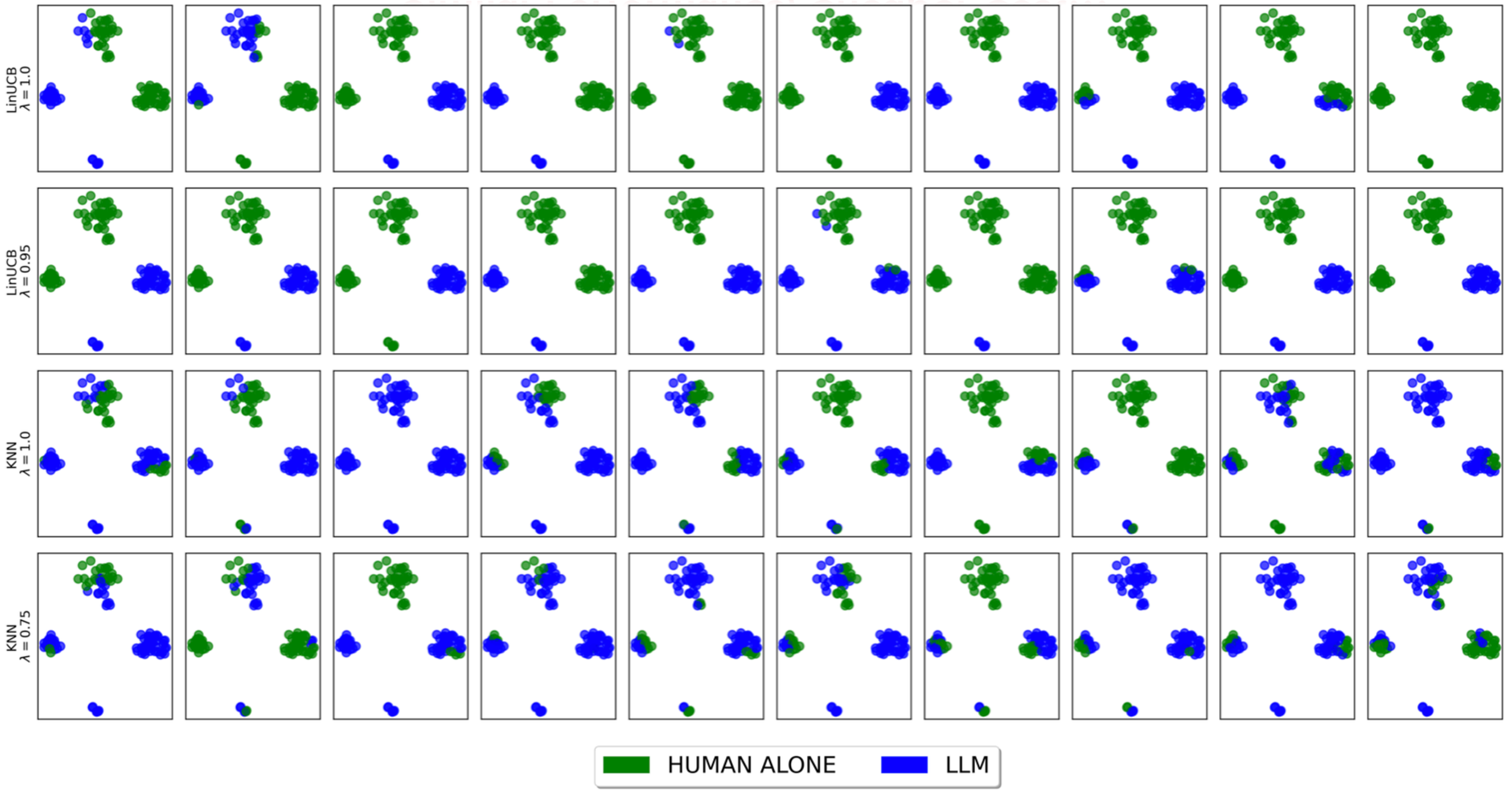
Personalize Access

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.

Takeaways

Resignation is a new mechanism for self-regulating (e.g., corporate compliance can establish policies to restrict AI use).

Algorithmic resignation 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).

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

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

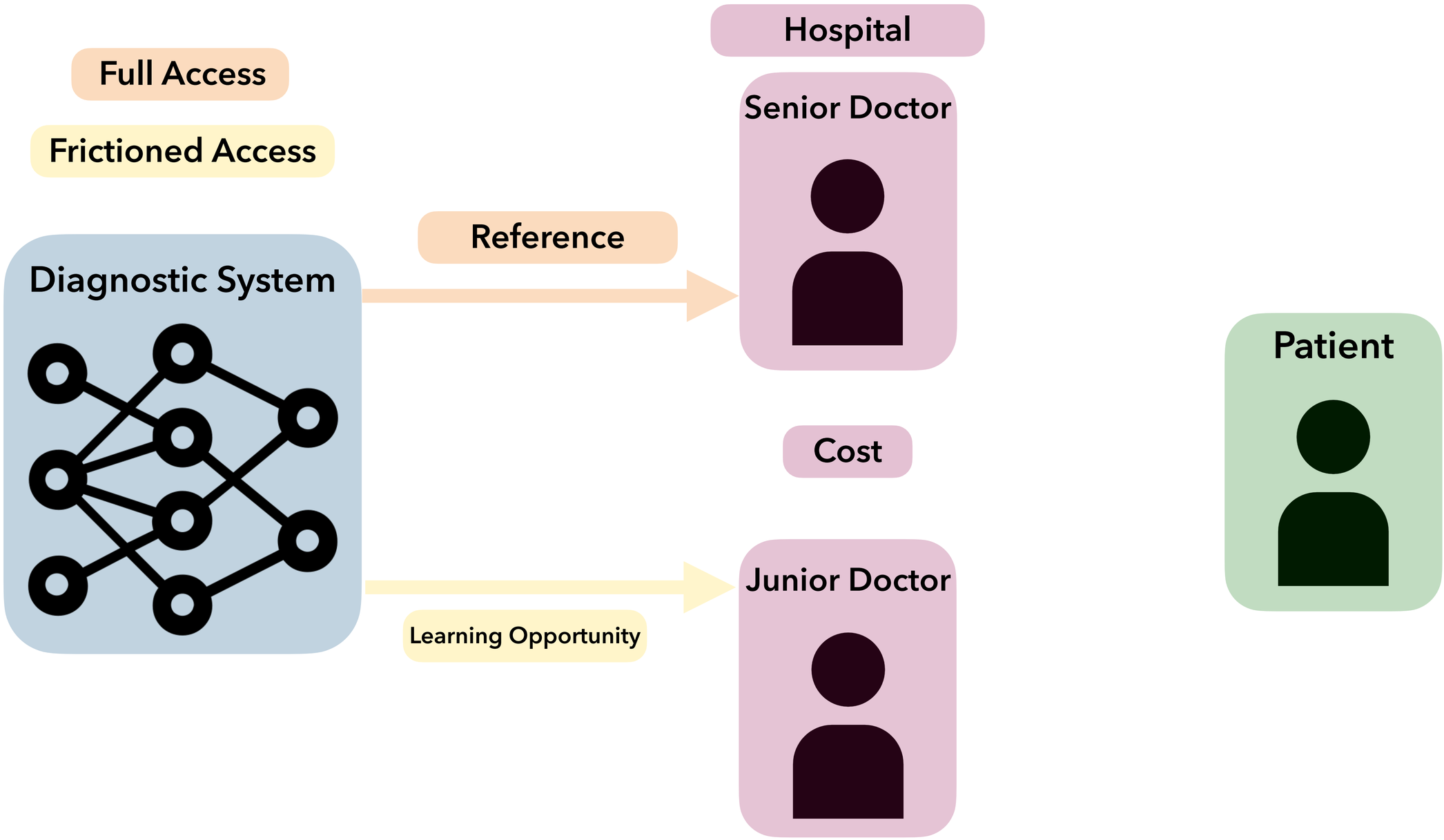

Valerie Chen
CMU


Ameet Talwalkar
CMU

@umangsbhatt

umangbhatt@nyu.edu

Appendix



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

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	$\mathbf{0.05 \pm 0.03}$	$\mathbf{0.05 \pm 0.03}$

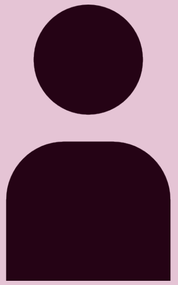
If a decision-maker benefits from having support some of the time, we can learn their policy online

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

Learning Personalized Decision Support Policies

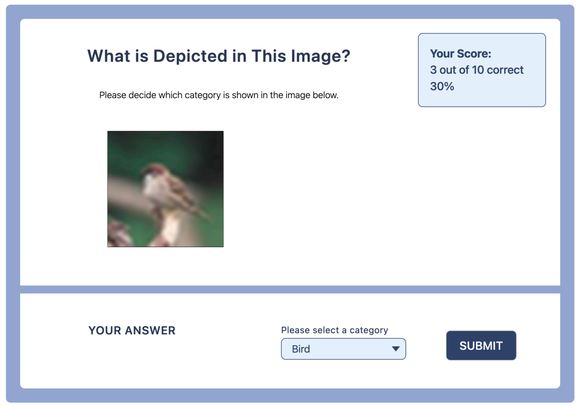
User Studies

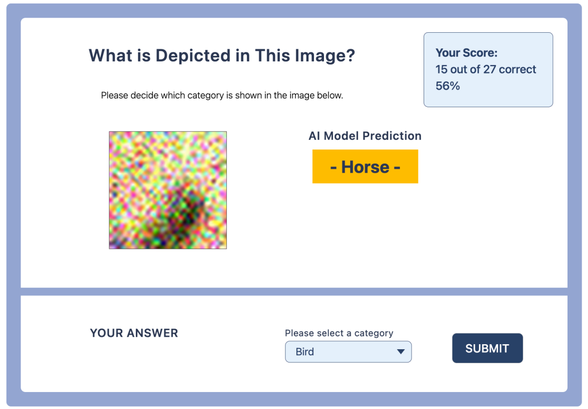
Decision Maker

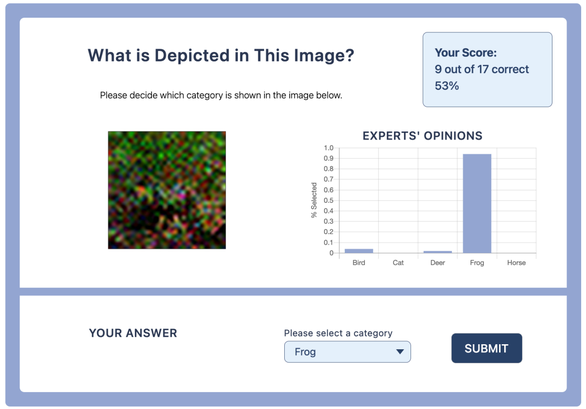


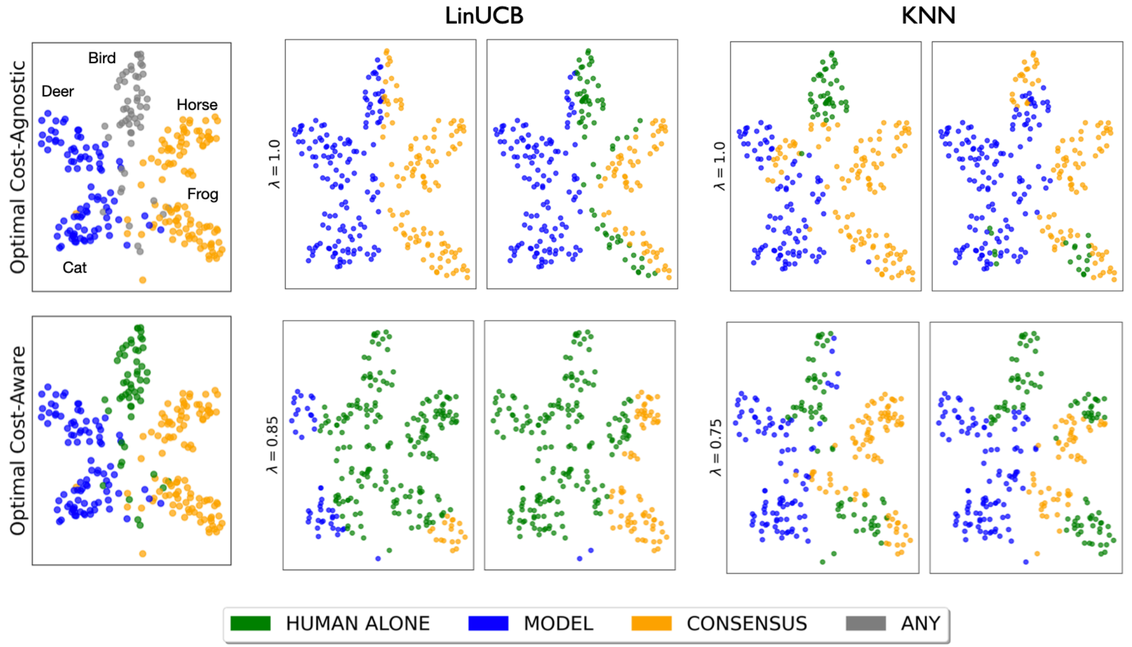
Personalize Access

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.