Introduction to Human Centered Artificial Intelligence

Swati Mishra

Human Centered Artificial Intelligence

Graduate Course - CAS 783

Winter 2025

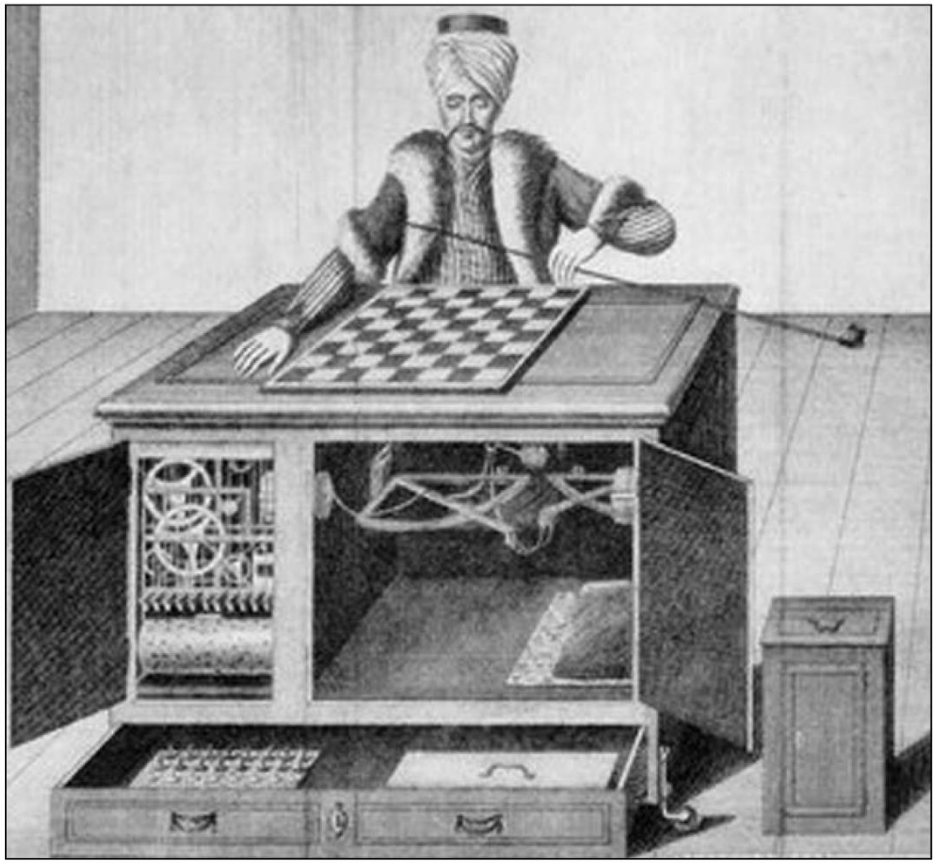
McMaster University
ENGINEERING

A Brief History of AI

"The history of AI is a history of fantasies, possibilities, demonstrations, and promise."

Source: *A (Very) Brief History of Artificial Intelligence, ACM Magazine*

A Brief History of AI



The Turk, from a 1789 Engraving by Freiherr Joseph Friedrich zu Racknitz.

" Chess playing machines of the Eighteenth relied on mechanical operations to simulate human thinking."

Picture Source: *A (Very) Brief History of Artificial Intelligence, ACM Magazine*

A Brief History of AI

Augusta Ada King, Countess of Lovelace

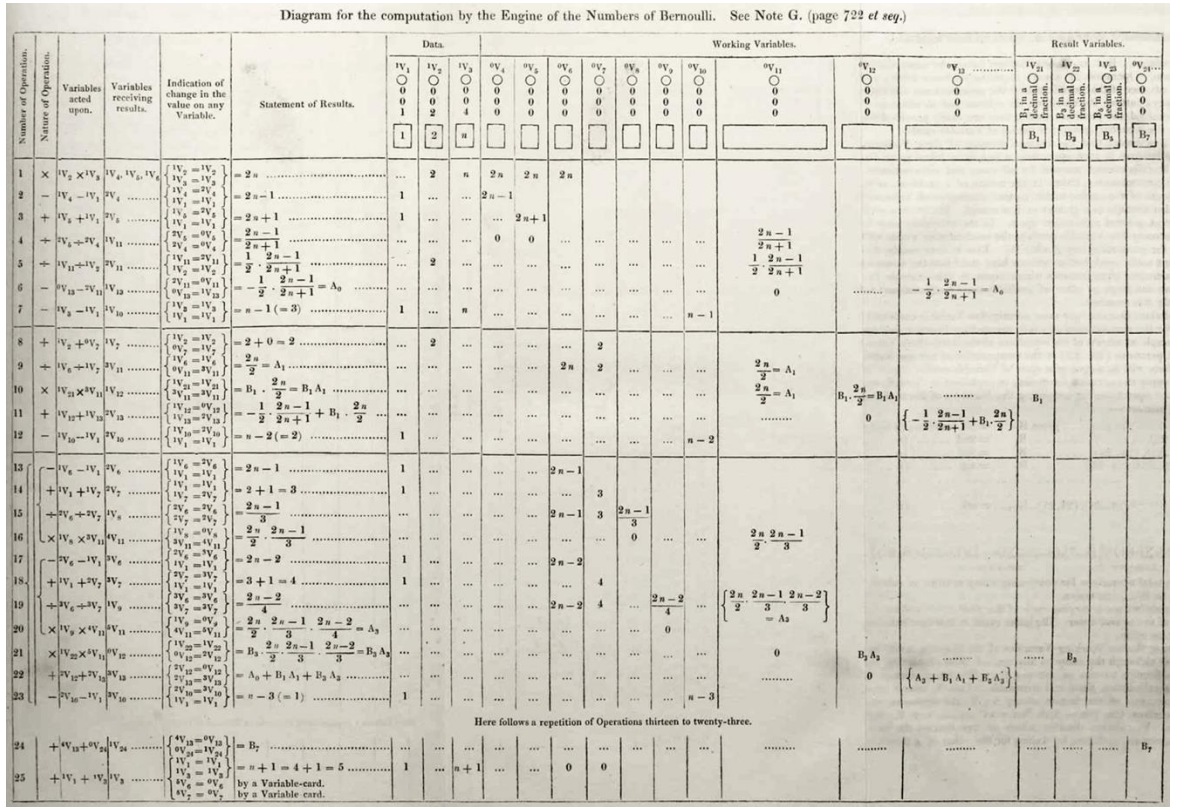


Diagram for the computation by the Engine of the Numbers of Bernoulli. See Note G. (page 722 *et seq.*)

Picture Source: *Wikipedia*

A Brief History of AI

On October 8, 2005, the Stanford Racing Team's Autonomous Robotic Car, Stanley, Won the Defense Advanced Research Projects Agency's (DARPA) Grand Challenge.

Photo courtesy, DARPA.

A Brief History of AI

Mars Rover.

Photo Courtesy, NASA

A Brief History of AI

"… AI is not just about robots (or machines).
It is also about understanding the nature of
intelligent thought and action using computers
as experimental devices…

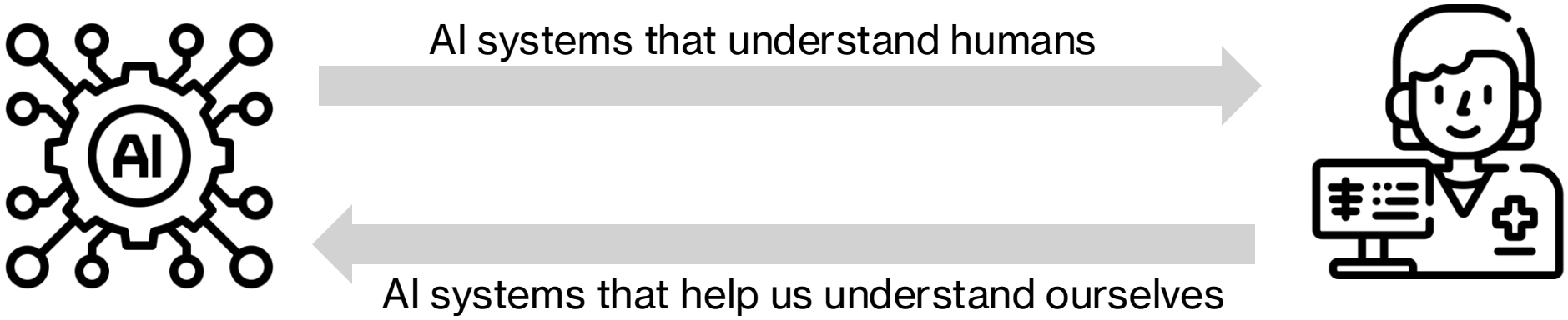
Source: *A (Very) Brief History of Artificial Intelligence, ACM Magazine*

Human Centered Artificial Intelligence

"Human-centered artificial intelligence is a perspective on AI and ML that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stake-holders, such as users, operators, clients, and other people in close proximity."

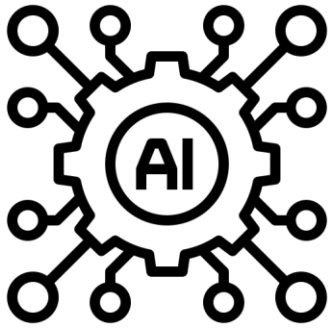
Source: Human-Centered Artificial Intelligence and Machine Learning, Mark O.Riedl

Human Centered Artificial Intelligence



Human Centered Artificial Intelligence

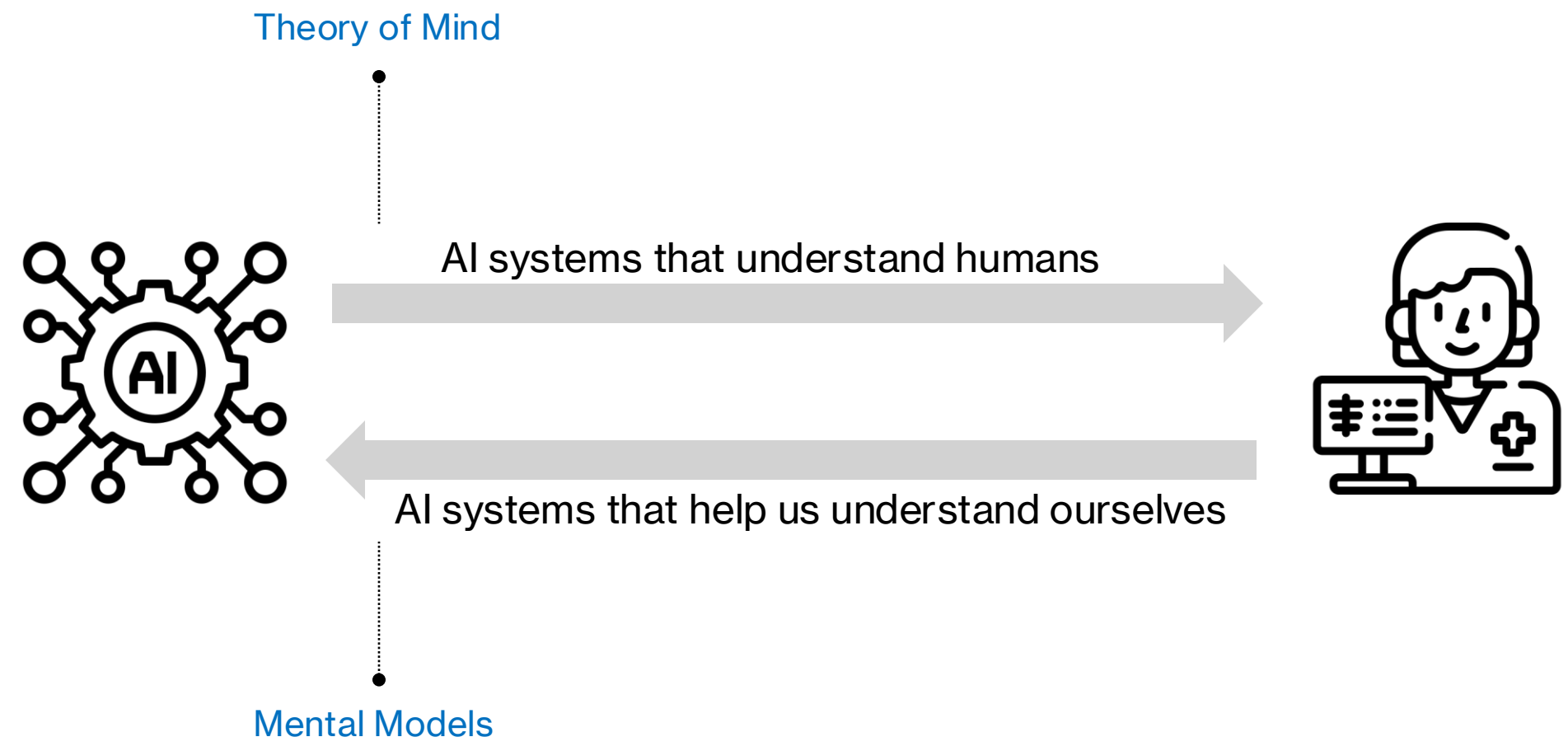
Theory of Mind



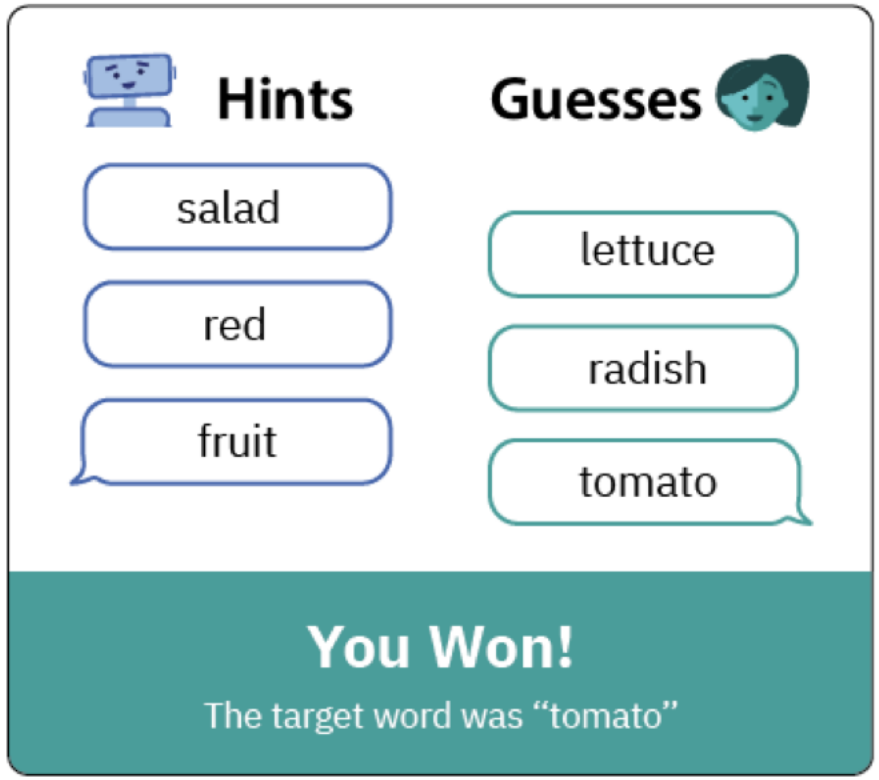
AI systems that understand humans

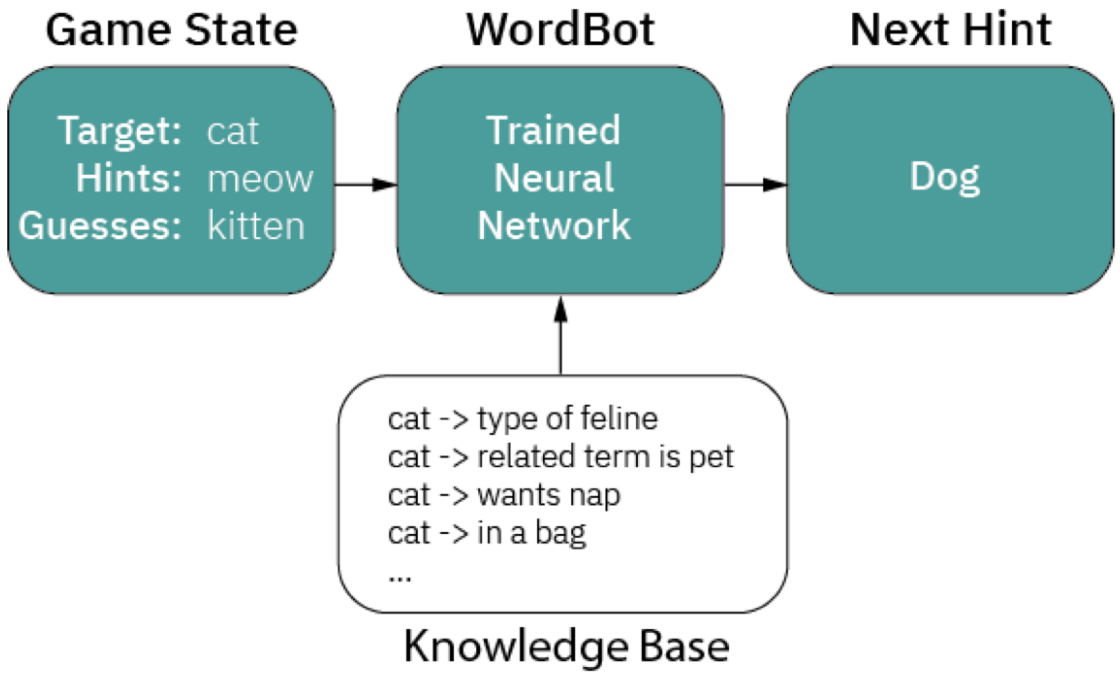
AI systems that help us understand ourselves

Human Centered Artificial Intelligence



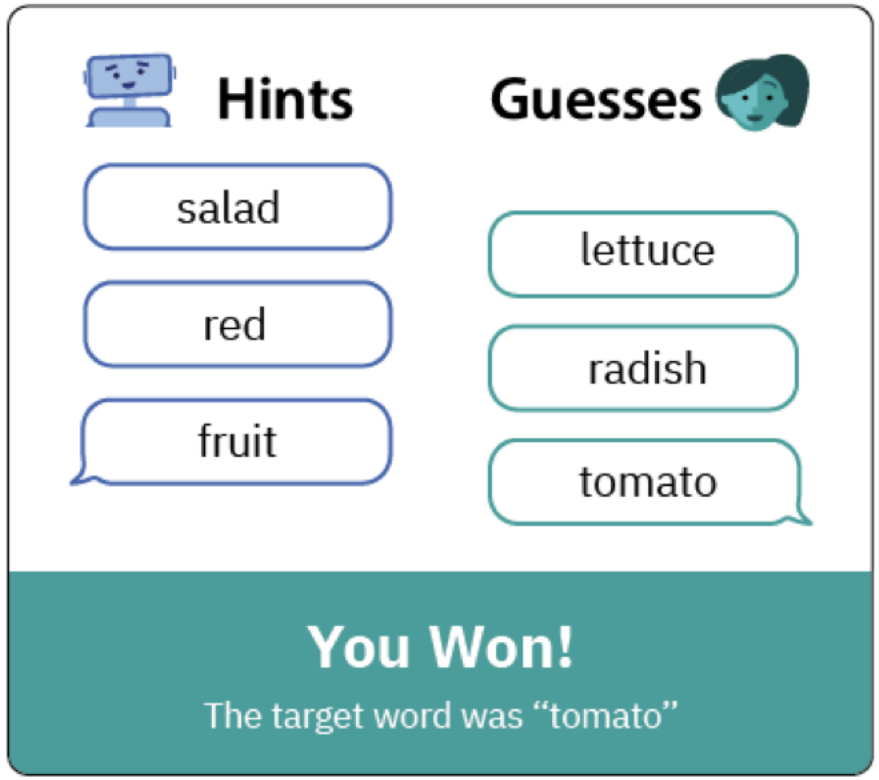
Human Centered Artificial Intelligence





Reference: Mental Models of AI Agents in a Cooperative Game Setting

Human Centered Artificial Intelligence



pattern seeking

synonyms/antonyms

steering

need for explanation

perspective taking

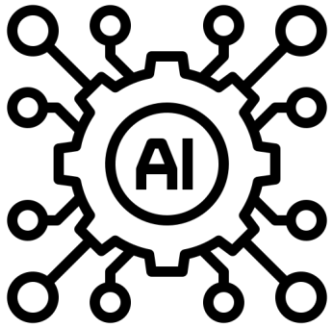
Mental Models ~120 participants

Reference: Mental Models of AI Agents in a Cooperative Game Setting

Human Centered Artificial Intelligence

Theory of Mind

Transparency

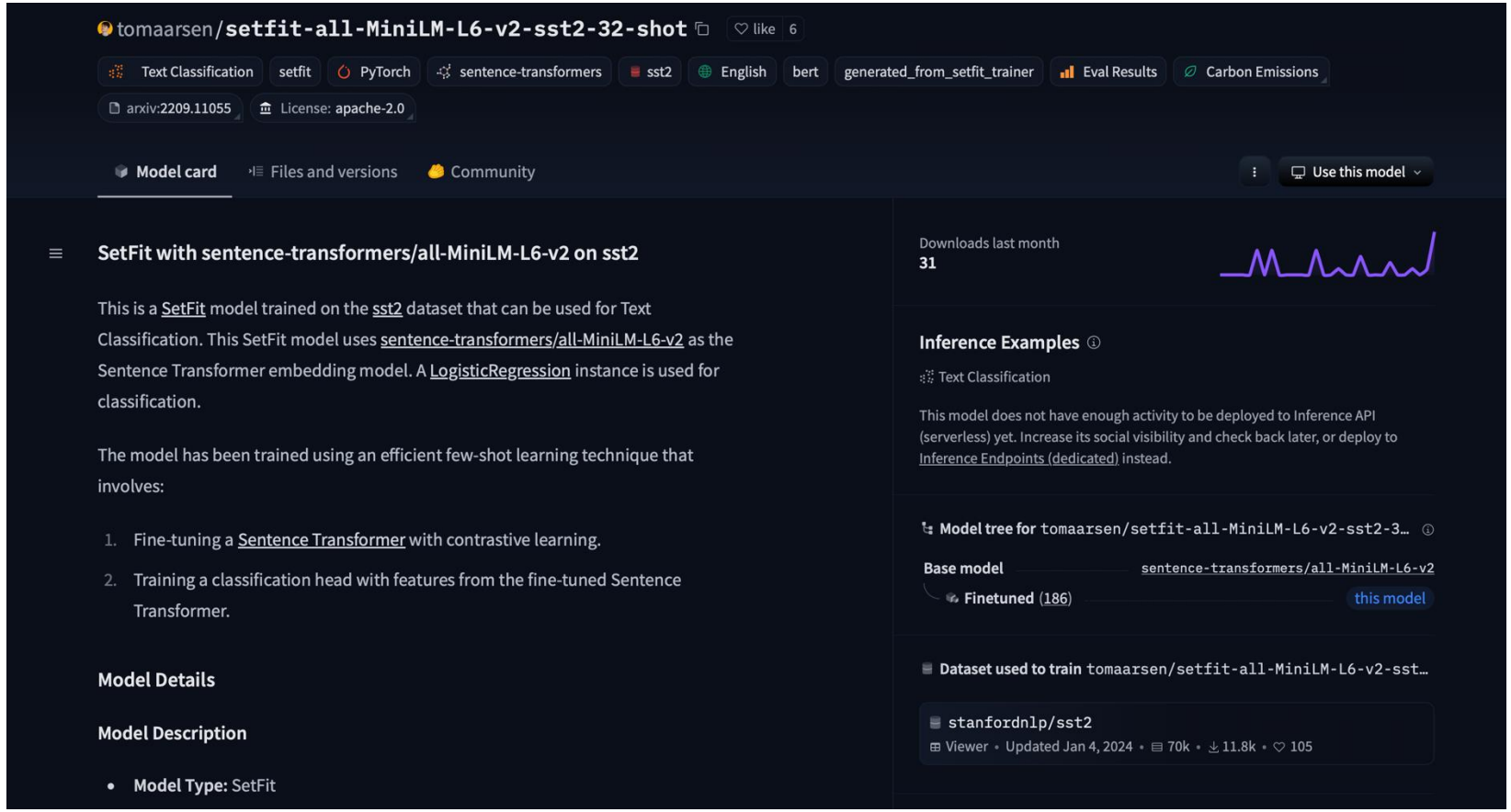


AI systems that understand humans

AI systems that help us understand ourselves

Mental Models

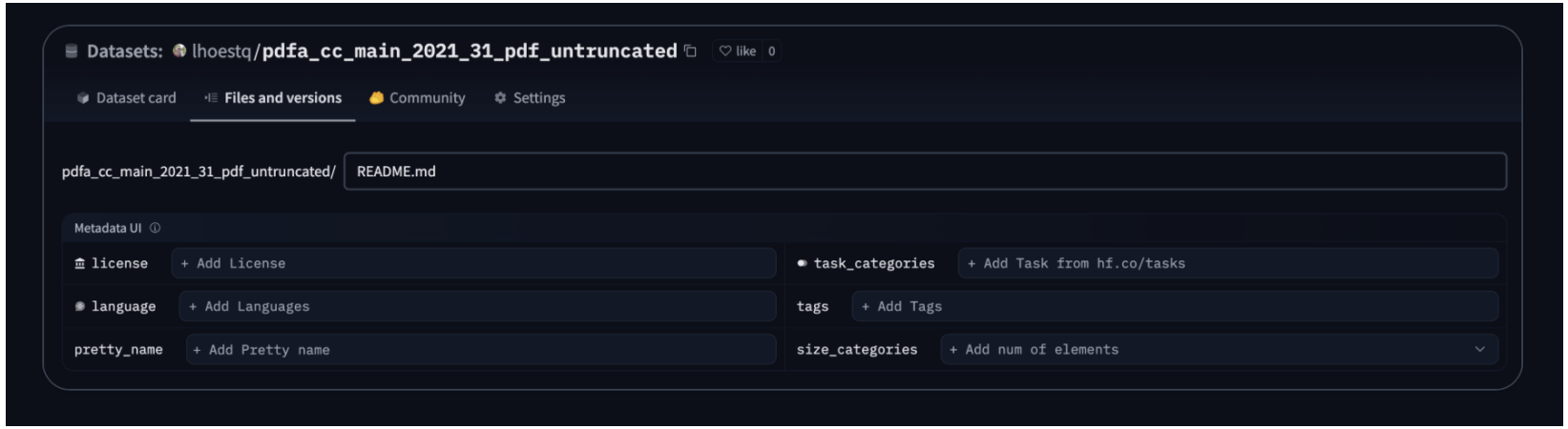
Human Centered Artificial Intelligence



Hugging face: Model Cards

McMaster University

Human Centered Artificial Intelligence

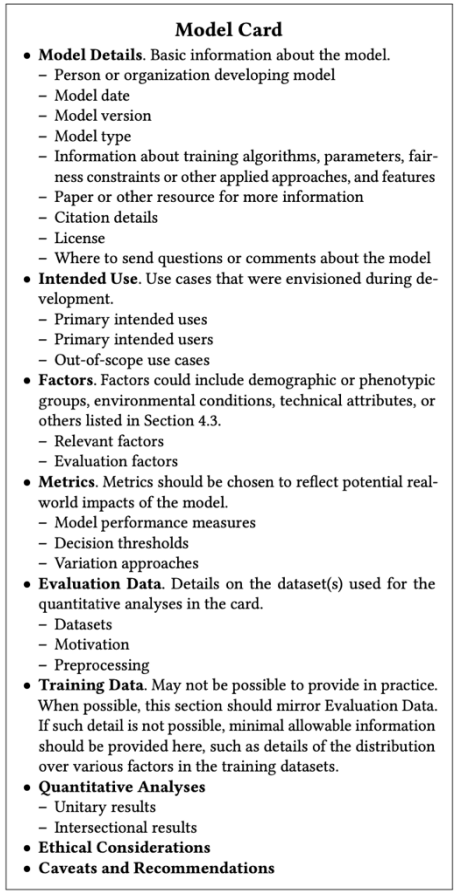


Hugging face: Datasheet Cards

Human Centered Artificial Intelligence

- **Why was the dataset created?** (e.g., was there a specific intended task gap that needed to be filled?)
- **Who funded the creation of the dataset?**
- **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances)
- **If it relates to people, were they told what the dataset would be used for and did they consent?** If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?
- **Will the dataset be updated?** How often, by whom?

Reference: Datasheet for Datasets



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

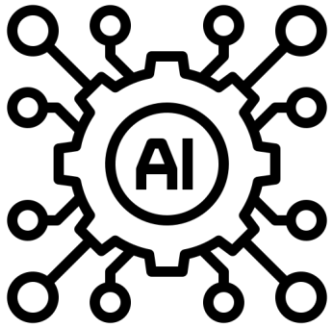
Reference: Model Cards

Human Centered Artificial Intelligence

Theory of Mind

Explainability

Transparency



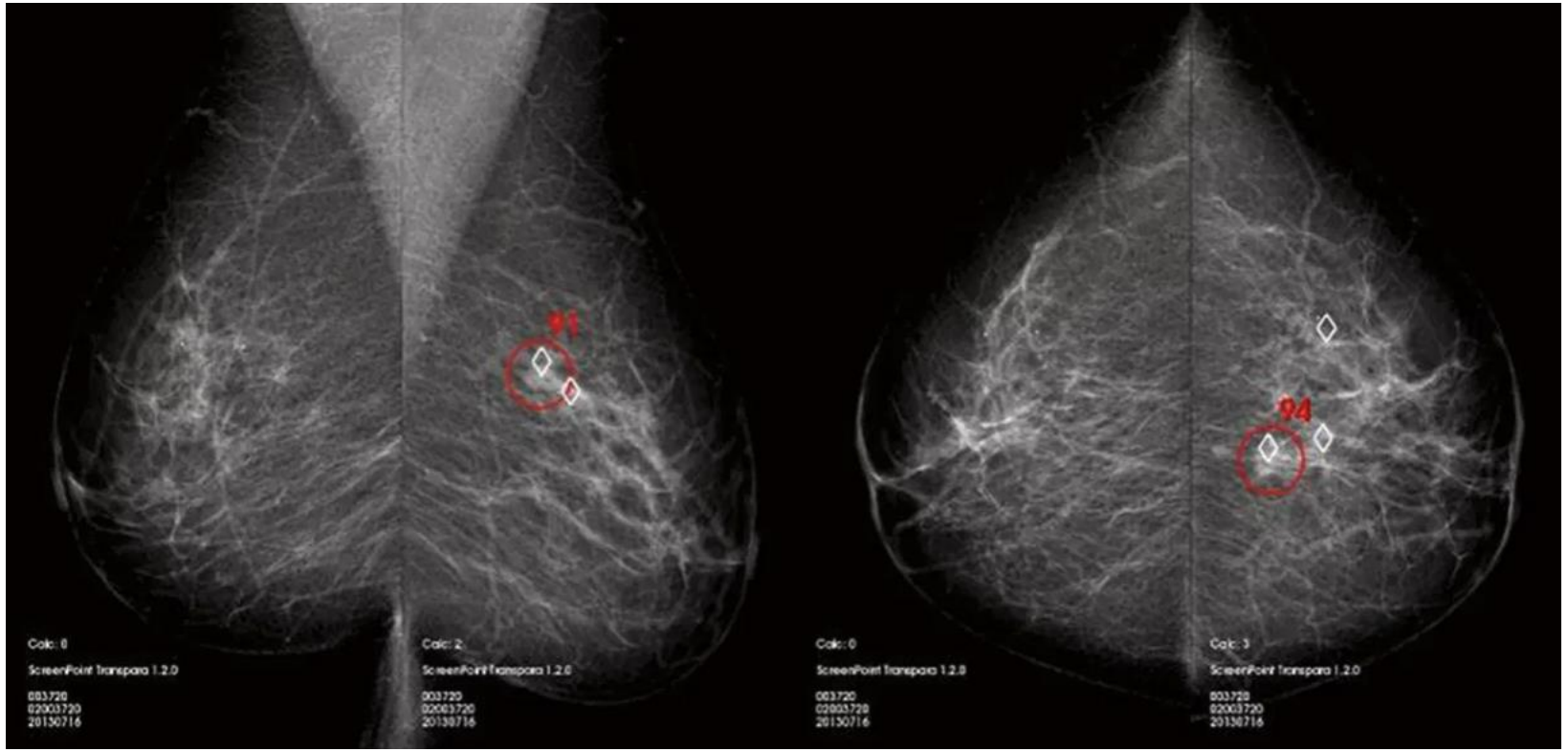
AI systems that understand humans

AI systems that help us understand ourselves

Mental Models

Human Centered Artificial Intelligence

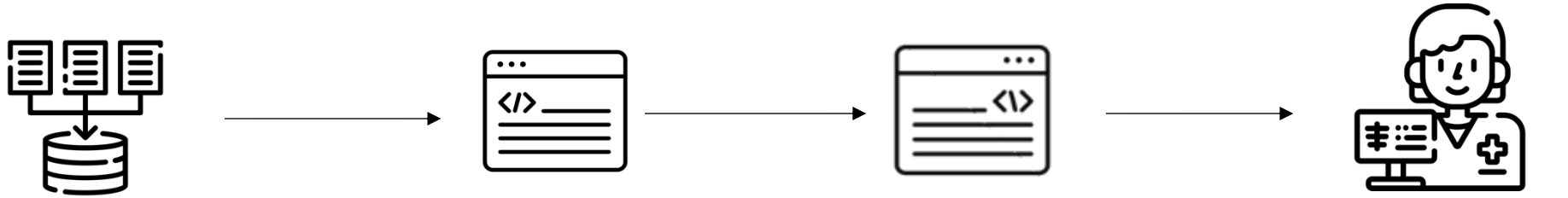
Computer Aided Diagnostics (CAD): Detecting and predicting abnormalities in breast tissue



McMaster University

Human Centered Artificial Intelligence

Computer Aided Diagnostics (CAD): Detecting and predicting abnormalities in breast tissue



Data Input

CAD program **identifies** abnormalities.

CAD program **classifies** abnormalities.

Radiologist

Goal: Discriminate between benign and malignant lesions with high accuracy.

Human Centered Artificial Intelligence

Computer Aided Diagnostics (CAD): Detecting and predicting abnormalities in breast tissue

Comparison of Accurate Classification of Microcalcifications between the Computerized Method and Radiologists

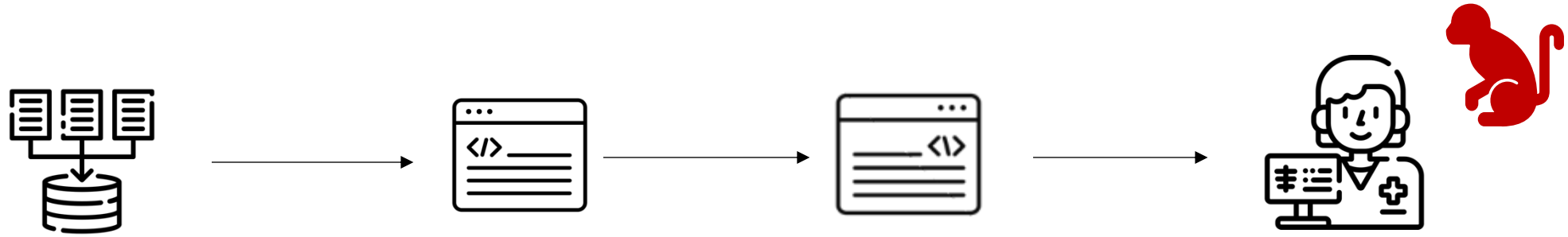
	Correct Classifications (%)		Positive Biopsy Yield (%)	Area Under ROC Curve	
	Malignant	Benign		Full Curve	TPF ≥0.90
Computerized method	100	82	76	0.92	0.082
Radiologists	100	27	45	0.89	0.042
P value	. . .	0.009	0.003	0.21	0.03

Note.—TPF = true-positive fraction. Biopsy in all patients at prospective clinical evaluation. Hypothetical positive biopsy yield = 36%.

Malignant and benign clustered microcalcifications: automated feature analysis and classification (1996)
Y Jiang, R M Nishikawa, D E Wolverton, C E Metz, M L Giger, R A Schmidt, C J Vyborny, K Doi

Human Centered Artificial Intelligence

Computer Aided Diagnostics (CAD): Detecting and predicting abnormalities in breast tissue



Goal: Discriminate between benign and malignant lesions with high accuracy.

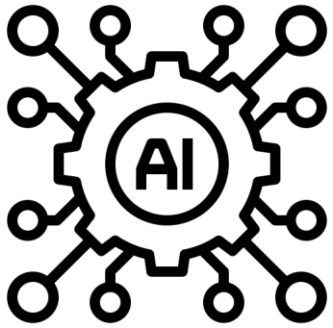
Human Centered Artificial Intelligence

Theory of Mind

Explainability

Transparency

Interpretability



AI systems that understand humans

AI systems that help us understand ourselves

Mental Models

A Note on Interpretability

Cervical Cancer (Risk Factors)

Donated on 3/2/2017

This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Health and Medicine	Classification
Feature Type	**# Instances**	**# Features**
Integer, Real	858	36

A Note on Interpretability

- Using logistic regression $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$

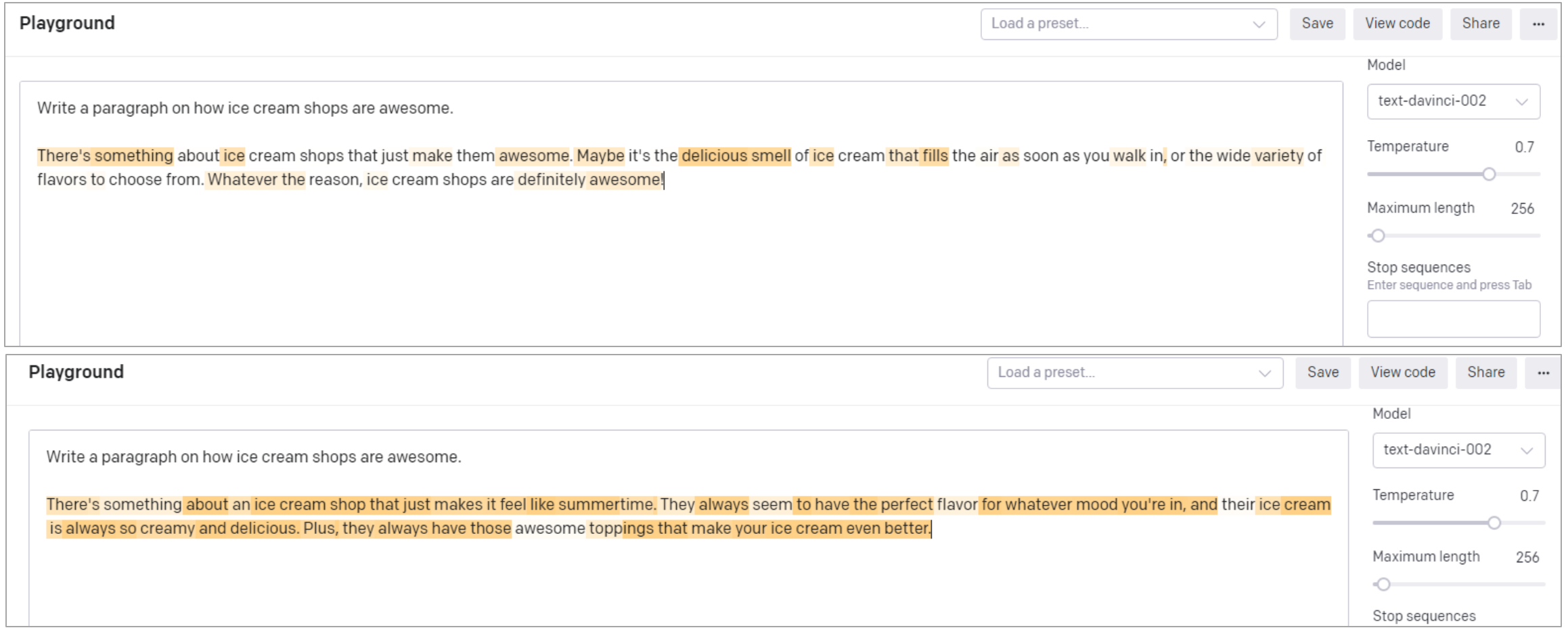
	Weight	Odds ratio	Std. Error
Intercept	-2.91	0.05	0.32
Hormonal contraceptives y/n	-0.12	0.89	0.30
Smokes y/n	0.26	1.30	0.37
Num. of pregnancies	0.04	1.04	0.10
Num. of diagnosed STDs	0.82	2.27	0.33
Intrauterine device y/n	0.62	1.86	0.40

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$P(cancer = Yes | \, patient\ smokes\ = Yes)$

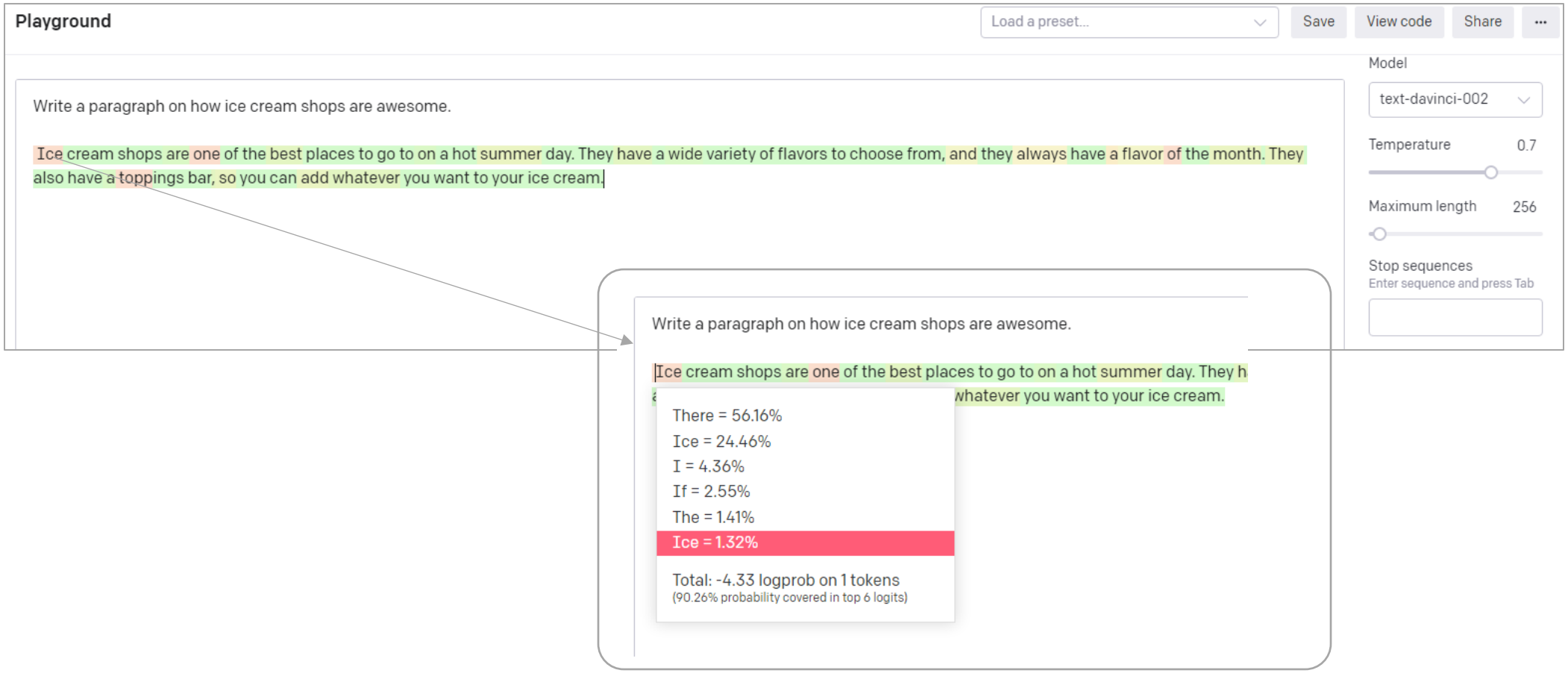
0	Age	Feature	Integer	Age
1	Number of sexual partners	Feature	Continuous	Other
2	First sexual intercourse	Feature	Continuous	None
3	Num of pregnancies	Feature	Continuous	None
4	Smokes	Feature	Continuous	None
5	Smokes (years)	Feature	Continuous	None
6	Smokes (packs/year)	Feature	Continuous	None
7	Hormonal Contraceptives	Feature	Continuous	None
8	Hormonal Contraceptives (years)	Feature	Continuous	None
9	IUD	Feature	Continuous	None
10	IUD (years)	Feature	Continuous	None
11	STDs	Feature	Continuous	None
12	STDs (number)	Feature	Continuous	None
13	STDs:condylomatosis	Feature	Continuous	None
14	STDs:cervical condylomatosis	Feature	Continuous	None
15	STDs:vaginal condylomatosis	Feature	Continuous	None
16	STDs:vulvo-perineal condylomatosis	Feature	Continuous	None
17	STDs:syphilis	Feature	Continuous	None
18	STDs:pelvic inflammatory disease	Feature	Continuous	None
19	STDs:genital herpes	Feature	Continuous	None
20	STDs:molluscum contagiosum	Feature	Continuous	None
21	STDs:AIDS	Feature	Continuous	None
22	STDs:HIV	Feature	Continuous	None

A Note on Interpretability



GPT Ancestor

A Note on Interpretability



A Note on Interpretability

Thinking Exercise!

A Note on Interpretability

Thinking Exercise!

What does VGG-16 think this is?

1) Broom
2) Honeycomb
3) Plunger
4) Tennis Ball
5) Panpipe

A Note on Interpretability

Thinking Exercise!

What does VGG-16 think this is?

1) Broom
2) Honeycomb
3) Plunger
4) Tennis Ball
5) Panpipe

Explanation of why this is a broom

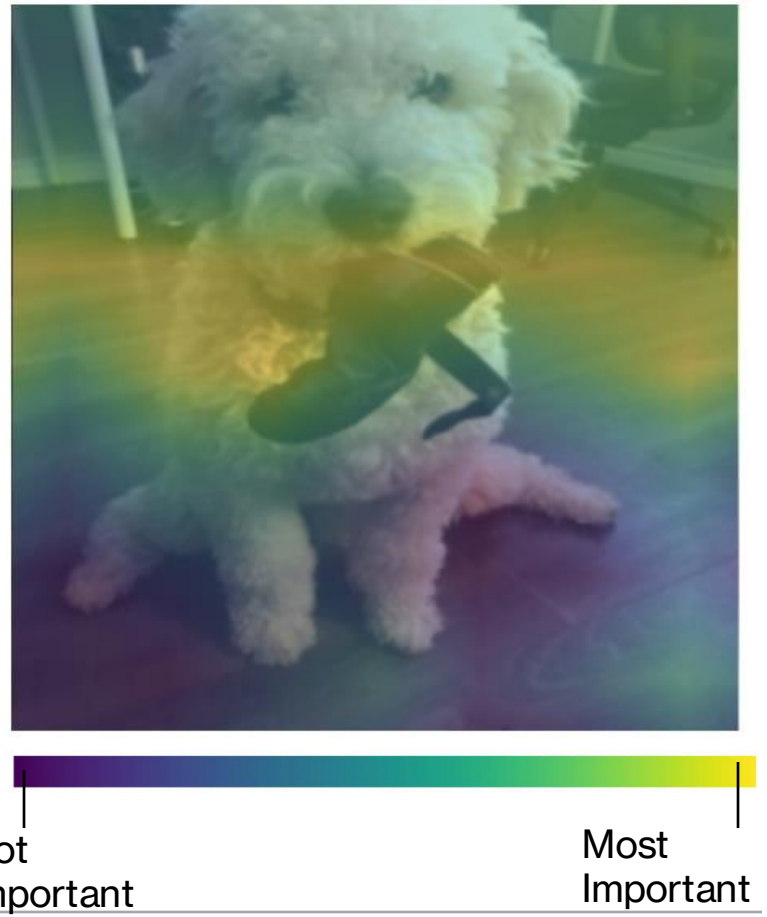
A Note on Interpretability

Thinking Exercise!

What does VGG-16 think this is?

1) Broom
2) Honeycomb
3) Plunger
4) Tennis Ball
5) Panpipe

Explanation of why this is a broom



A Note on Interpretability

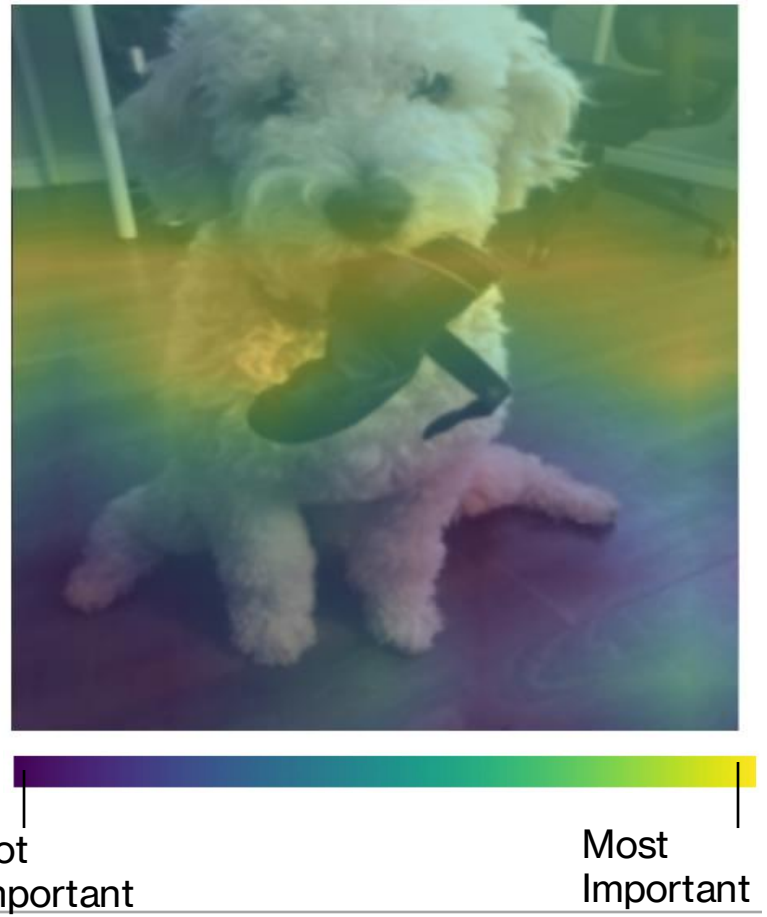
Thinking Exercise!

What does VGG-16 think this is?

1) Broom
2) Honeycomb
3) Plunger
4) Tennis Ball
5) Panpipe

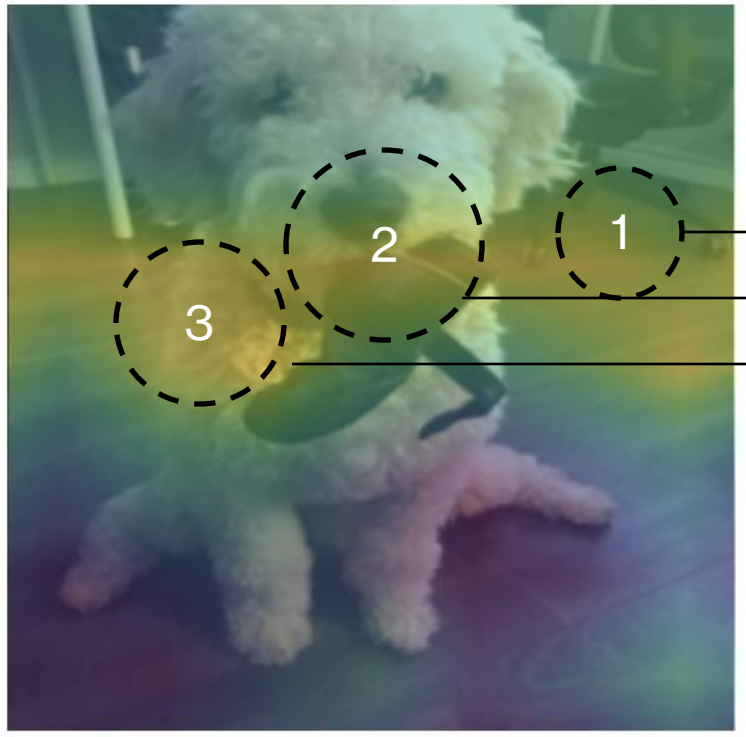
How confident are you that VGG-16 is good at identifying dogs and why?

Explanation of why this is a broom



A Note on Interpretability

How confident are you that VGG-16 is good at identifying dogs and why?



Wooden Floor

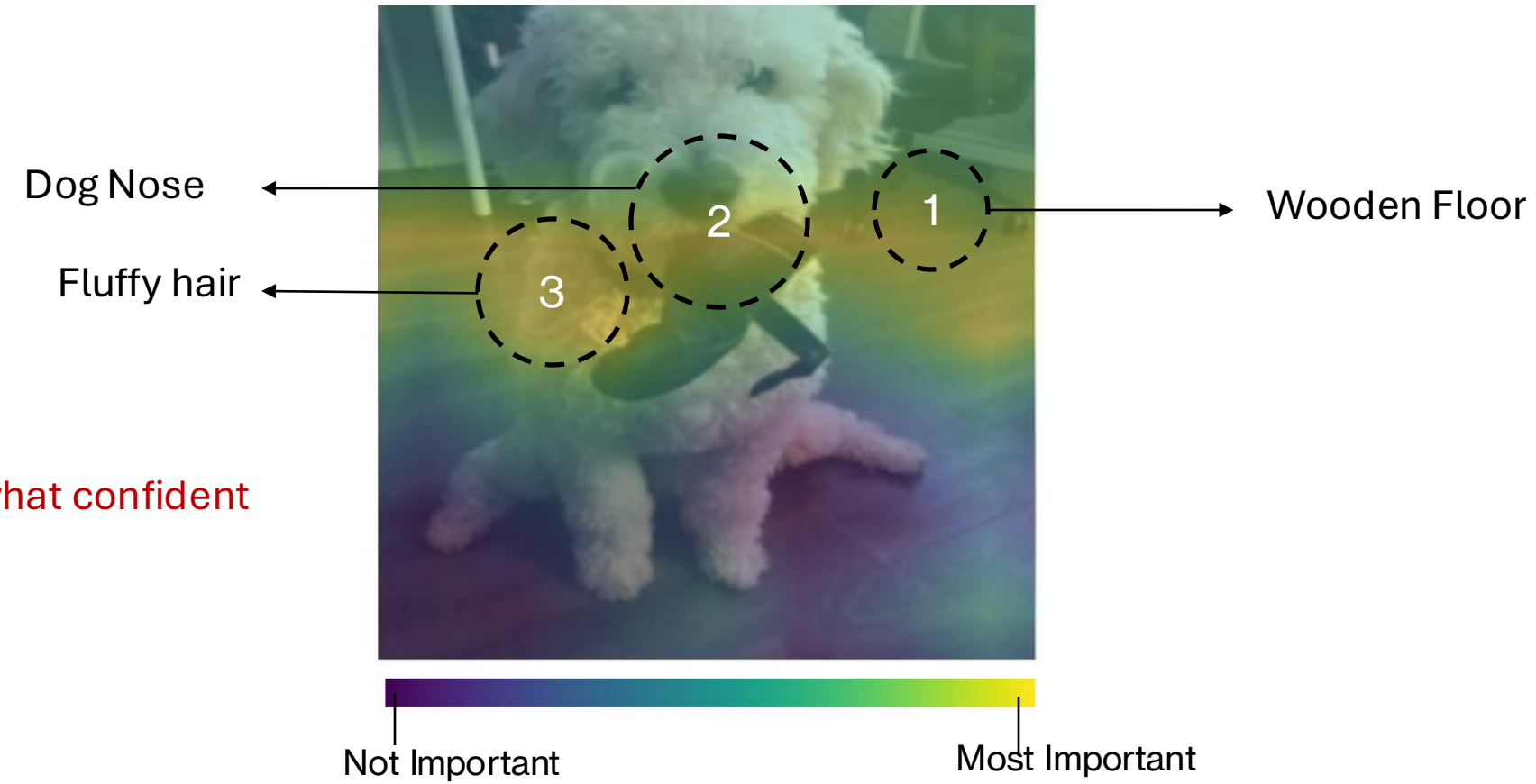
Shoe

Soft “bristles”

Not confident at all

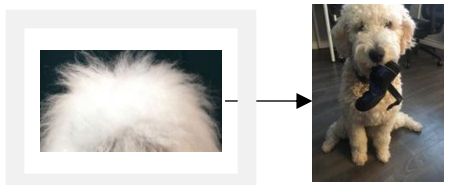
A Note on Interpretability

How confident are you that VGG-16 is good at identifying dogs and why?



Somewhat confident

A Note on Interpretability



Fluffy hair

Dog

Higher - level data features

Data Concepts
(e.g., textures in image, sentiment of text)

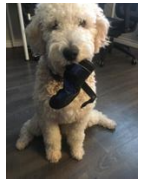
Data Regions
(e.g., segments in image, sentences in text)

Data Representations
(e.g., pixels in image, characters in text)

Human Reasoning is Top-Down

Machine Reasoning is Bottom-up

Pixel Weights



Dog

Lower - level data features

Big Book of Concepts, Gregory L. Murphy, 2004, MIT Press

A Note on Interpretability

A concept is a mental representation of classes or categories of things which help individuals reason about it.

Yellow Stripes

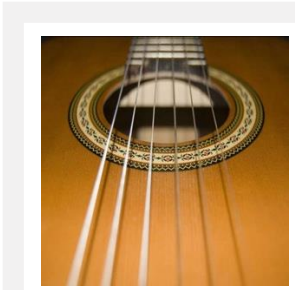
Tiger

Dots

Ladybug

Black Stripes

Zebra



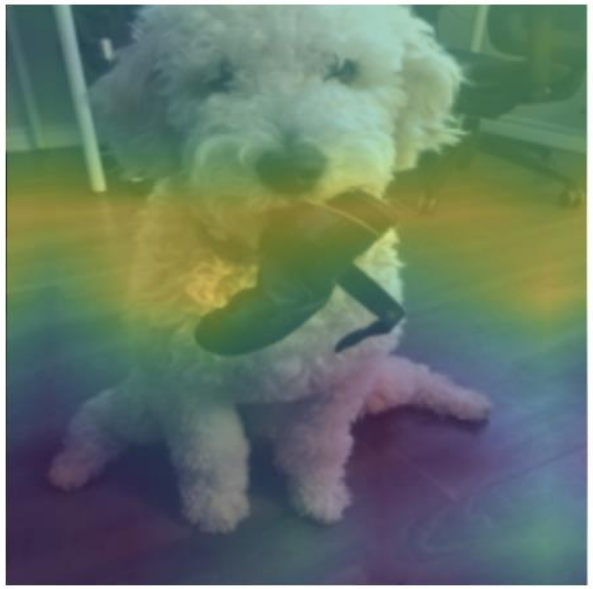
Strings

Guitar

A Note on Interpretability

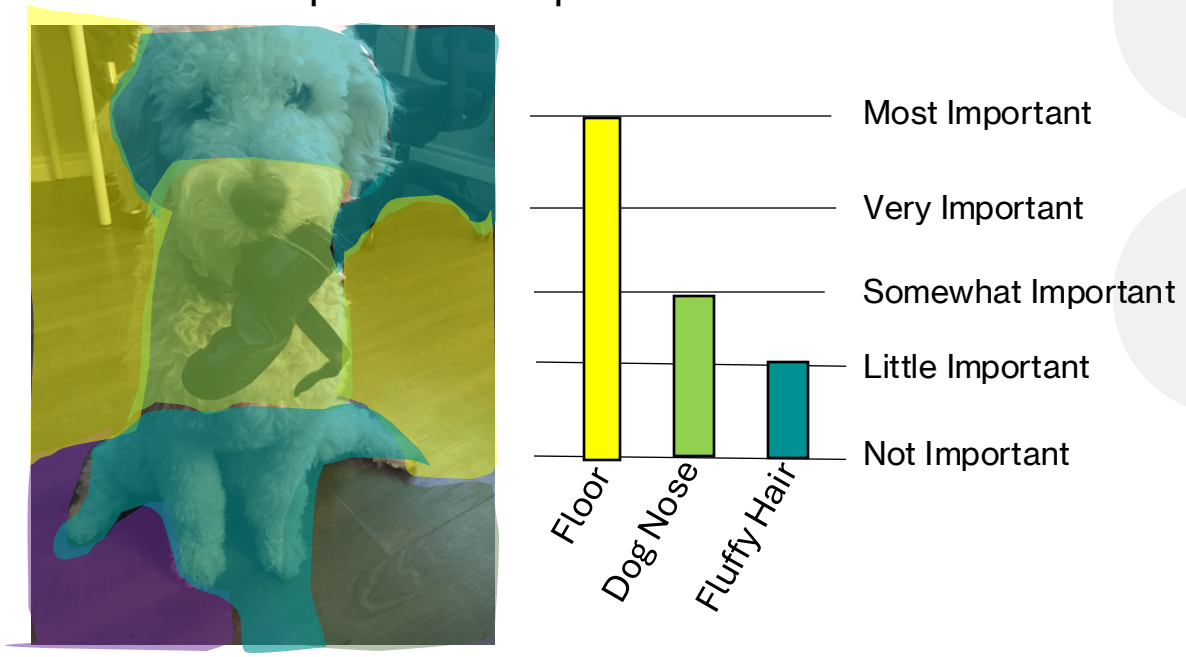
A concept is a mental representation of classes or categories of things which help individuals reason about it.

Machine Feature-based Explanation



Translate to

Human Concept-based Explanation



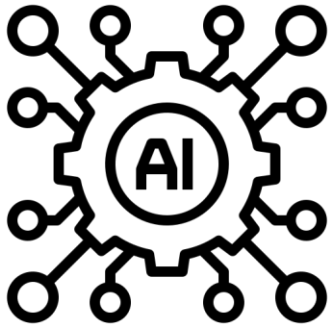
Human Centered Artificial Intelligence

Theory of Mind

Explainability

Transparency

Interpretability

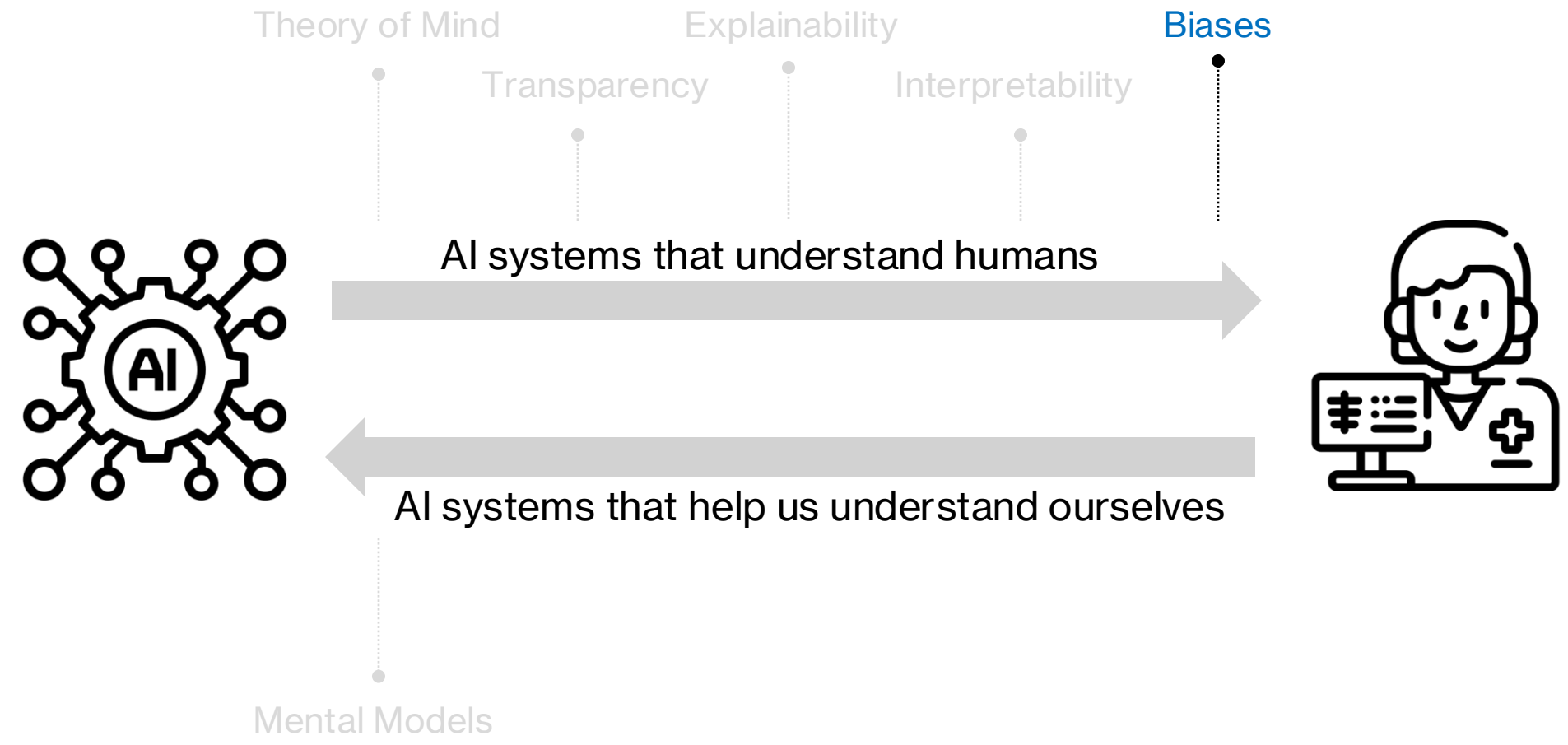


AI systems that understand humans

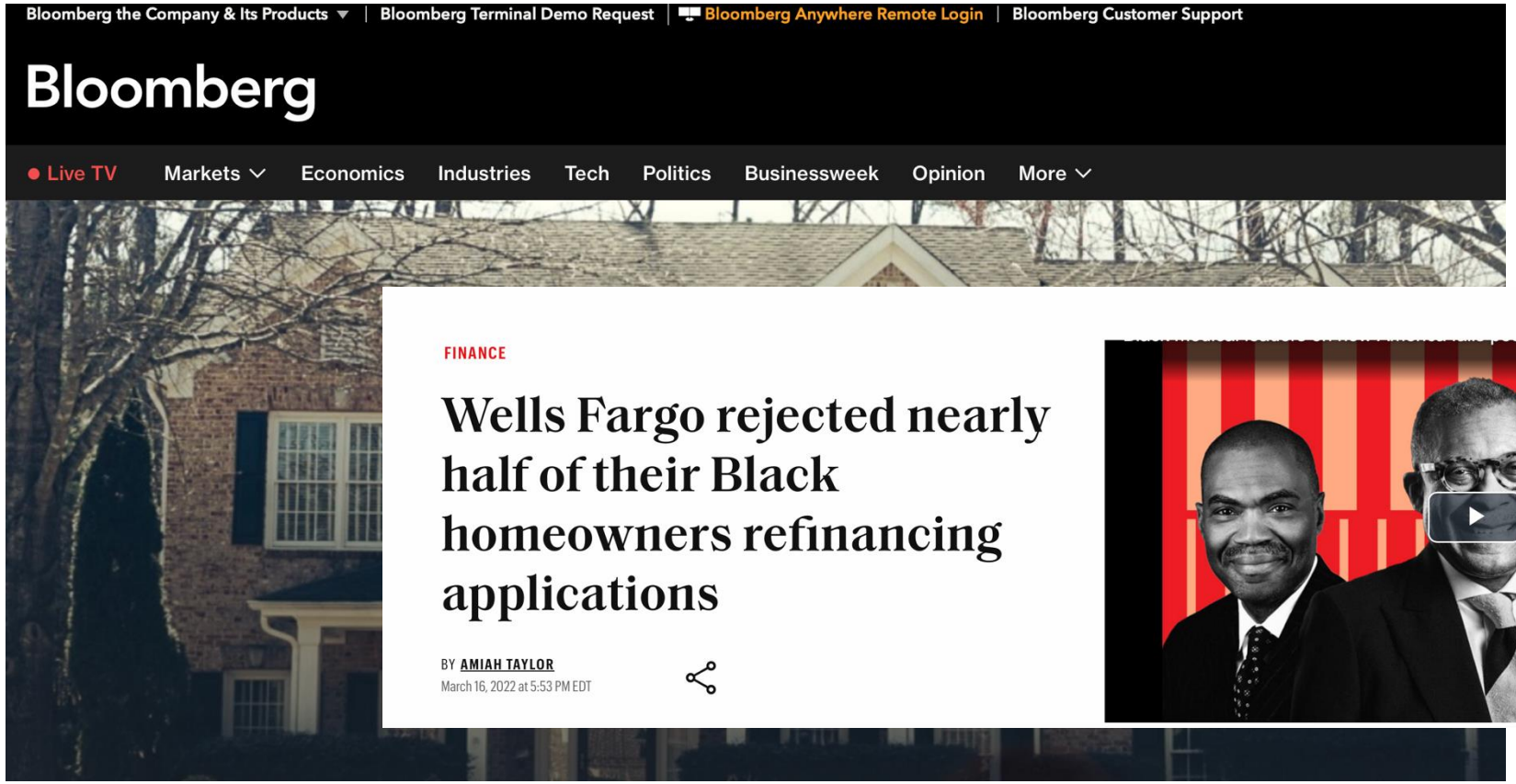
AI systems that help us understand ourselves

Mental Models

Human Centered Artificial Intelligence



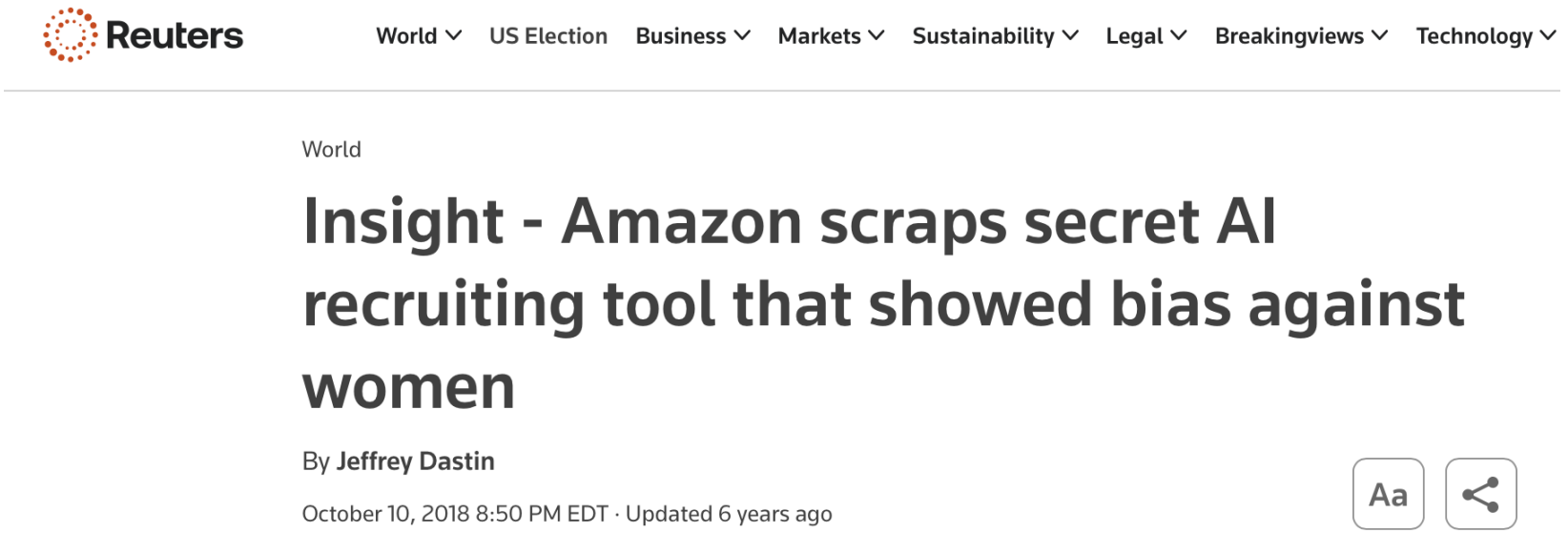
Human Centered Artificial Intelligence



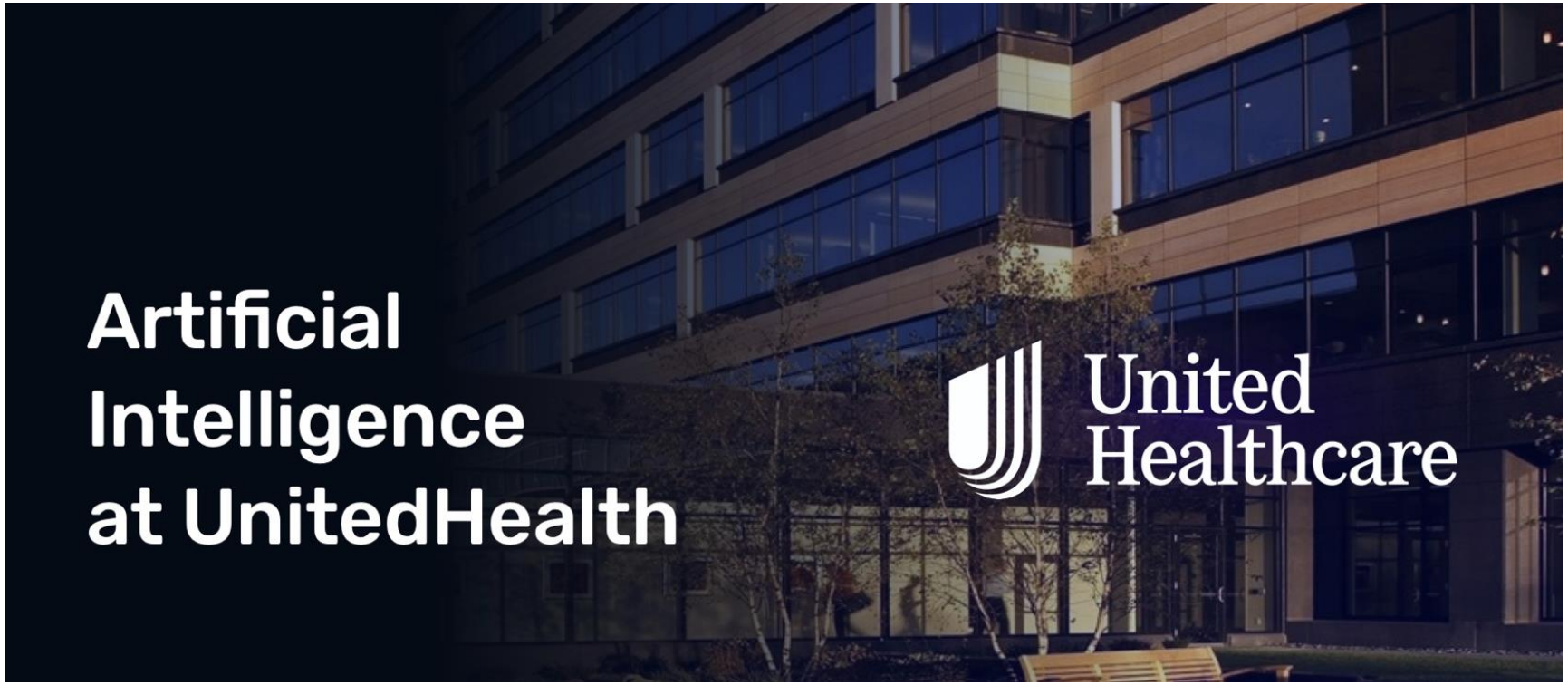
Flitter 2022, Donnan et al. 2022

McMaster University

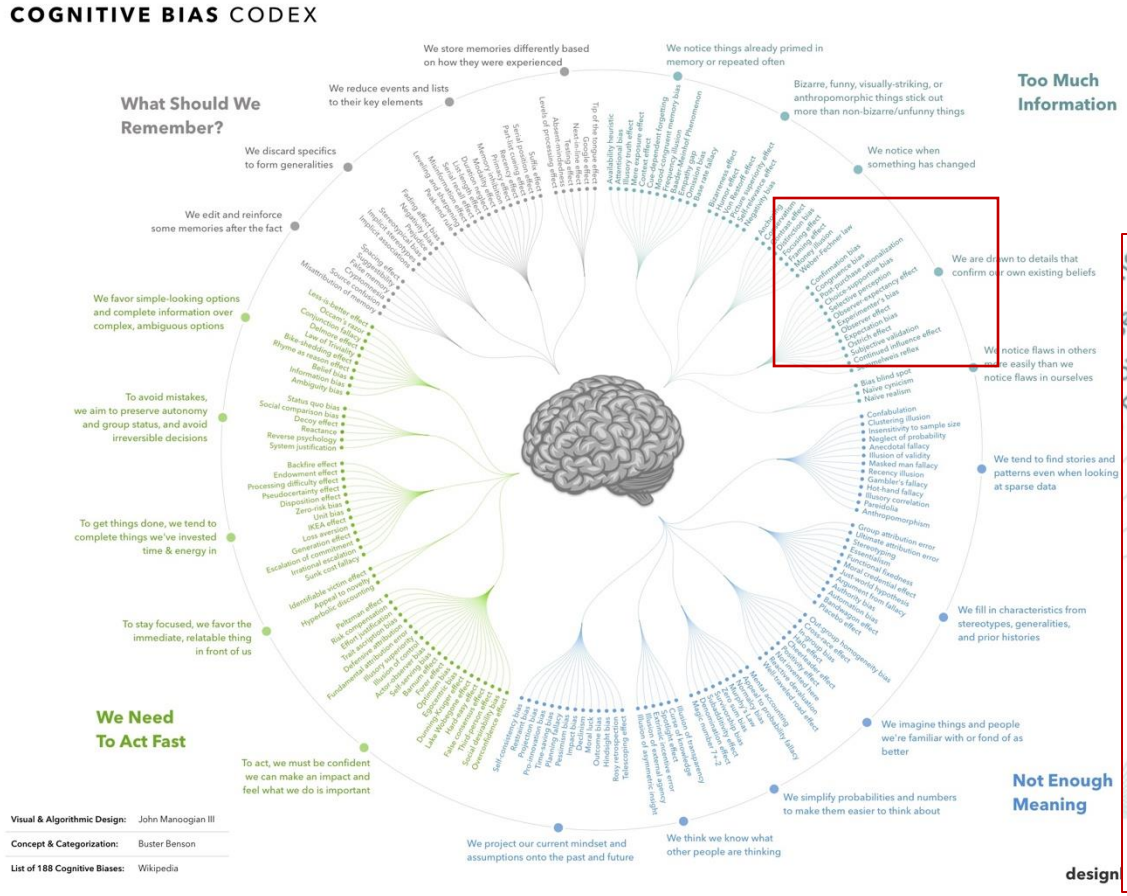
Human Centered Artificial Intelligence



Human Centered Artificial Intelligence

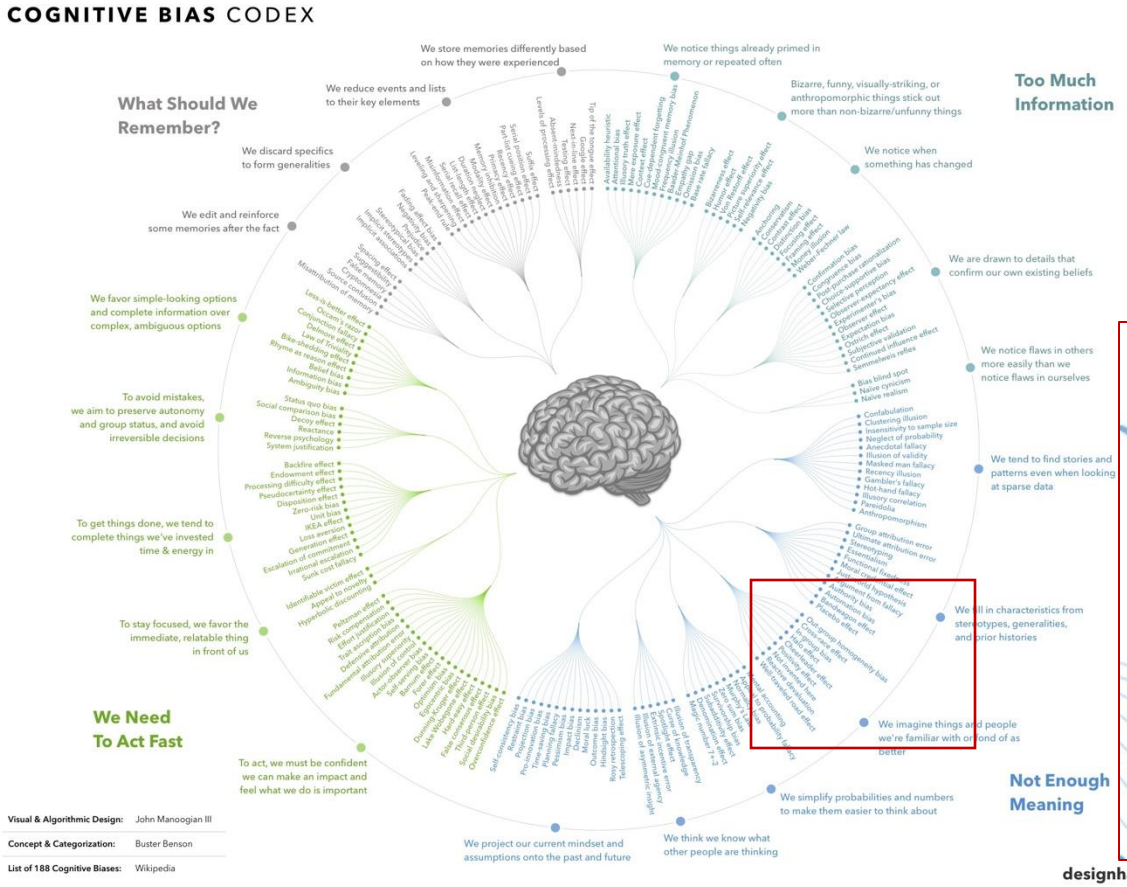


Types of Biases

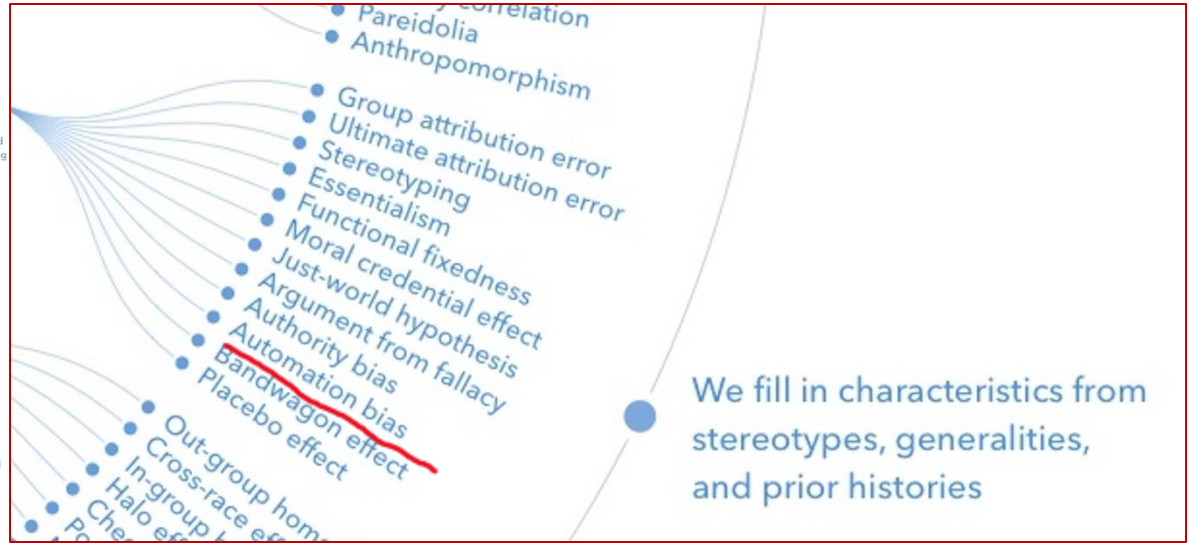


Confirmation Bias: Confirmation bias is the tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypothesis while giving disproportionately less attention to information that contradicts it.

Human Centered Artificial Intelligence

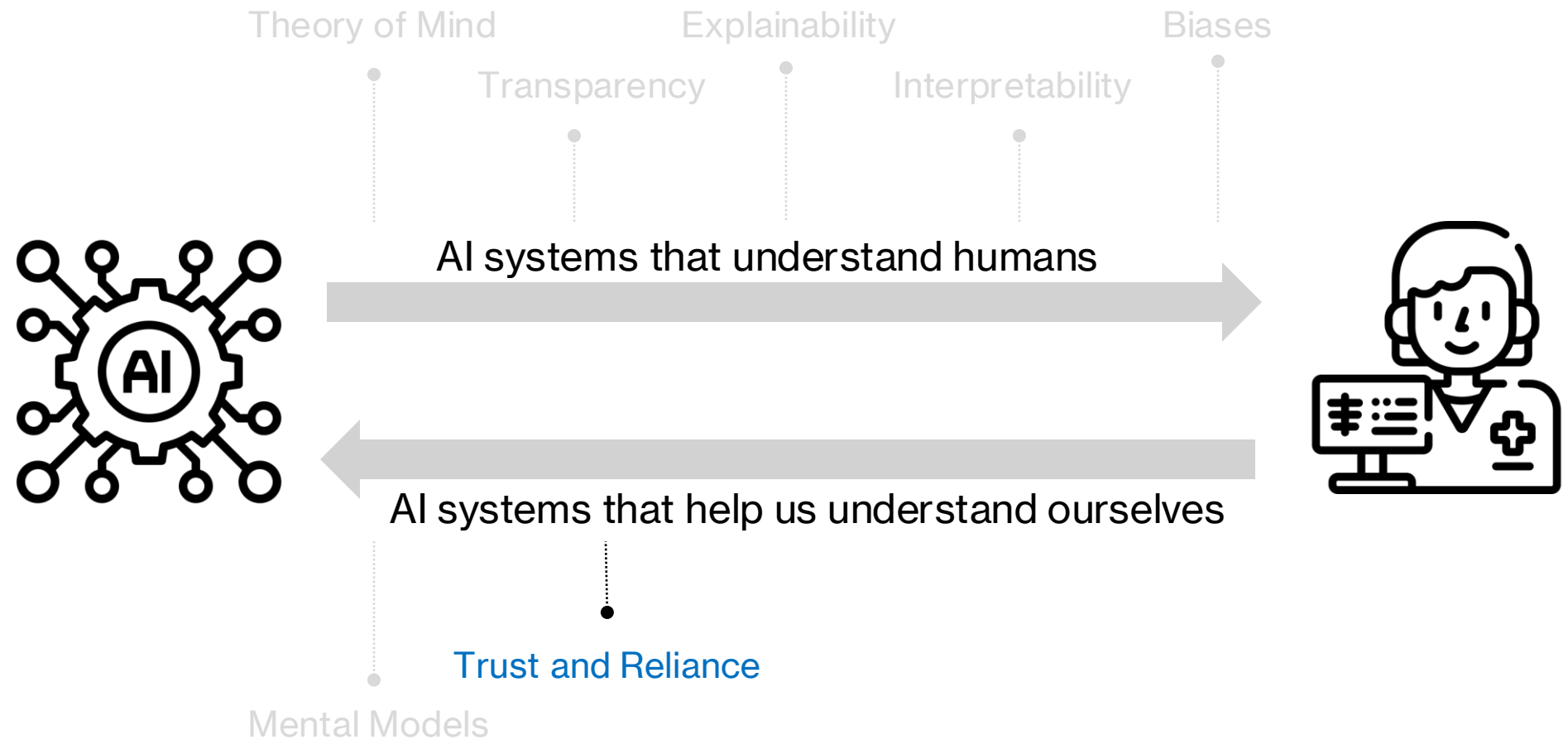


Automation bias: It is the human tendency to favor suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct.



designhacks.co

Human Centered Artificial Intelligence



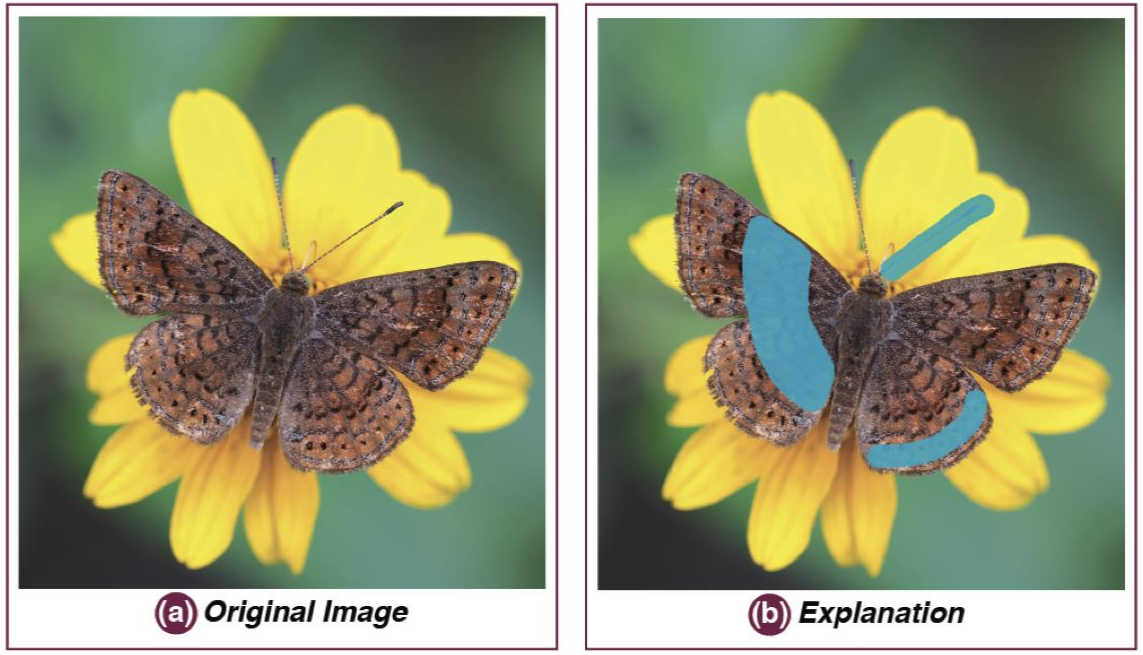
Human Centered Artificial Intelligence

(a) Desert Tarantula

(b) Ecuadorian Lubber Grasshopper

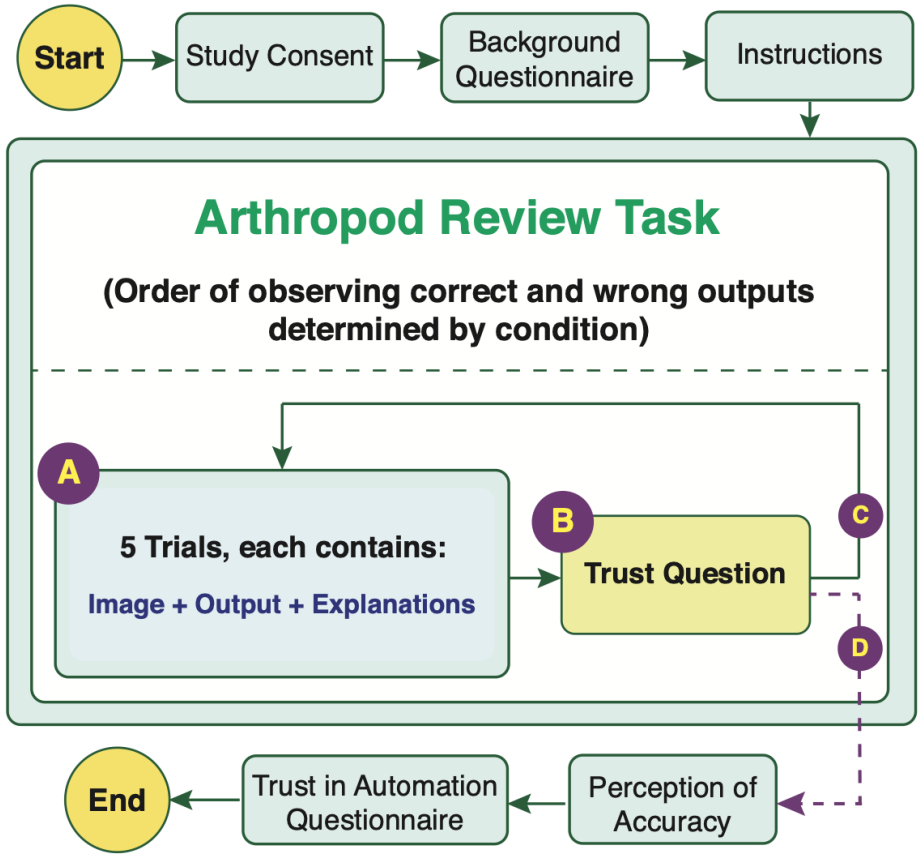
(c) Rhopalid (scentless plant bug)

(d) Metalmark Butterfly (*Calephelis* sp.)



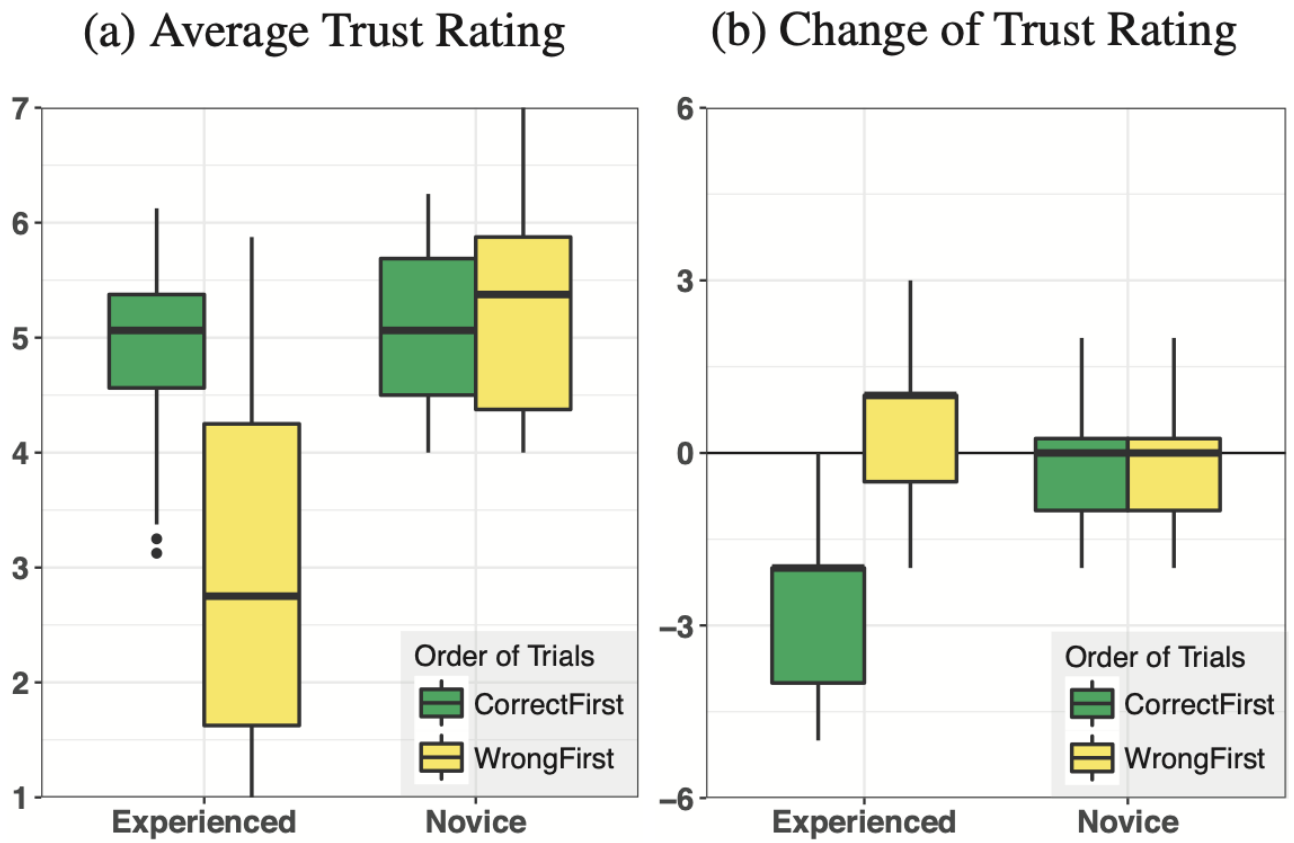
Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

Human Centered Artificial Intelligence



Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

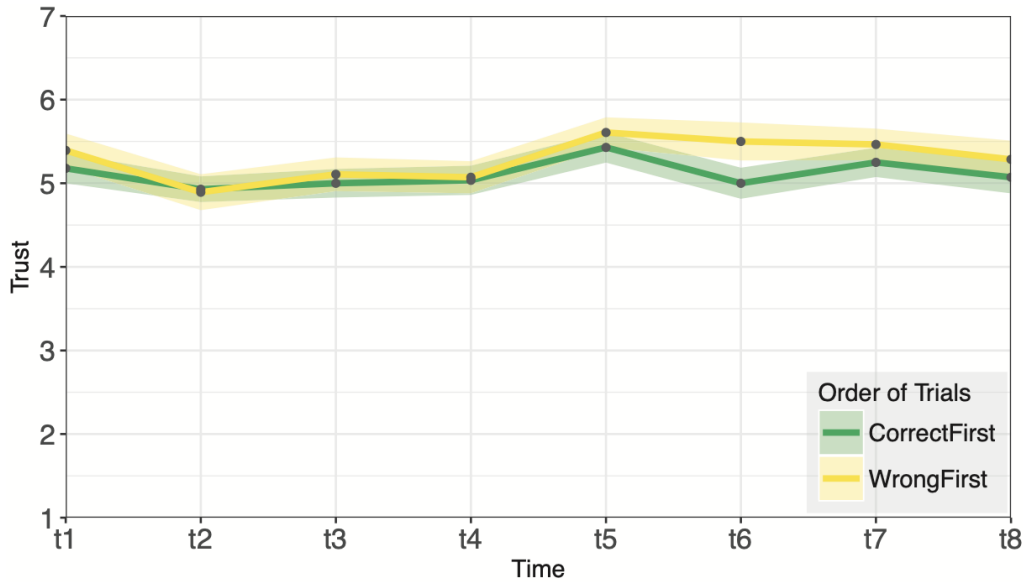
Human Centered Artificial Intelligence



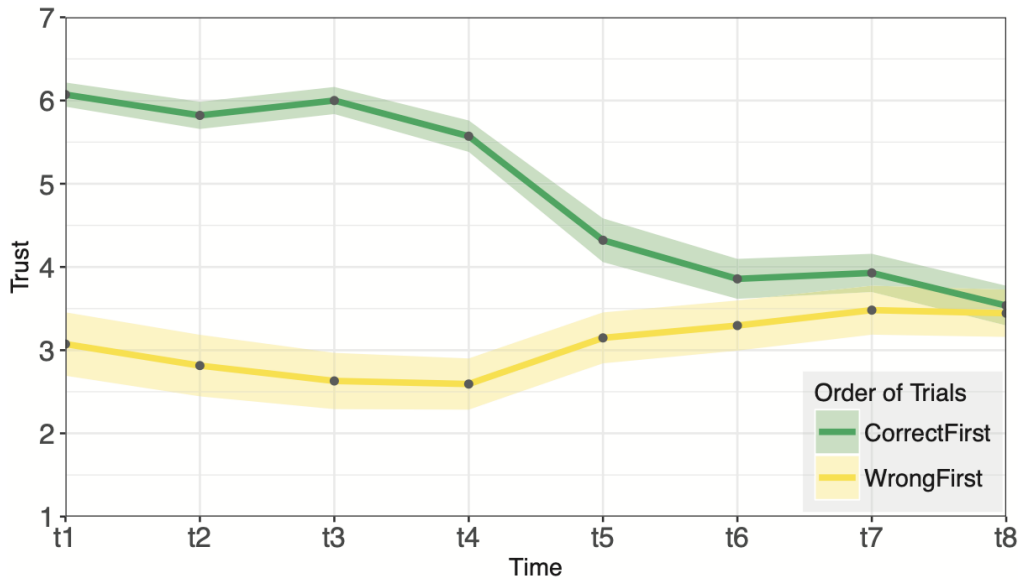
Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

Human Centered Artificial Intelligence

(a) Changes of trust over time for *novice* participants.



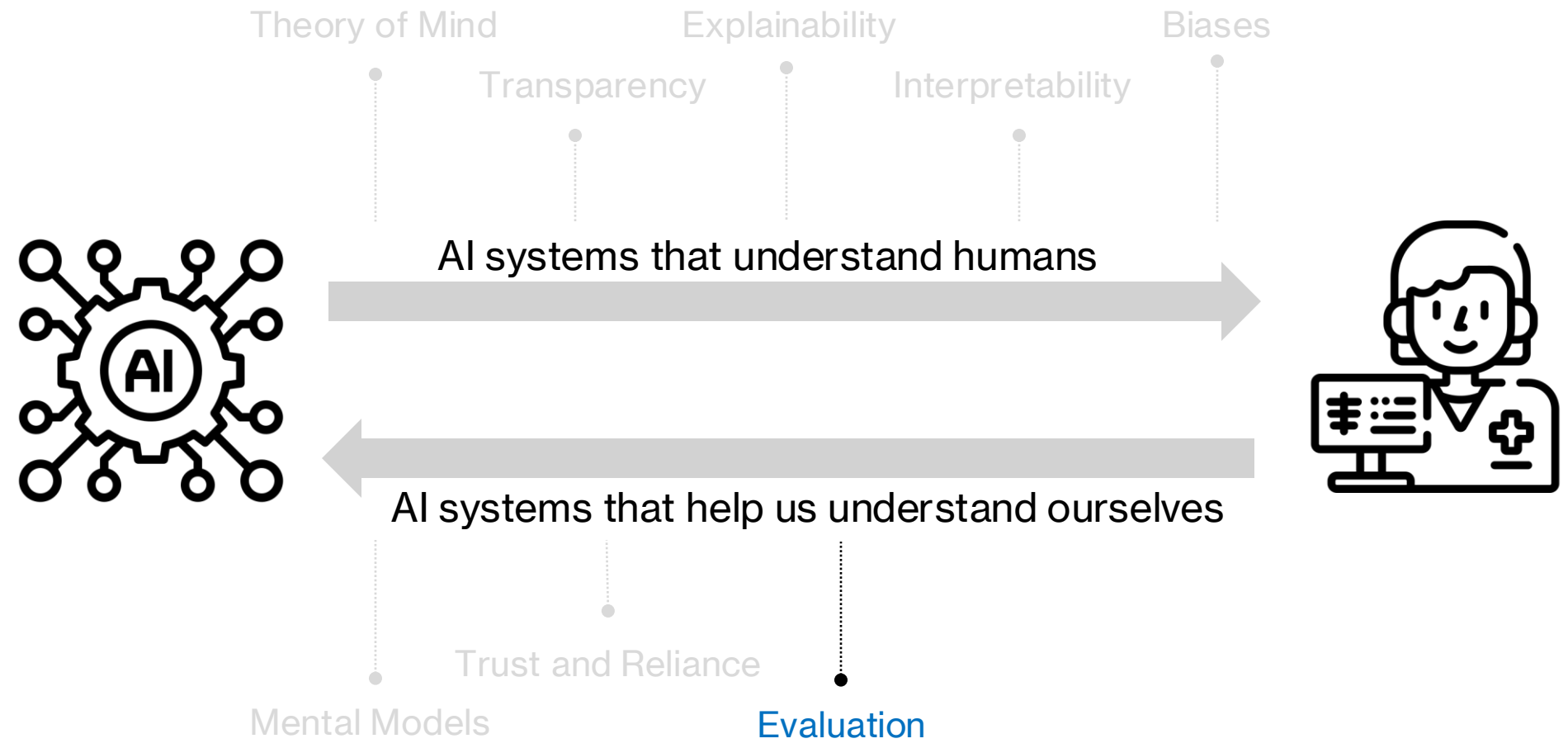
(b) Changes of trust over time for *experienced* participants.



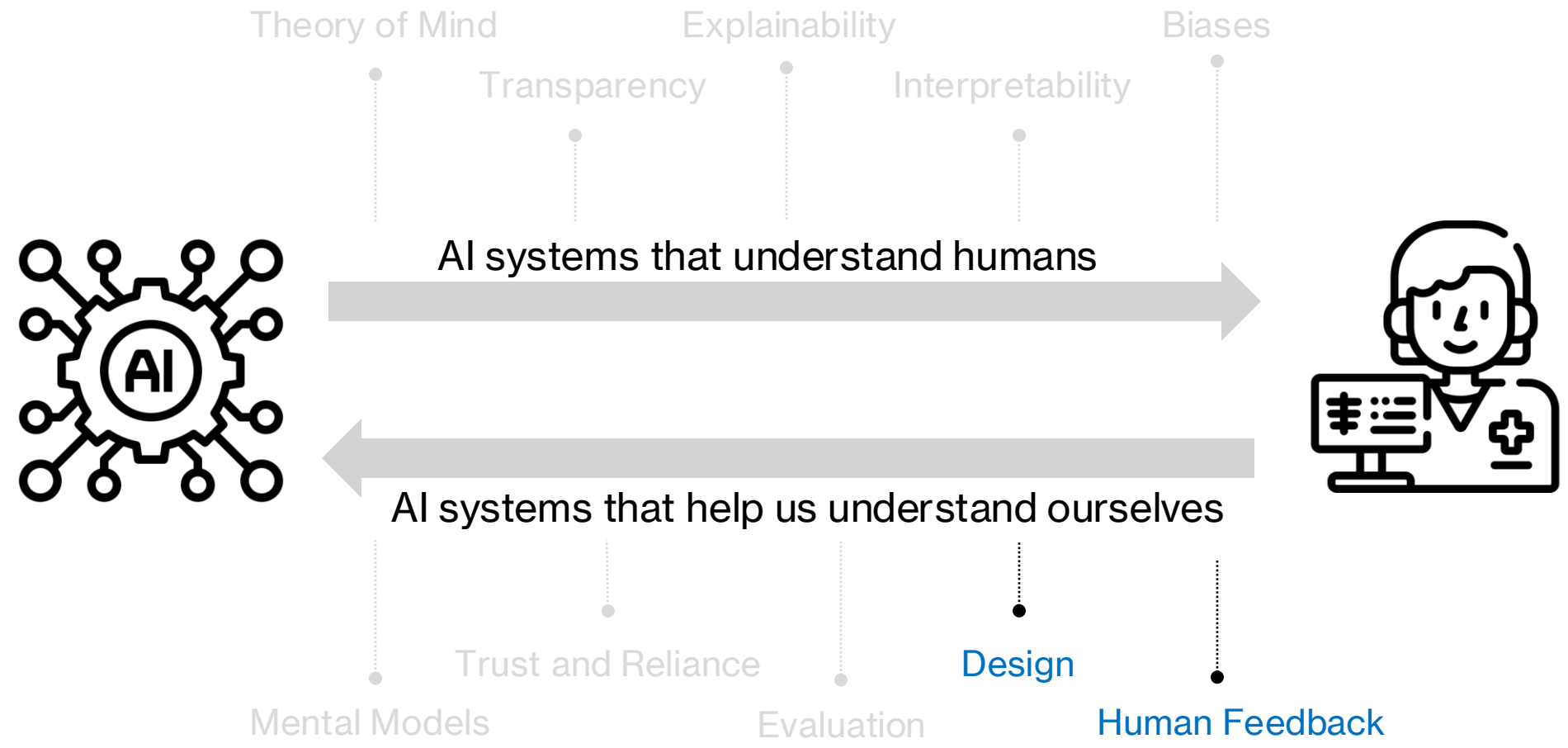
Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

McMaster University

Human Centered Artificial Intelligence

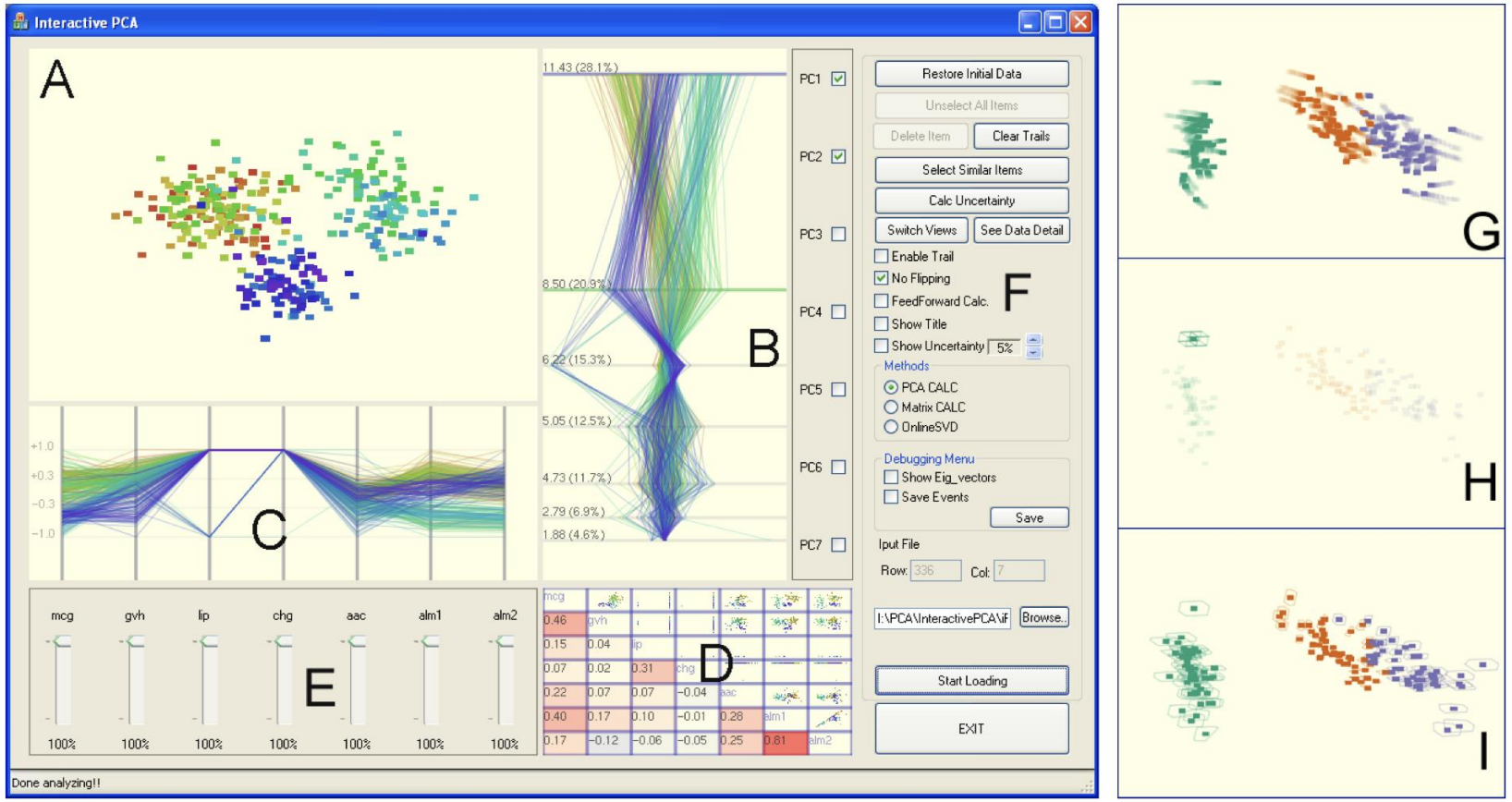


Human Centered Artificial Intelligence

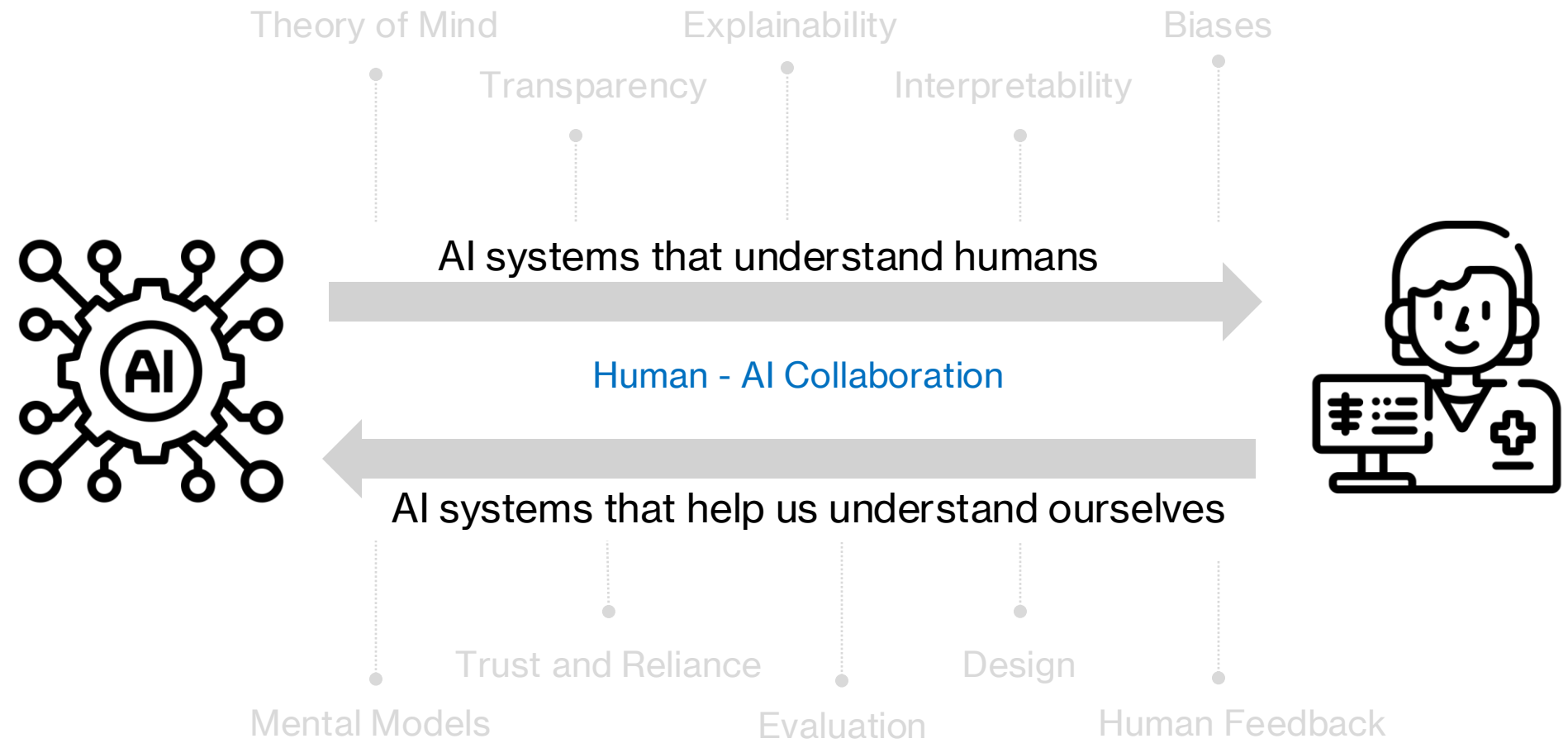


Human Centered Artificial Intelligence

D.H. Jeong et. al / iPCA: An Interactive System for PCA-based Visual Analytics



Human Centered Artificial Intelligence



Human Centered Artificial Intelligence

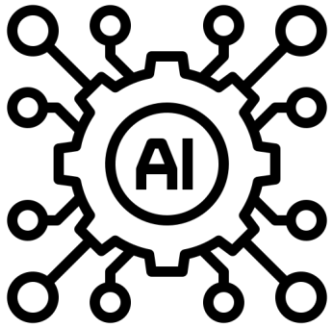
Module 1
Theory of Mind

Module 2
Transparency

Module 3
Explainability

Module 4
Interpretability

Module 5
Biases



AI systems that understand humans

Human - AI Collaboration
Module 11 and 12

AI systems that help us understand ourselves

Module 6
Mental Models

Module 7
Trust and Reliance

Module 8
Evaluation

Module 9
Design

Module 10
Human Feedback

*not necessarily in that order

Goals

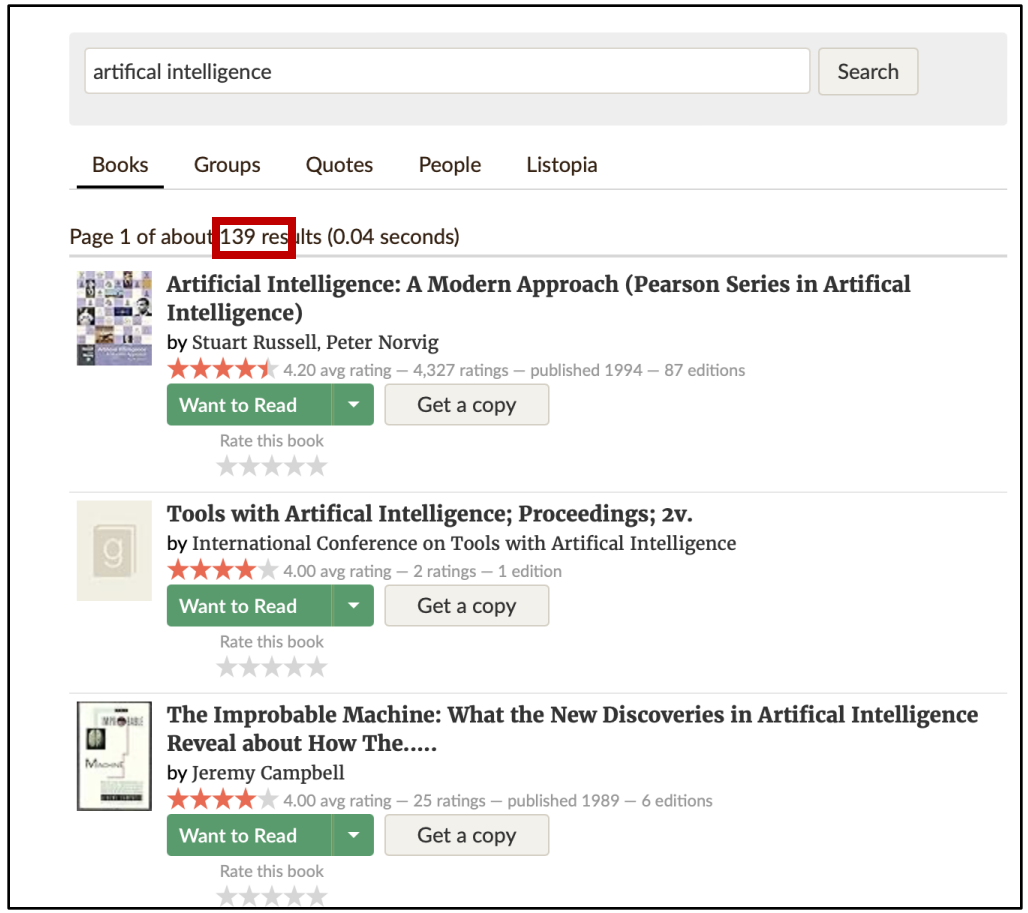
This course will focus on:

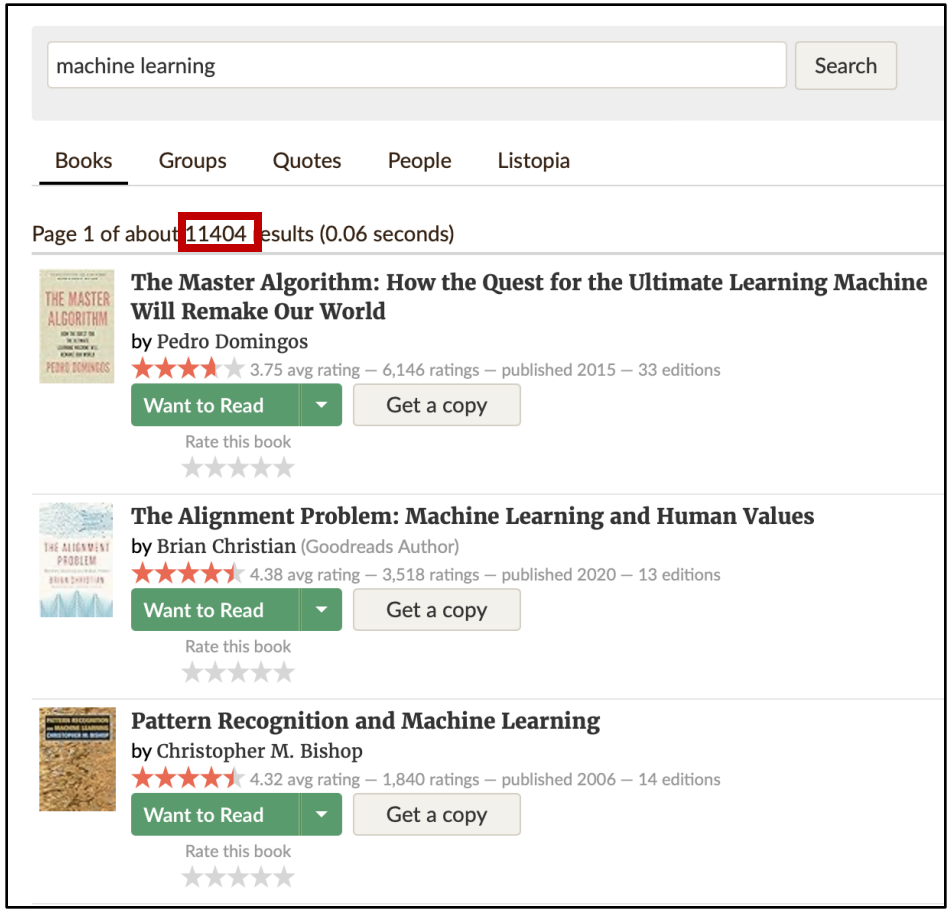
- Practical aspects of training and evaluating Artificial Intelligence Systems.
- Designing and building Human – AI Collaborative Systems
- Getting familiar with some of the most pressing issues in AI research.

This course will not focus on:

- Machine Learning theory.
- Building Large Language Models/ Multimodal models
- Details of Artificial Intelligence techniques.

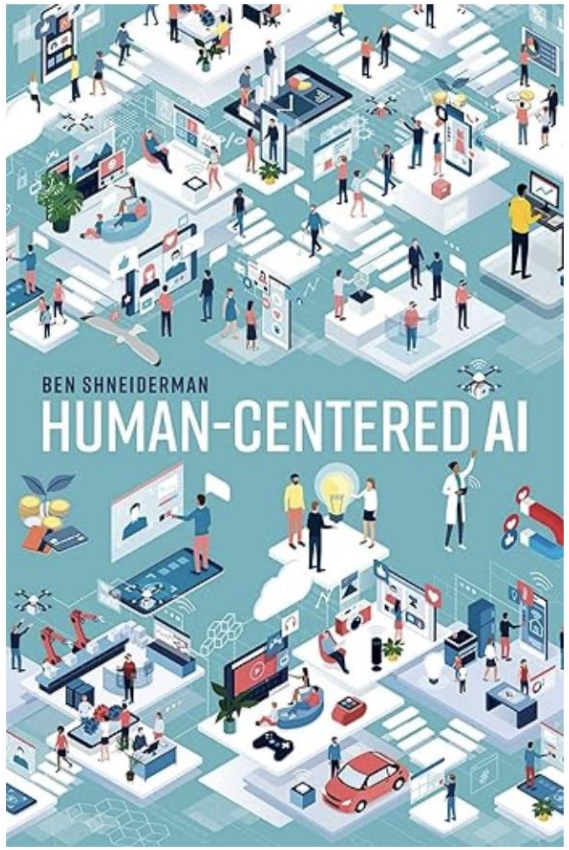
Resources

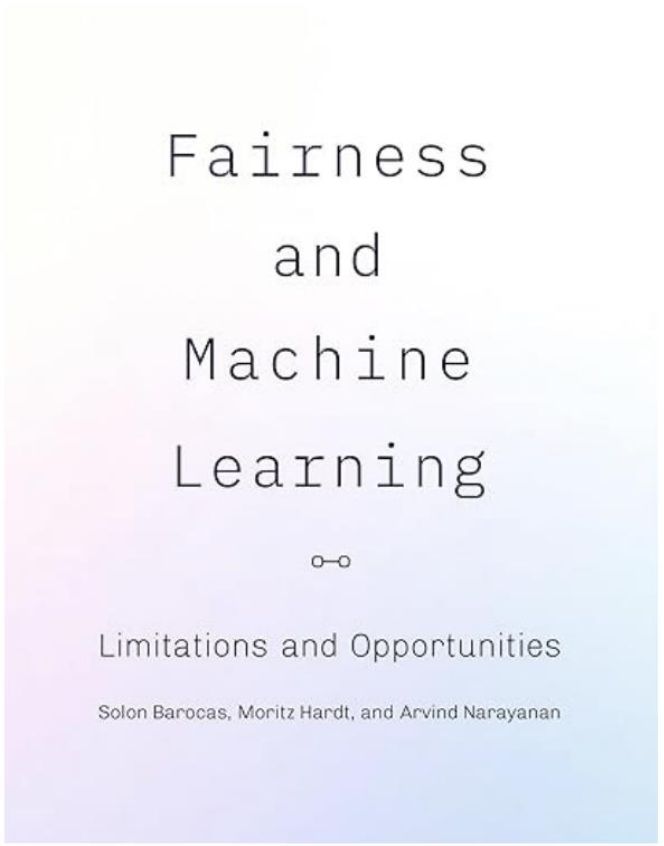


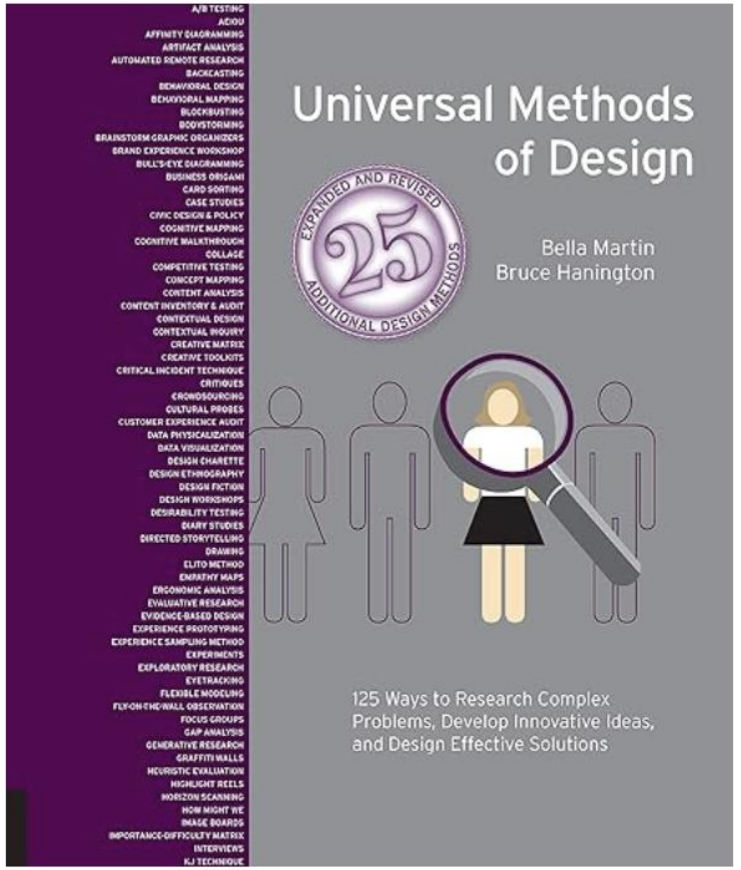


McMaster University

Resources


BEN SHNEIDERMAN
HUMAN-CENTERED AI


Fairness
and
Machine
Learning
Limitations and Opportunities
Solon Barocas, Moritz Hardt, and Arvind Narayanan


Universal Methods of Design
EXPANDED AND REVISED
25
ADDITIONAL DESIGN METHODS
Bella Martin
Bruce Hanington
125 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions

Learning Evaluation

Evaluation	Grade Weight	Responsibility
Assignment I	15%	Paper Presentation and Leading Discussion
Assignment II	15%	Paper Presentation and Leading Discussion
Class Participation	20%	Participating in Discussion (in-person / online)
Project Presentation I	10%	Final Project Proposal
Project Presentation II	10%	Project Results
Final Research Proposal	30%	Proposal Write up

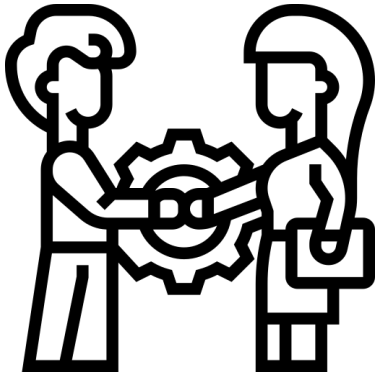
Learning Evaluation

Assignment Presentation Format:

- Sign up for any 2 papers listed on the shared excel sheet; 1 from Group A, and 1 from Group B
- Presentation Time: 20 minutes
- Questions Answer and Discussions: 20 minutes
- Presenter will post on allocated Discussion board on Avenue:
    - a summary of the paper
    - 3-5 questions worth discussing about the paper
- Non-presenters will post their thoughts and responses on the questions on Discussion thread.

Code of Conduct

Be Honest



Be Respectful

Thank you