



Introduction to Human Centered Artificial Intelligence

Swati Mishra

Human Centered Artificial Intelligence

Graduate Course - CAS 783

Winter 2025



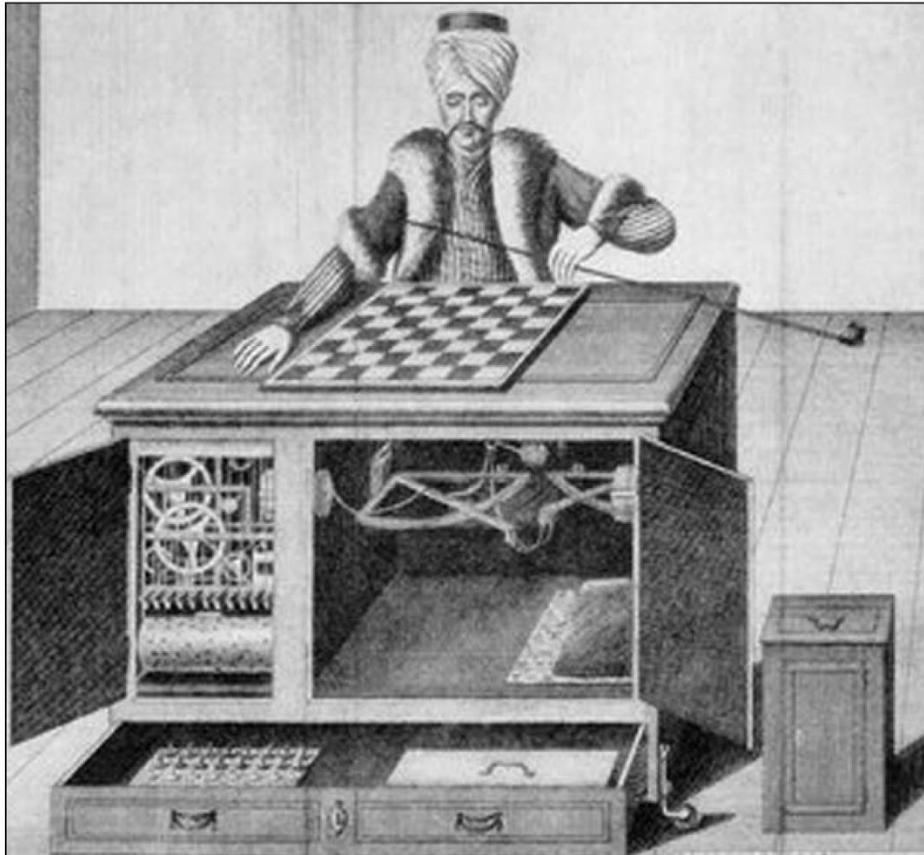
ENGINEERING

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

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...

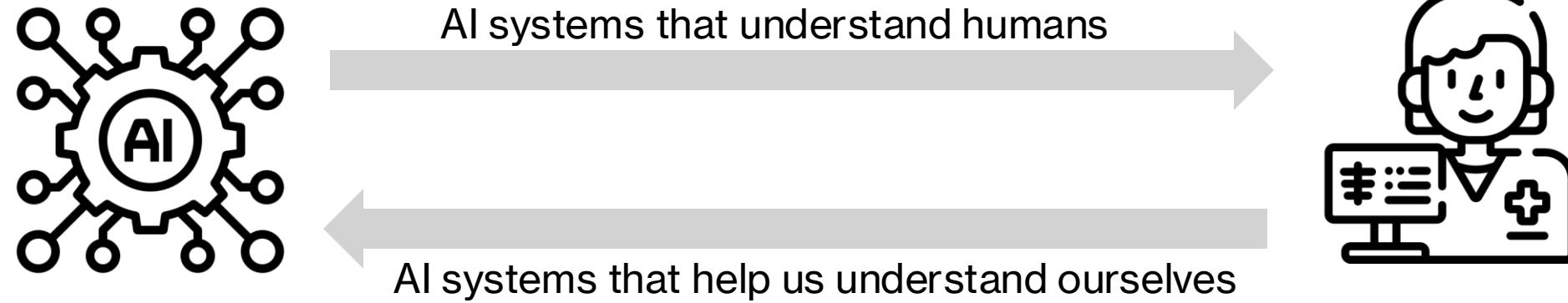
Source: *A (Very) Brief History of Artificial Intelligence*, ACM Magazine

Human Centered Artificial Intelligence

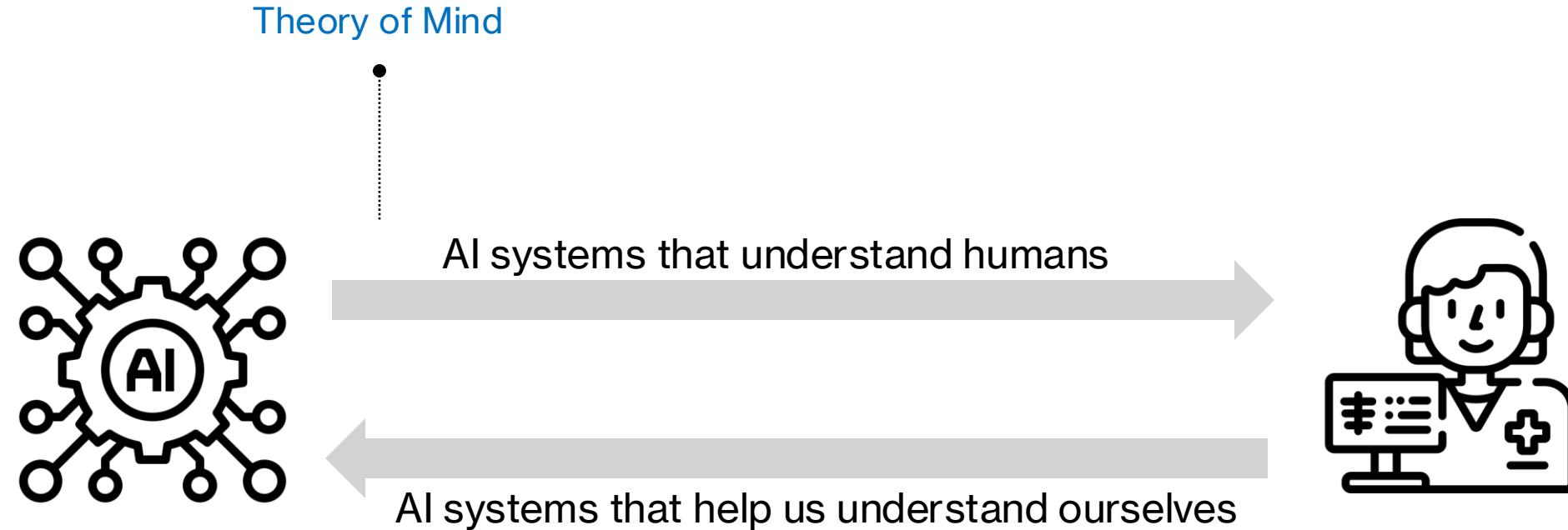
“Human-centered artificial intelligence is a perspective on AI and ML that intelligent systems must be designed with awareness that they are part of a larger system consisting of human stake-holders, such as users, operators, clients, and other people in close proximity.”

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

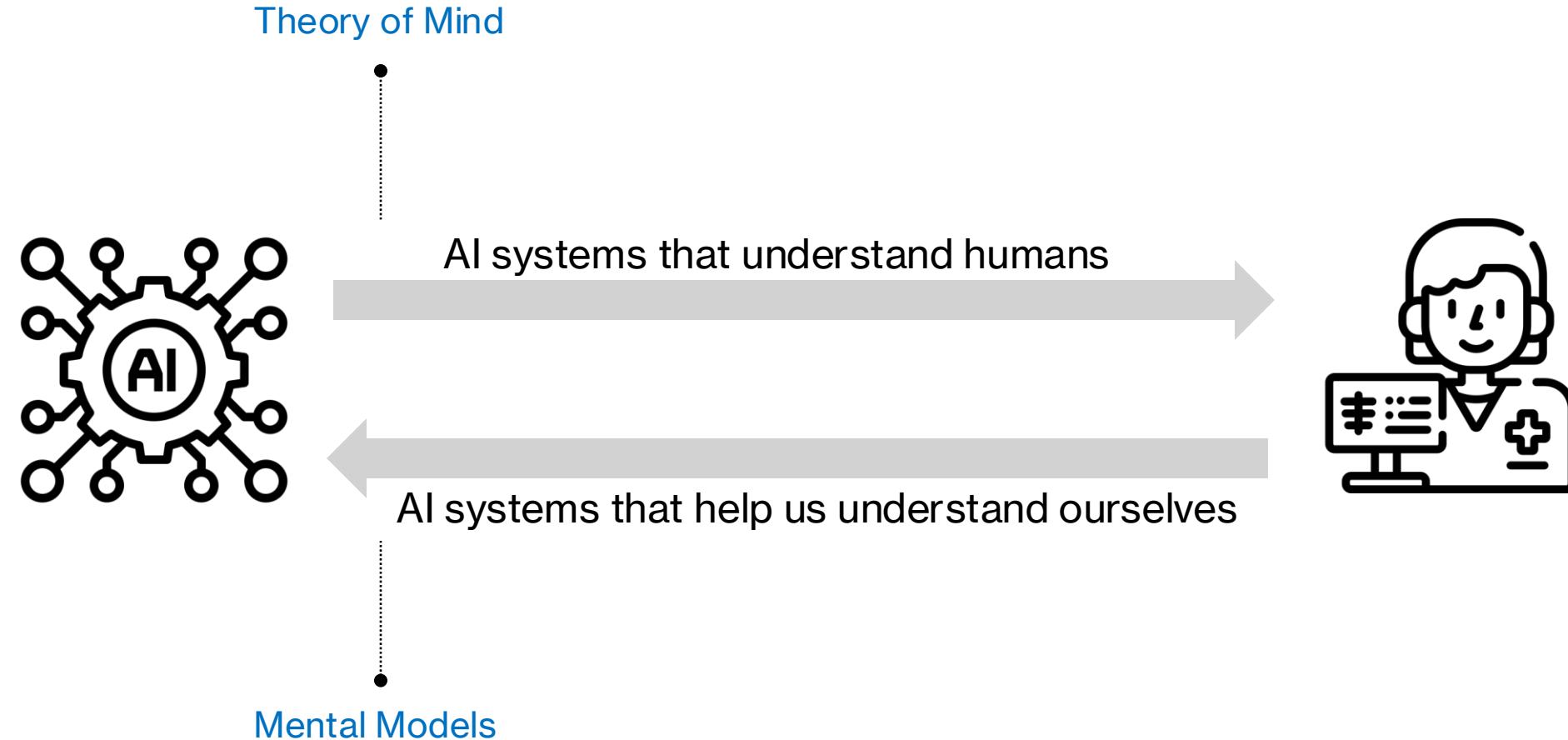
Human Centered Artificial Intelligence



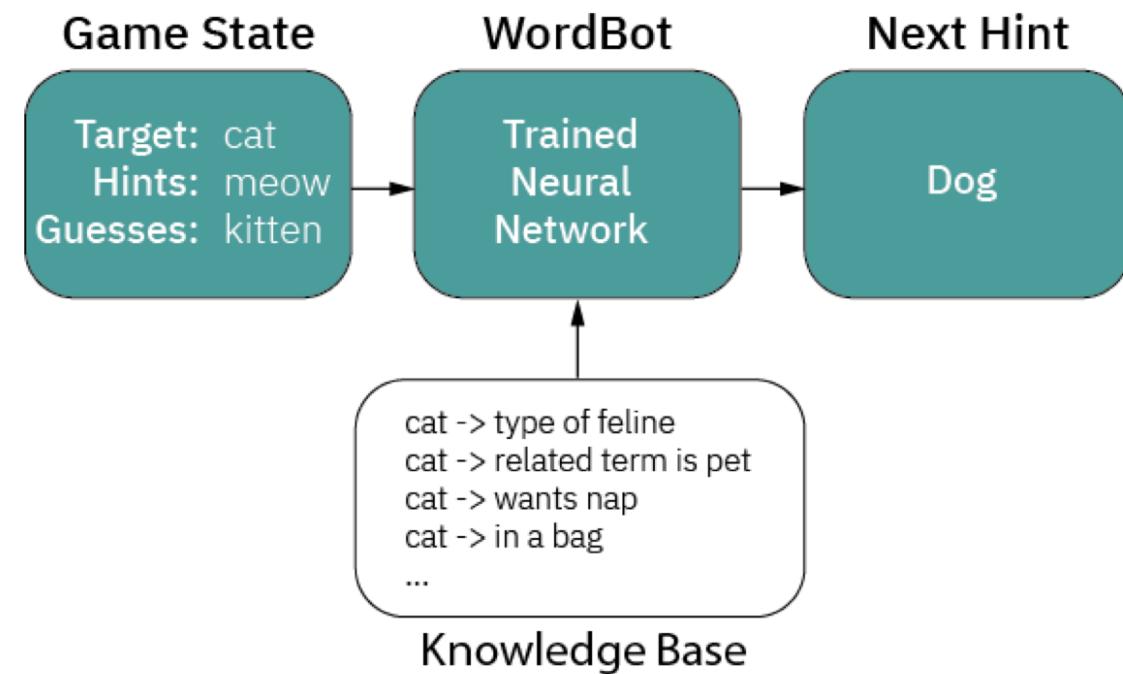
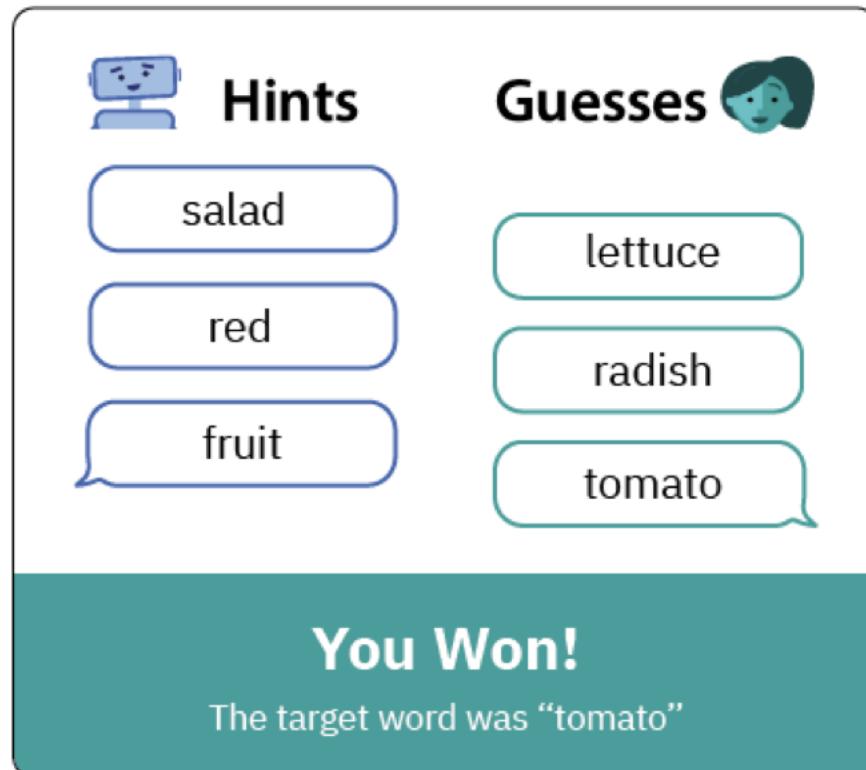
Human Centered Artificial Intelligence



Human Centered Artificial Intelligence



Human Centered Artificial Intelligence



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

Human Centered Artificial Intelligence

The image shows a game interface. On the left, under the heading "Hints" (with a robot icon), are three blue-outlined boxes containing "salad", "red", and "fruit". On the right, under the heading "Guesses" (with a person icon), are three teal-outlined boxes containing "lettuce", "radish", and "tomato". At the bottom, a teal banner displays the text "You Won!" in white, followed by "The target word was ‘tomato’" in white.

Hints	Guesses
salad	lettuce
red	radish
fruit	tomato

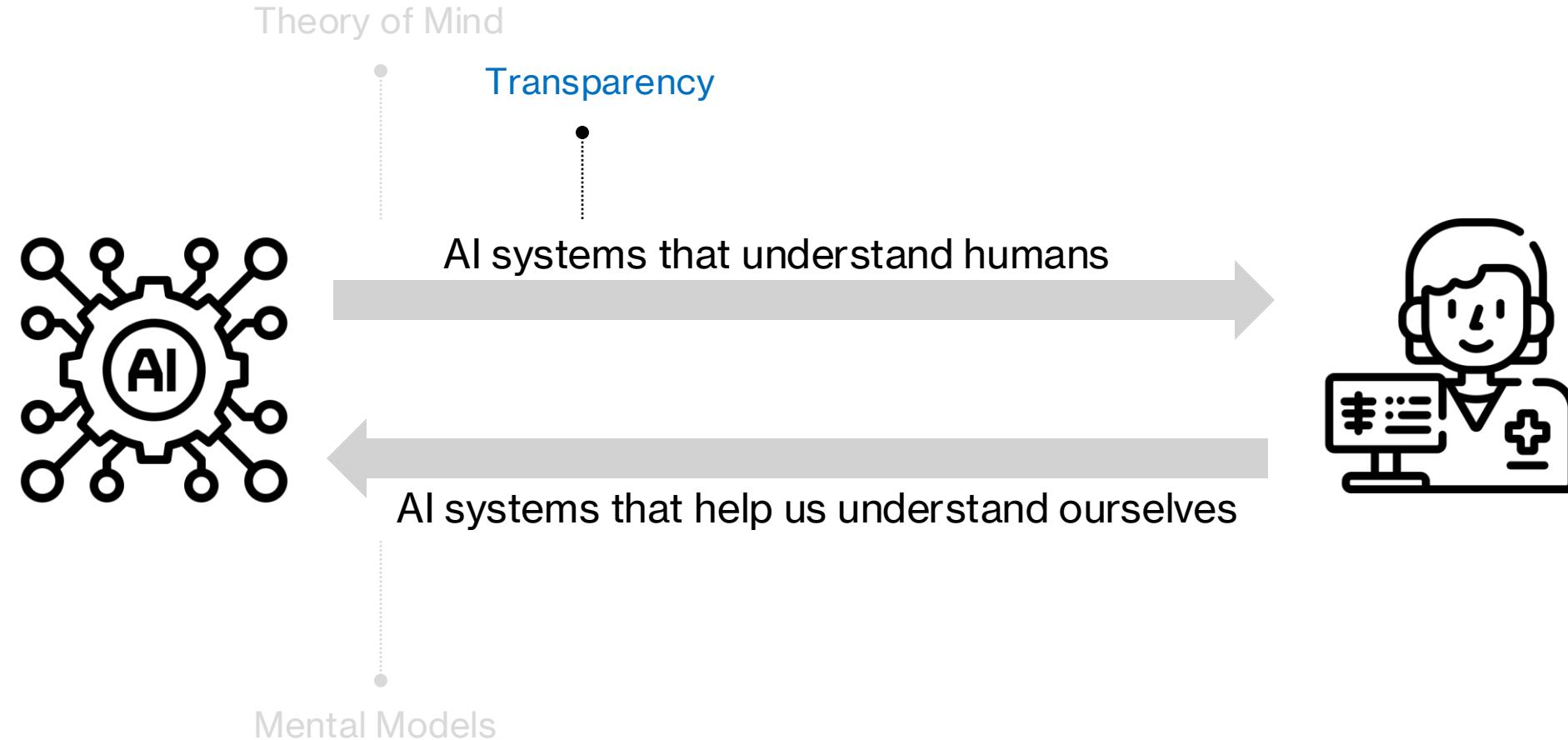
You Won!
The target word was “tomato”

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

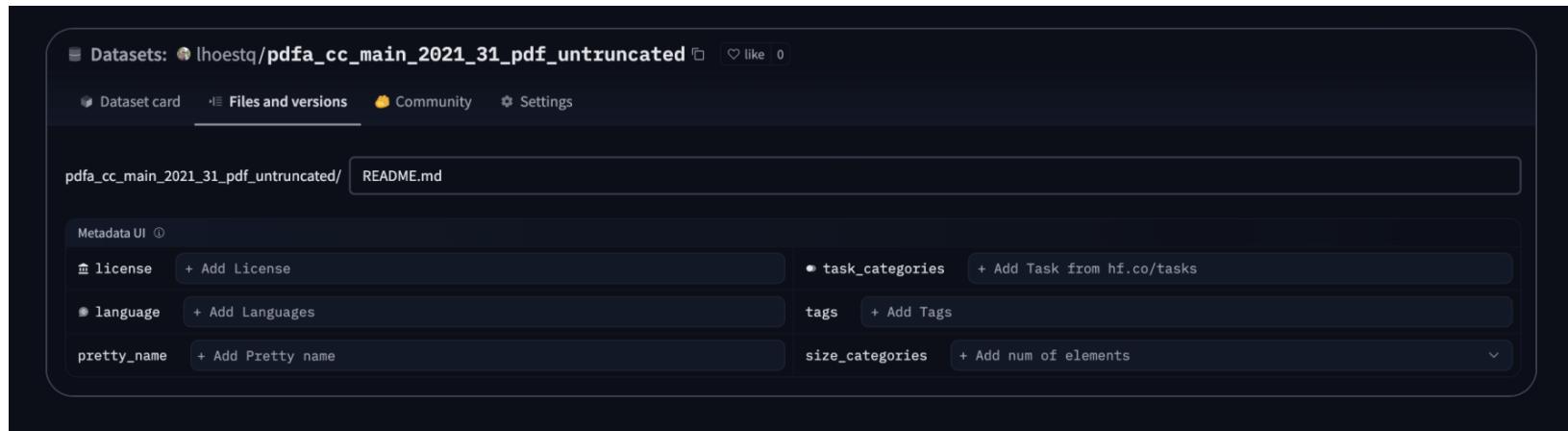


Human Centered Artificial Intelligence

The screenshot shows a Hugging Face Model Card for the model `tomaarsen/setfit-all-MiniLM-L6-v2-sst2-32-shot`. The card includes sections for Model card, Files and versions, and Community. It details the SetFit model trained on the `sst2` dataset, using `sentence-transformers/all-MiniLM-L6-v2` as the Sentence Transformer embedding model. The model has been trained using an efficient few-shot learning technique involving fine-tuning a Sentence Transformer and training a classification head. The card also lists the base model as `sentence-transformers/all-MiniLM-L6-v2`, which is finetuned (186) for this specific task. The dataset used for training is `stanfordnlp/sst2`, updated on Jan 4, 2024, with 70k samples, 11.8k validation samples, and 105 commits. The card also shows download statistics for the last month (31) and provides links for inference examples and the model tree.

Hugging face: Model Cards

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?

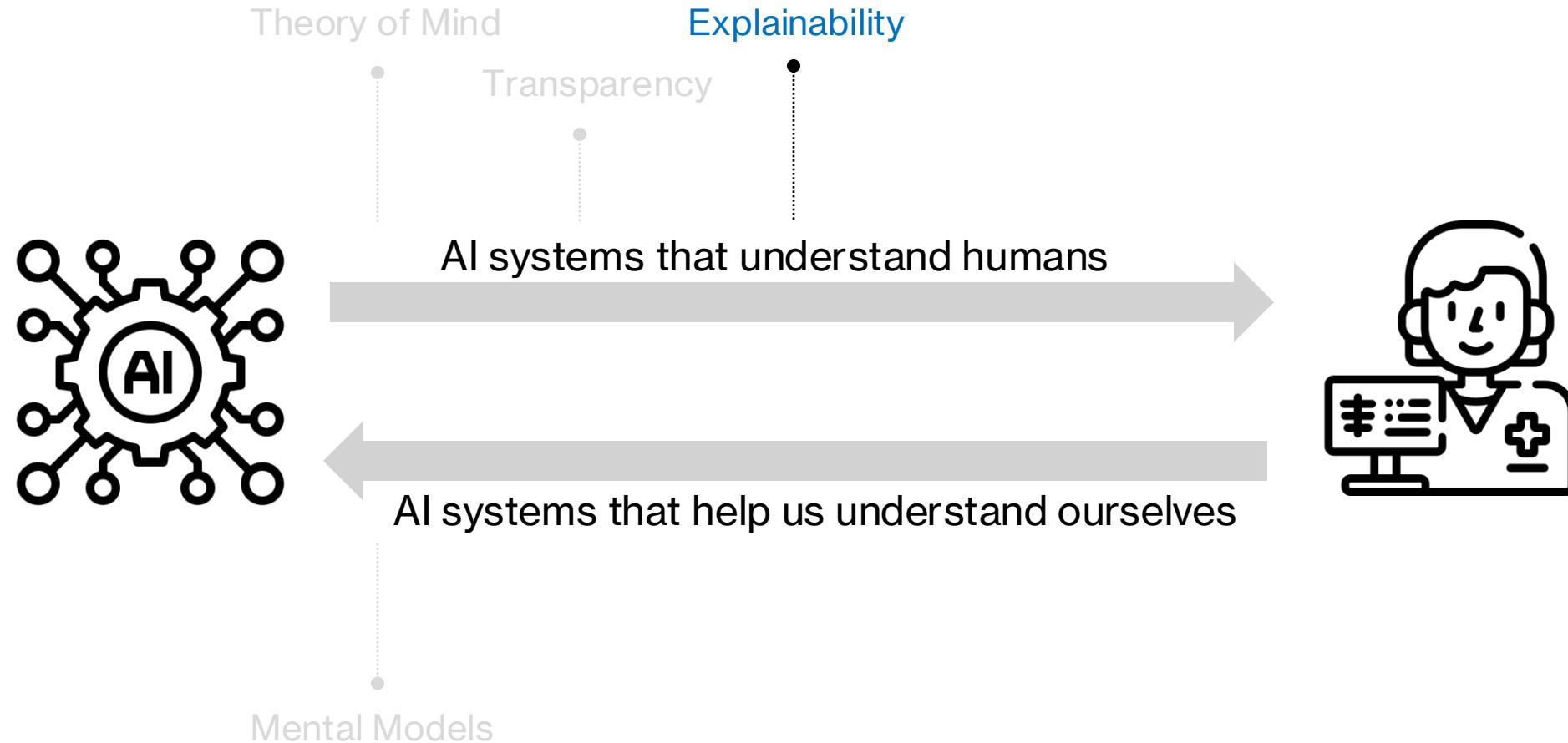
Model Card

- **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: Datasheet for Datasets

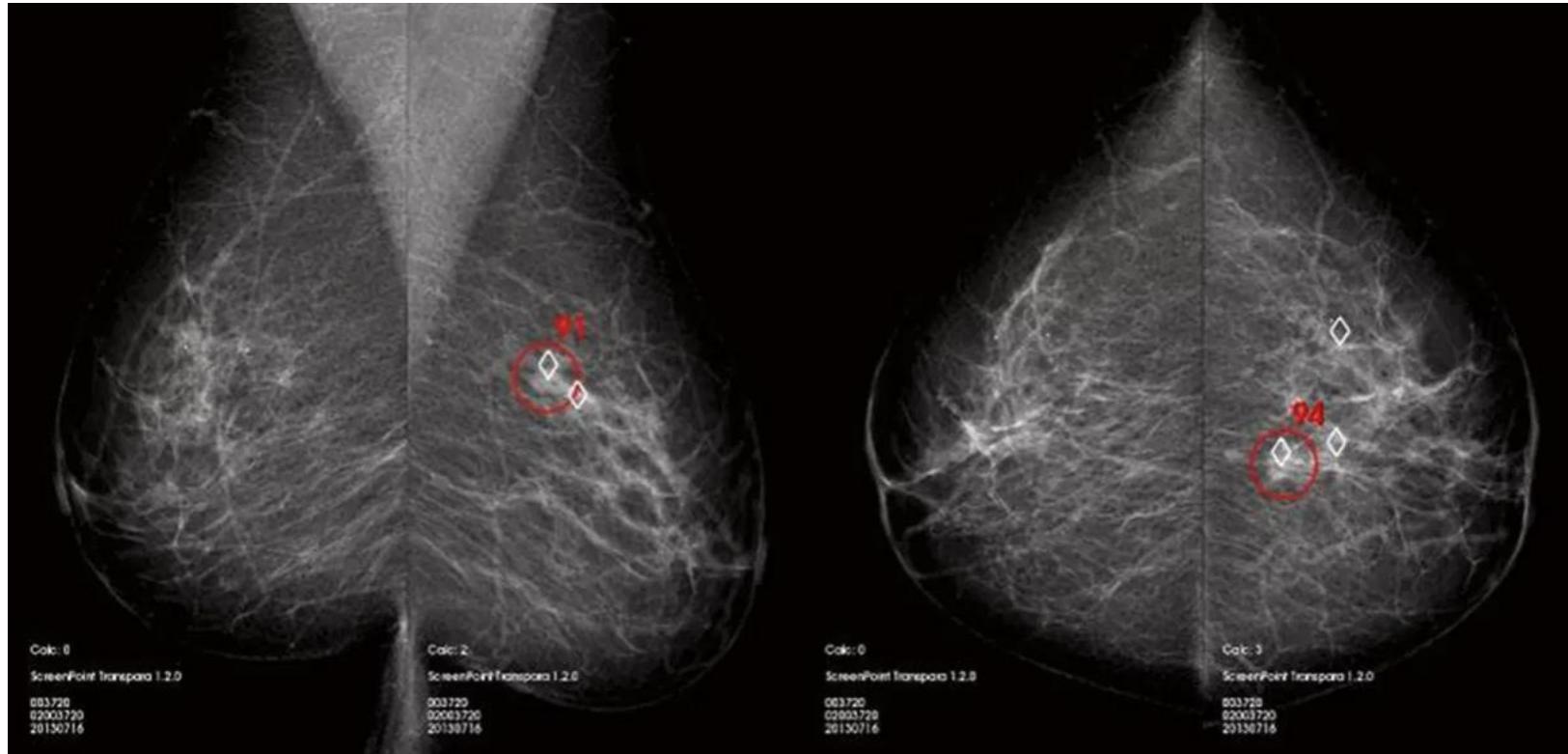
Reference: Model Cards

Human Centered Artificial Intelligence



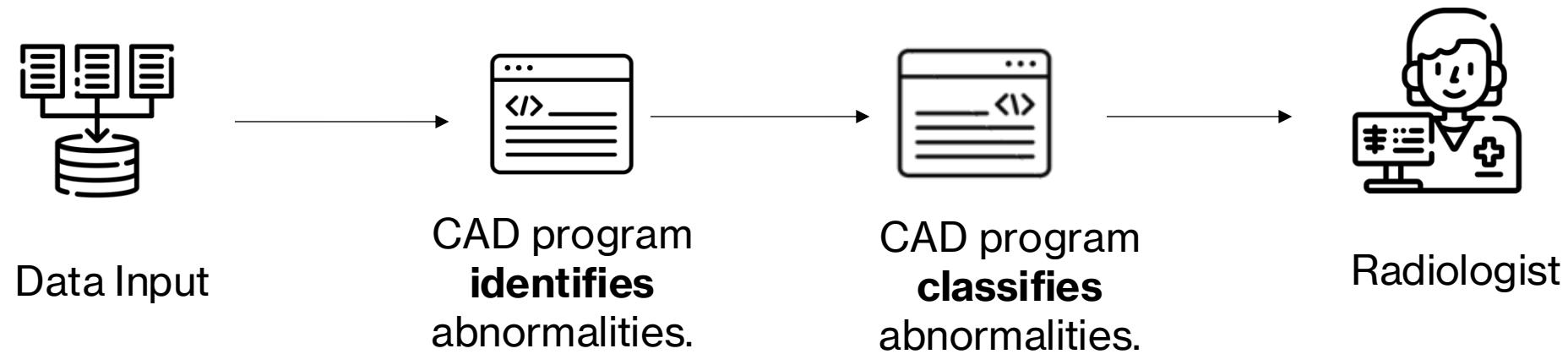
Human Centered Artificial Intelligence

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



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

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

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

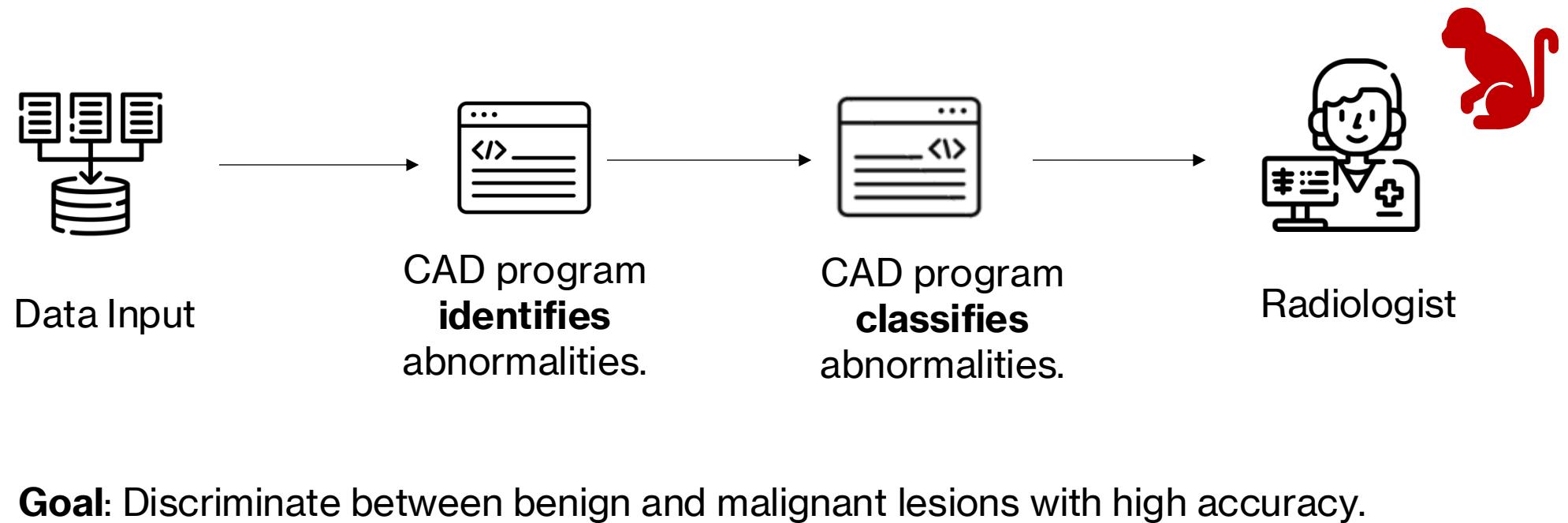
	Correct Classifications (%)		Positive Biopsy Yield (%)	Area Under ROC Curve	
	Malignant	Benign		Full Curve	TPF ≥ 0.90
Computerized method	100	82	76	0.92	0.082
Radiologists	100	27	45	0.89	0.042
P value	...	0.009	0.003	0.21	0.03

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

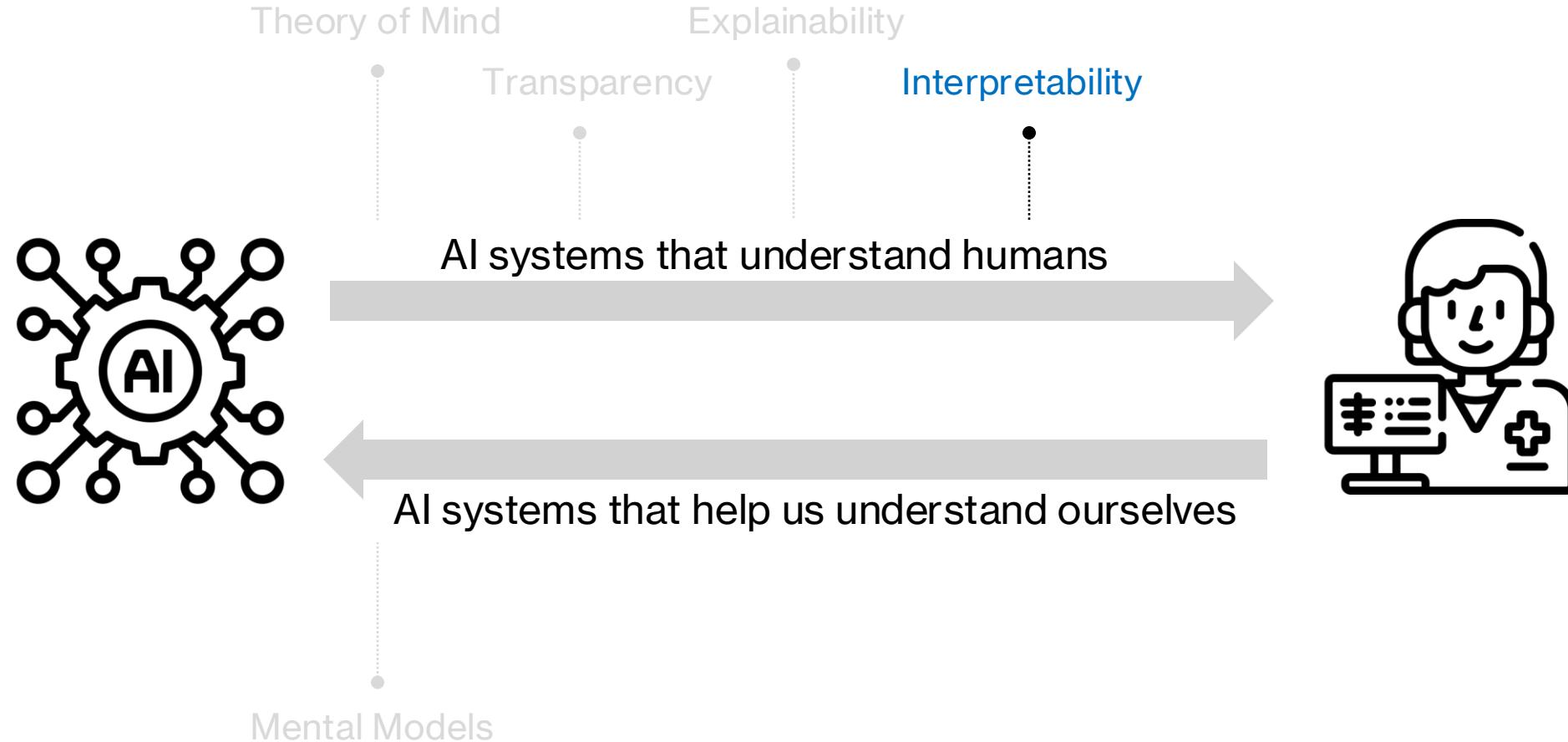
Malignant and benign clustered microcalcifications: automated feature analysis and classification (1996)
Y Jiang, R M Nishikawa, D E Wolverton, C E Metz, M L Giger, R A Schmidt, C J Vyborny, K Doi

Human Centered Artificial Intelligence

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



Human Centered Artificial Intelligence



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

Multivariate

Subject Area

Health and Medicine

Associated Tasks

Classification

Feature Type

Integer, Real

Instances

858

Features

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

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

P(cancer = Yes | patient smokes = Yes)

		Age	Feature	Integer	Age
0					
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

Playground

Load a preset... Save View code Share ...

Model

text-davinci-002

Temperature 0.7

Maximum length 256

Stop sequences Enter sequence and press Tab

Write a paragraph on how ice cream shops are awesome.

There's something about ice cream shops that just make them awesome. Maybe it's the delicious smell of ice cream that fills the air as soon as you walk in, or the wide variety of flavors to choose from. Whatever the reason, ice cream shops are definitely awesome!

Playground

Load a preset... Save View code Share ...

Model

text-davinci-002

Temperature 0.7

Maximum length 256

Stop sequences

Write a paragraph on how ice cream shops are awesome.

There's something about an ice cream shop that just makes it feel like summertime. They always seem to have the perfect flavor for whatever mood you're in, and their ice cream is always so creamy and delicious. Plus, they always have those awesome toppings that make your ice cream even better!

GPT Ancestor

A Note on Interpretability

Playground

Load a preset... Save View code Share ...

Model

text-davinci-002

Temperature 0.7

Maximum length 256

Stop sequences Enter sequence and press Tab

Write a paragraph on how ice cream shops are awesome.

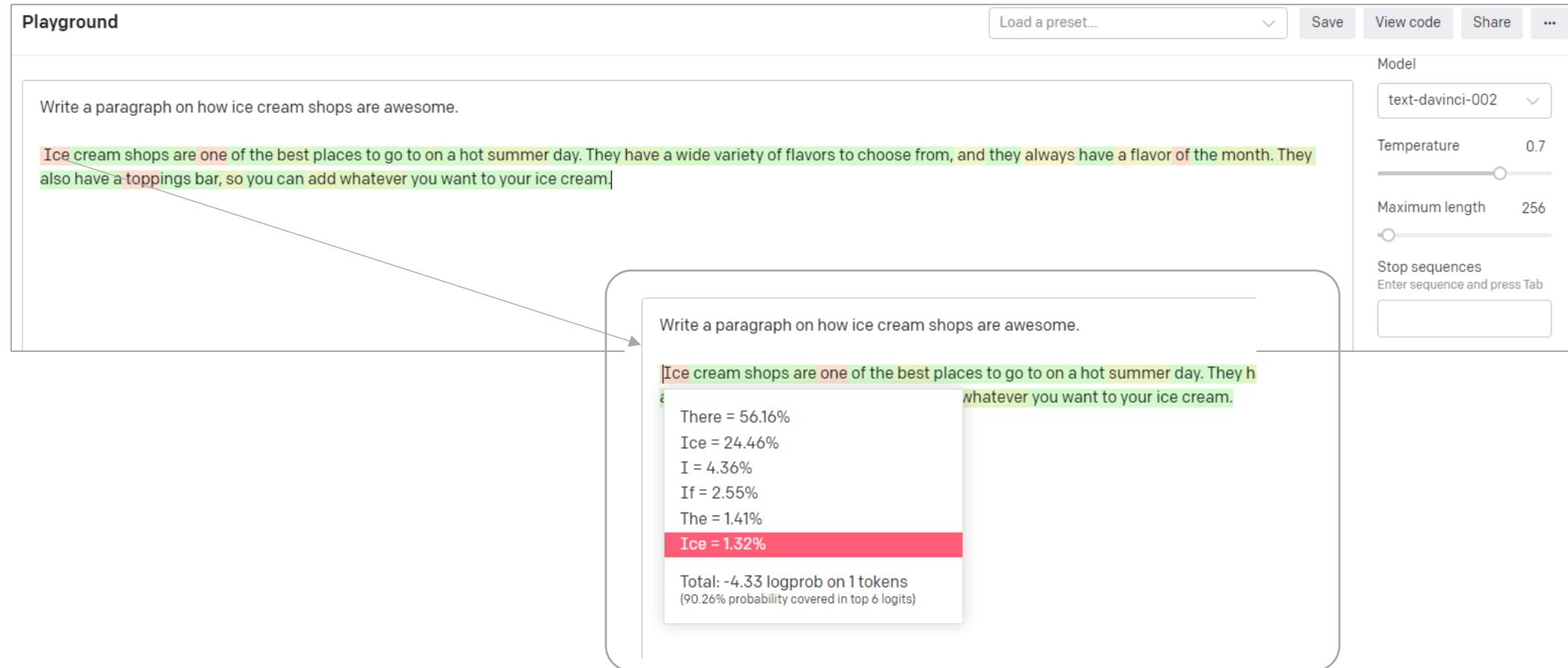
Ice cream shops are one of the best places to go to on a hot summer day. They have a wide variety of flavors to choose from, and they always have a flavor of the month. They also have a toppings bar, so you can add whatever you want to your ice cream.

Write a paragraph on how ice cream shops are awesome.

Ice cream shops are one of the best places to go to on a hot summer day. They h

There = 56.16%
Ice = 24.46%
I = 4.36%
If = 2.55%
The = 1.41%
Ice = 1.32%

Total: -4.33 logprob on 1 tokens
(90.26% probability covered in top 6 logits)



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



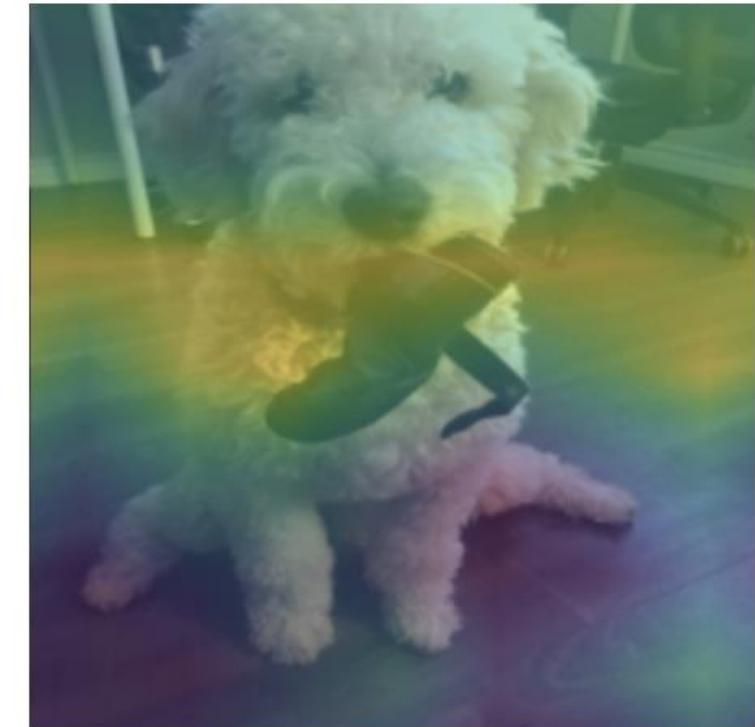
Thinking Exercise!

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



Not
Important

Most
Important

A Note on Interpretability



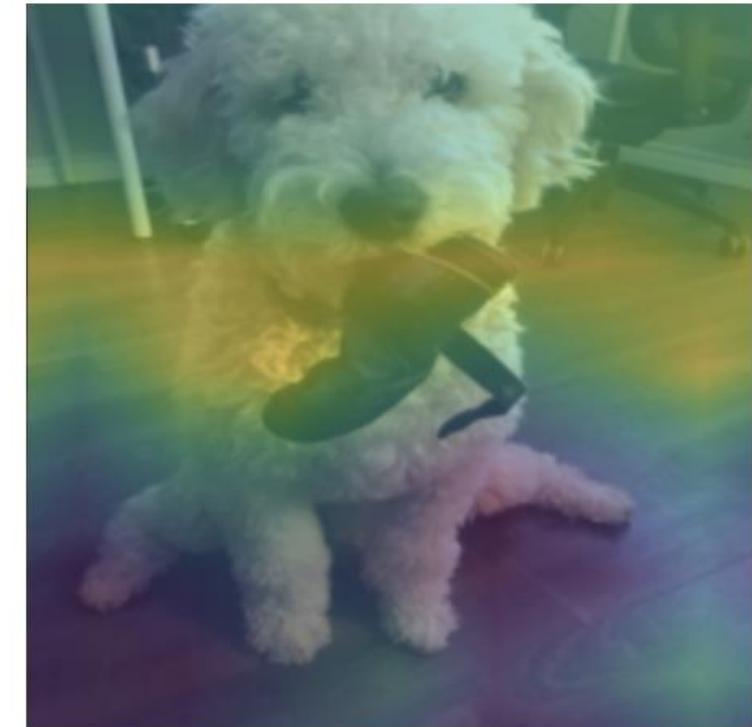
Thinking Exercise!

What does VGG-16 think this is?

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

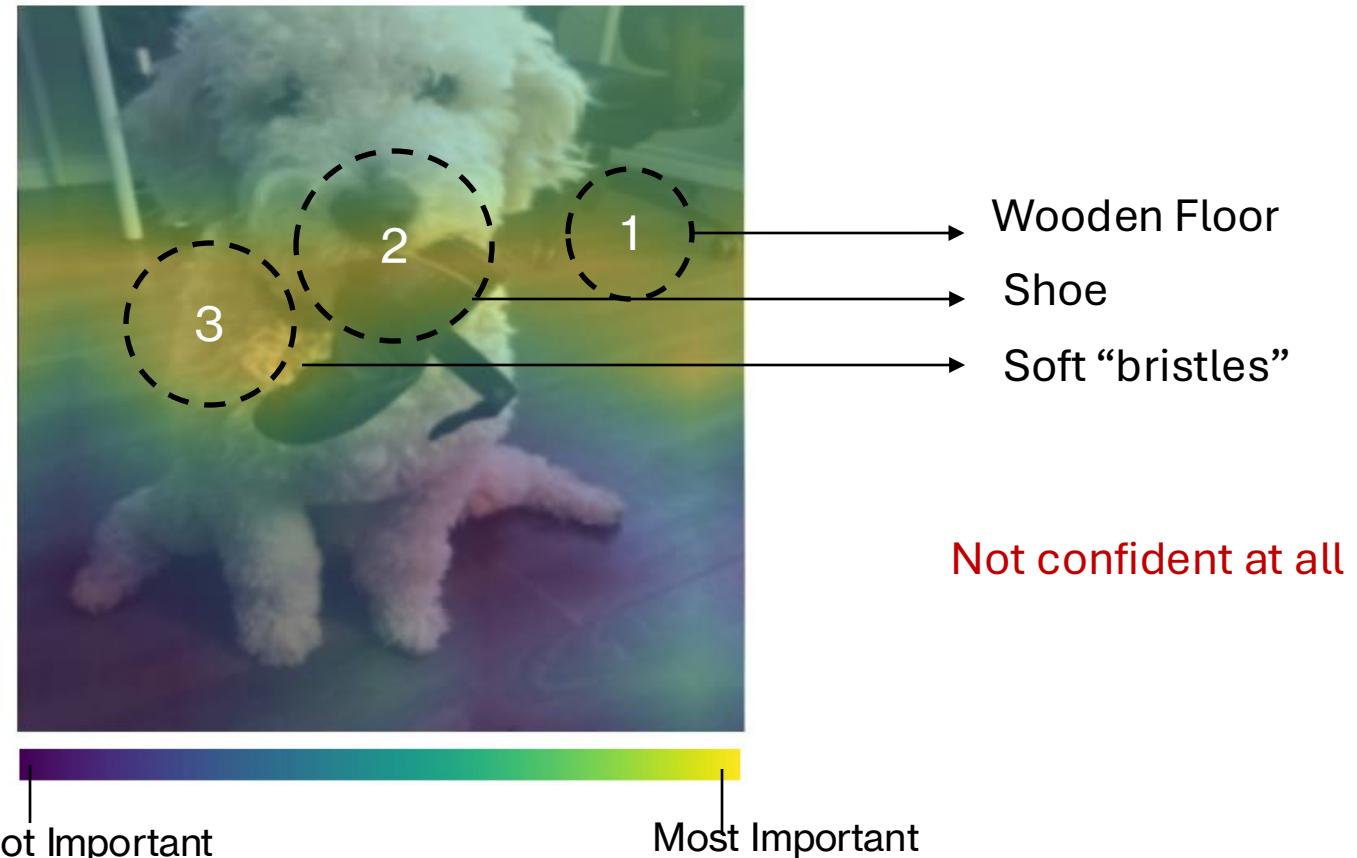
How confident are you that VGG-16 is good at identifying dogs and why?

Explanation of why this is a broom



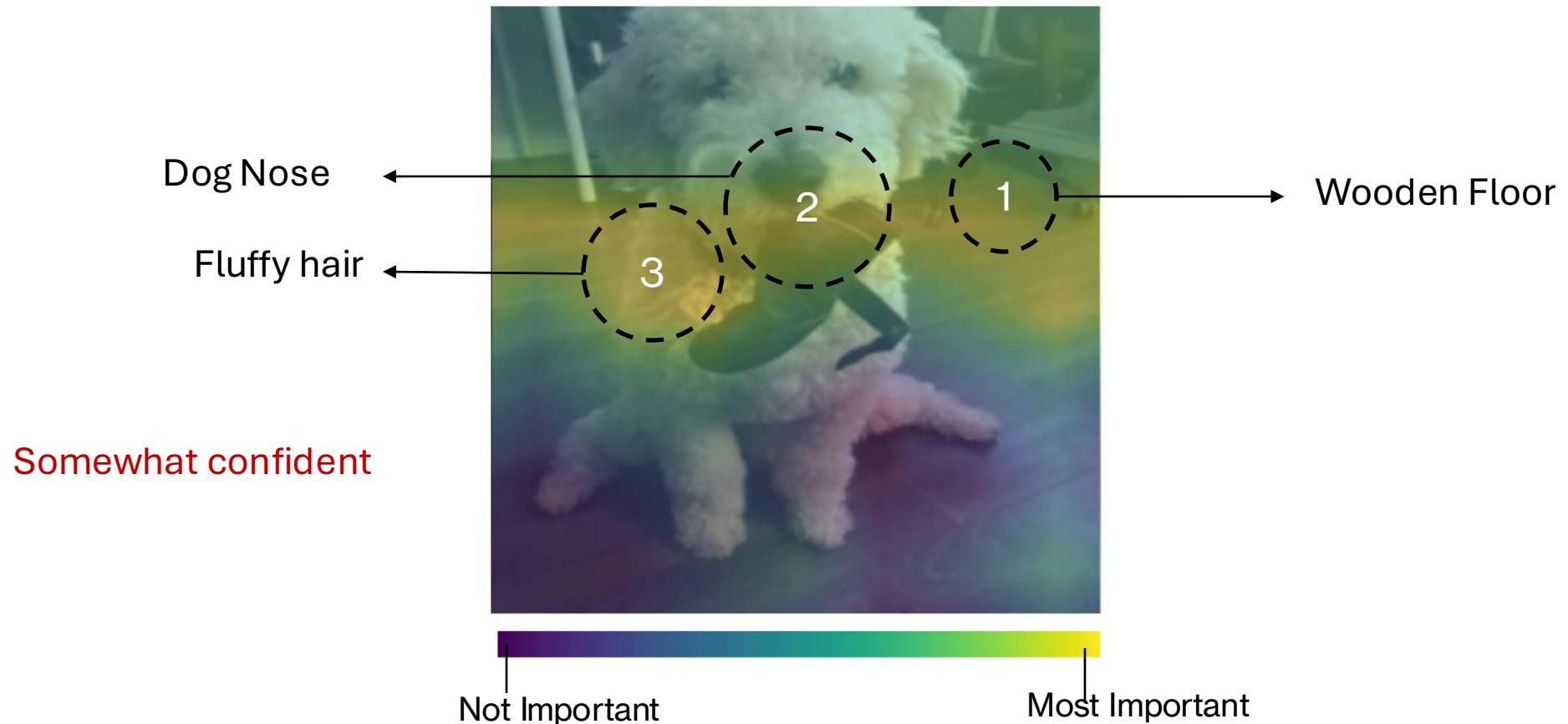
A Note on Interpretability

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

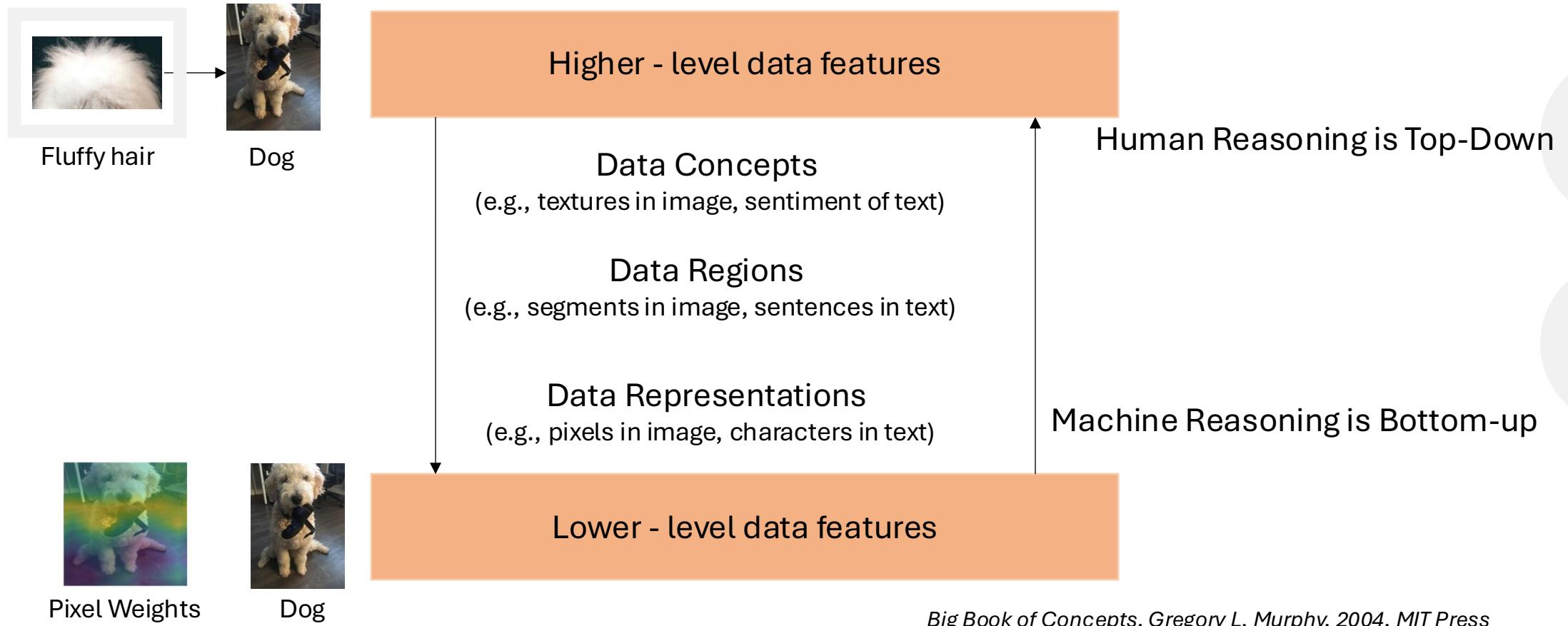


A Note on Interpretability

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



A Note on Interpretability



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

A Note on Interpretability

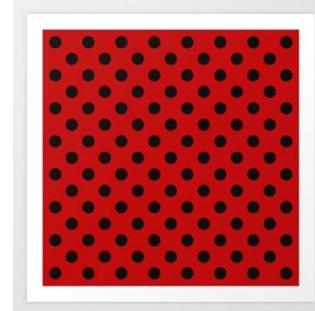
A concept is a mental representation of classes or categories of things which help individuals reason about it.



Yellow Stripes



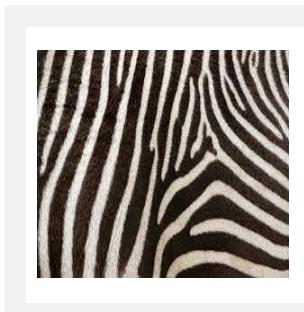
Tiger



Dots



Ladybug



Black Stripes



Zebra



Strings

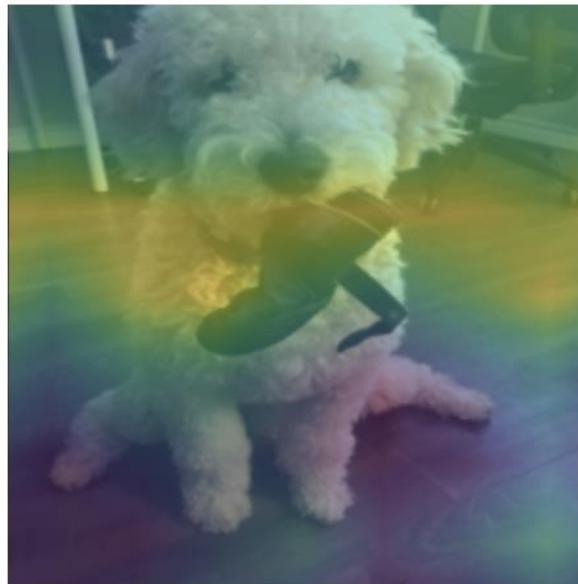


Guitar

A Note on Interpretability

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

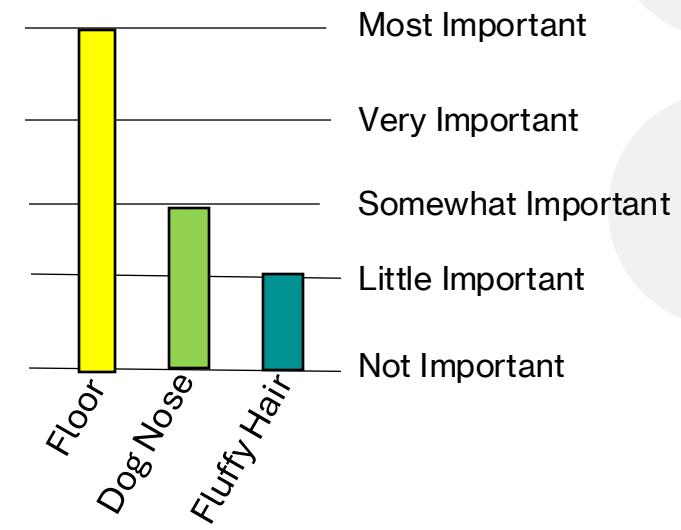
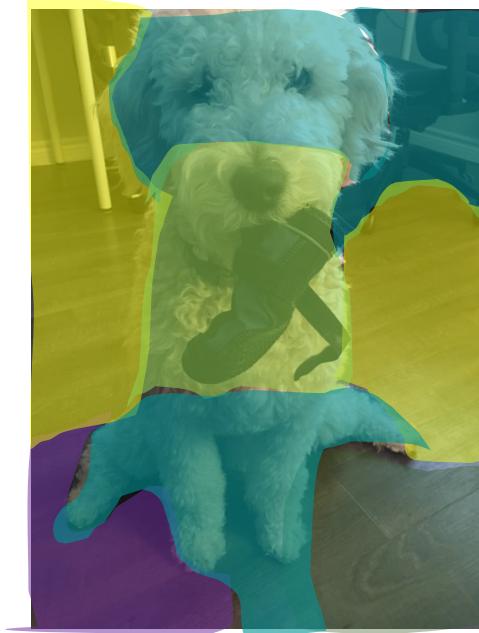
Machine Feature-based Explanation



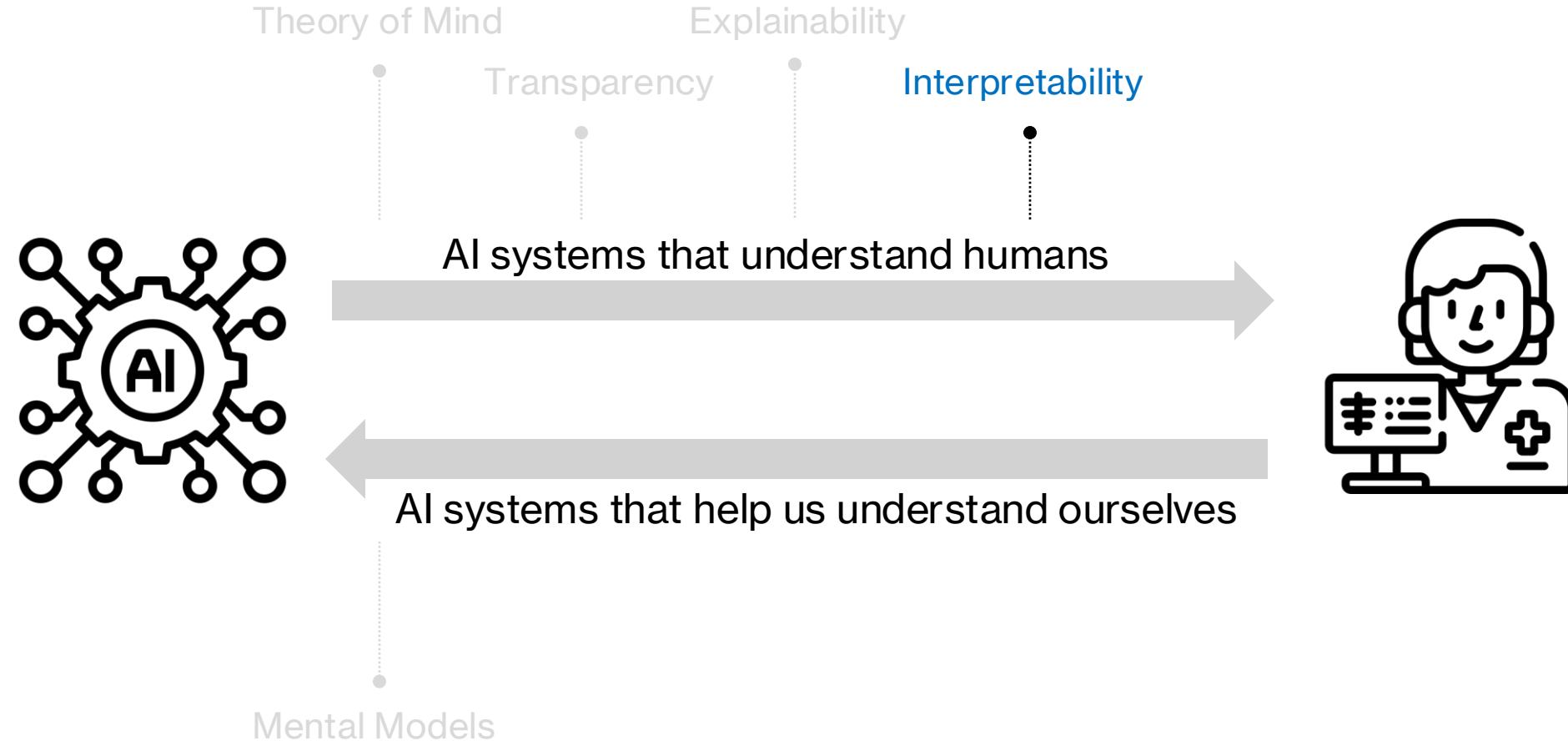
Translate to



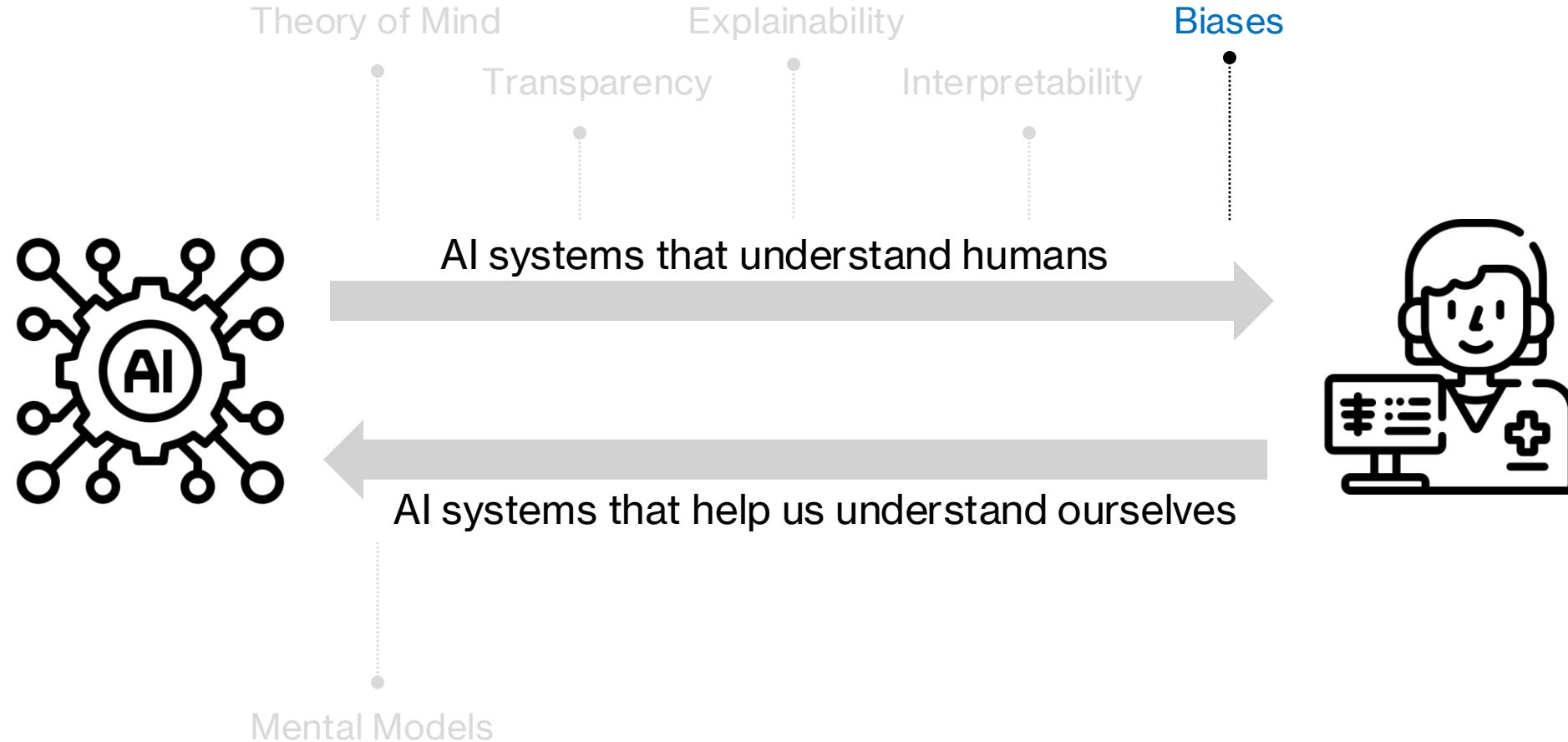
Human Concept-based Explanation



Human Centered Artificial Intelligence



Human Centered Artificial Intelligence



Human Centered Artificial Intelligence

Bloomberg the Company & Its Products ▾ | Bloomberg Terminal Demo Request |  Bloomberg Anywhere Remote Login | Bloomberg Customer Support

Bloomberg

● Live TV Markets ▾ Economics Industries Tech Politics Businessweek Opinion More ▾



FINANCE

Wells Fargo rejected nearly half of their Black homeowners refinancing applications

BY AMIAH TAYLOR
March 16, 2022 at 5:53 PM EDT



A news article from Bloomberg's Finance section. The headline discusses Wells Fargo's discriminatory lending practices against Black homeowners. The article is by Amiah Taylor and was published on March 16, 2022.



Flitter 2022, Donnan et al. 2022

Human Centered Artificial Intelligence



[World](#) ▾ [US Election](#) [Business](#) ▾ [Markets](#) ▾ [Sustainability](#) ▾ [Legal](#) ▾ [Breakingviews](#) ▾ [Technology](#) ▾

World

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

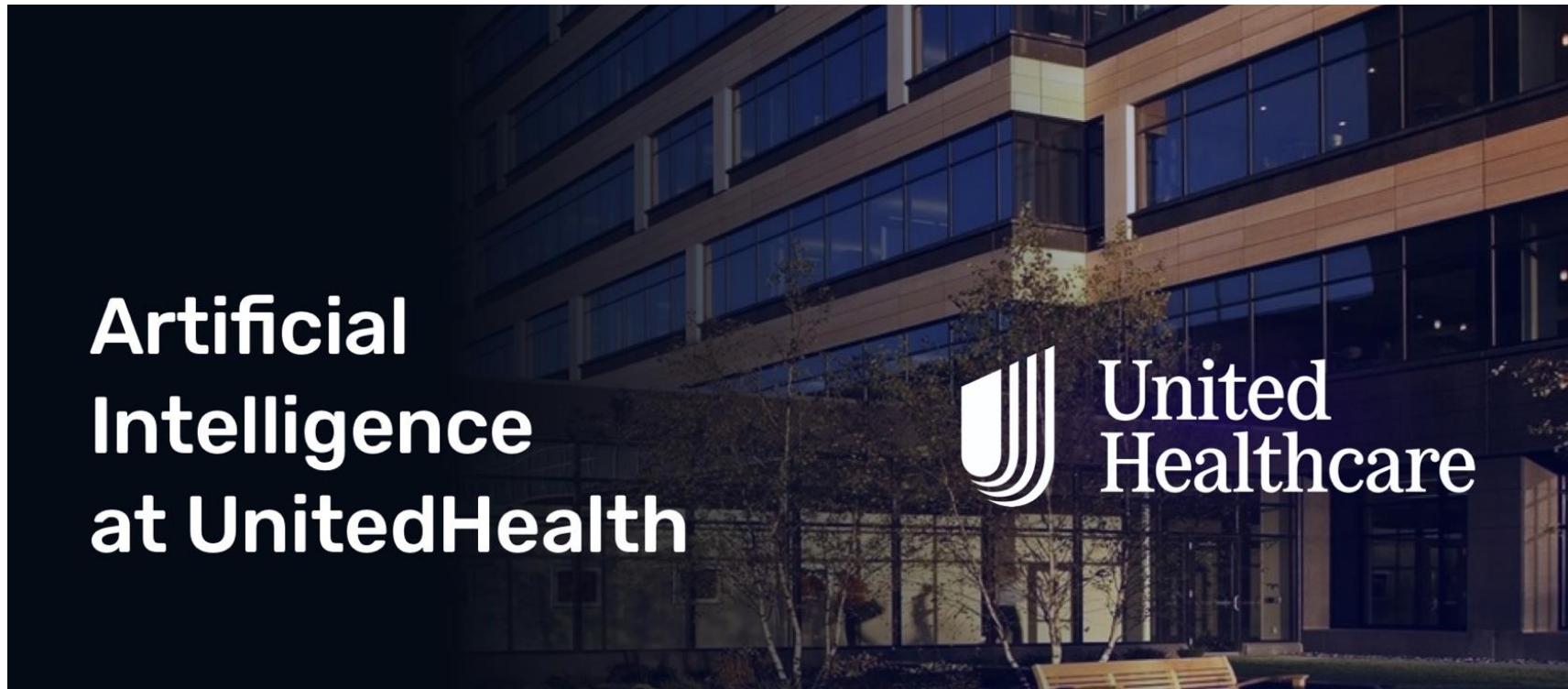
By Jeffrey Dastin

October 10, 2018 8:50 PM EDT · Updated 6 years ago

A small, light gray rounded square containing the letters "Aa", representing a font size adjustment icon.

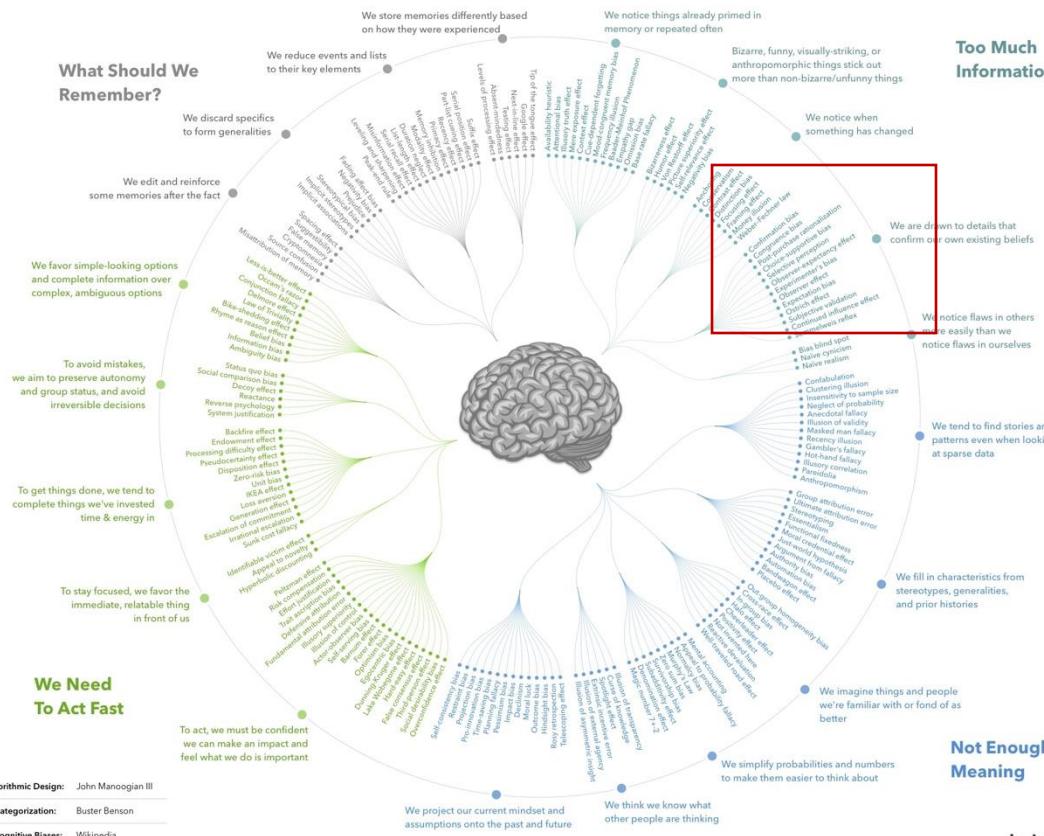


Human Centered Artificial Intelligence

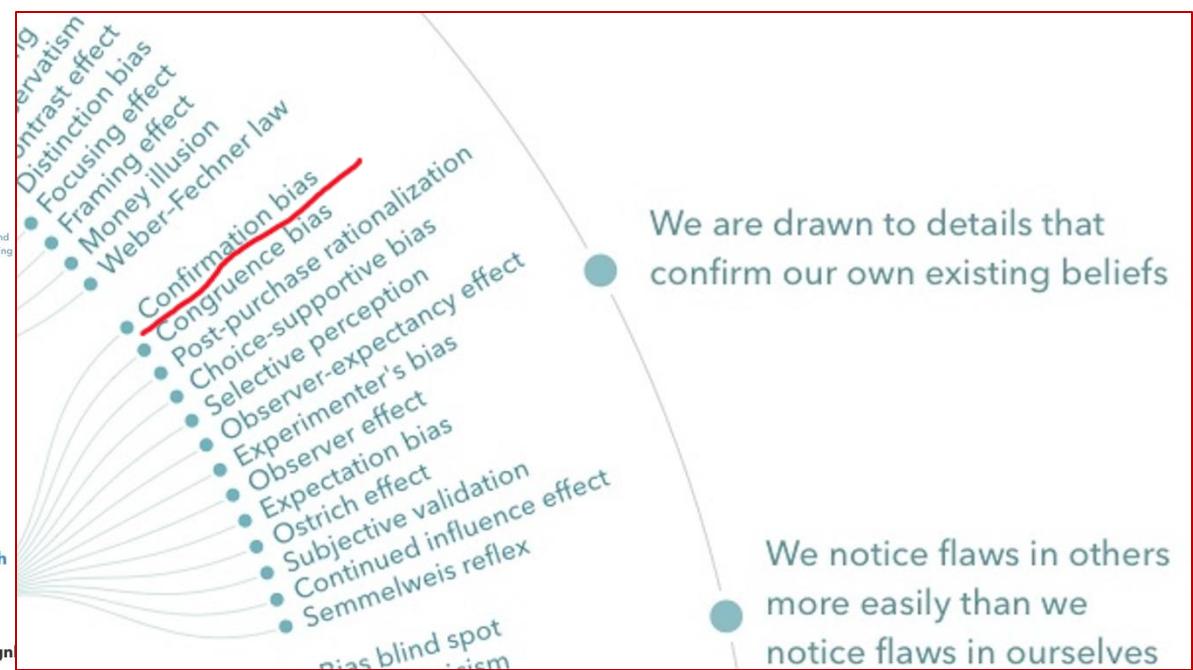


Types of Biases

COGNITIVE BIAS CODEX

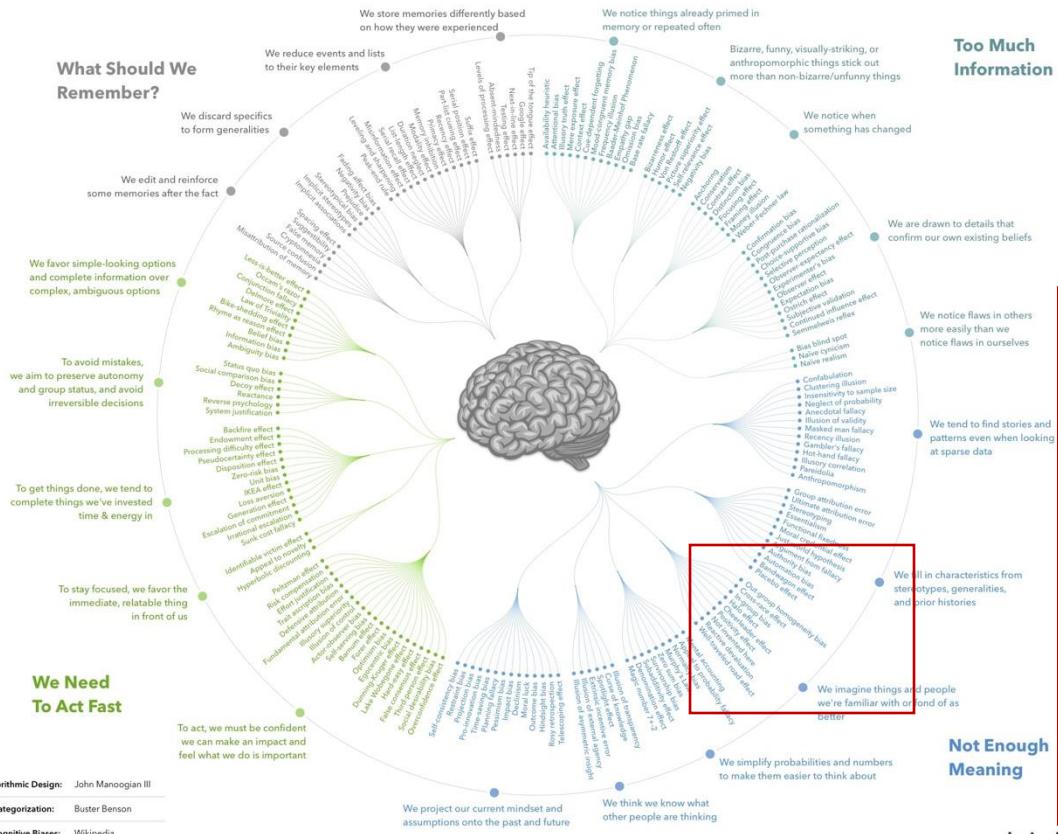


Confirmation Bias: Confirmation bias is the tendency to search for, interpret, favor, and recall information in a way that confirms one's beliefs or hypothesis while giving disproportionately less attention to information that contradicts it.

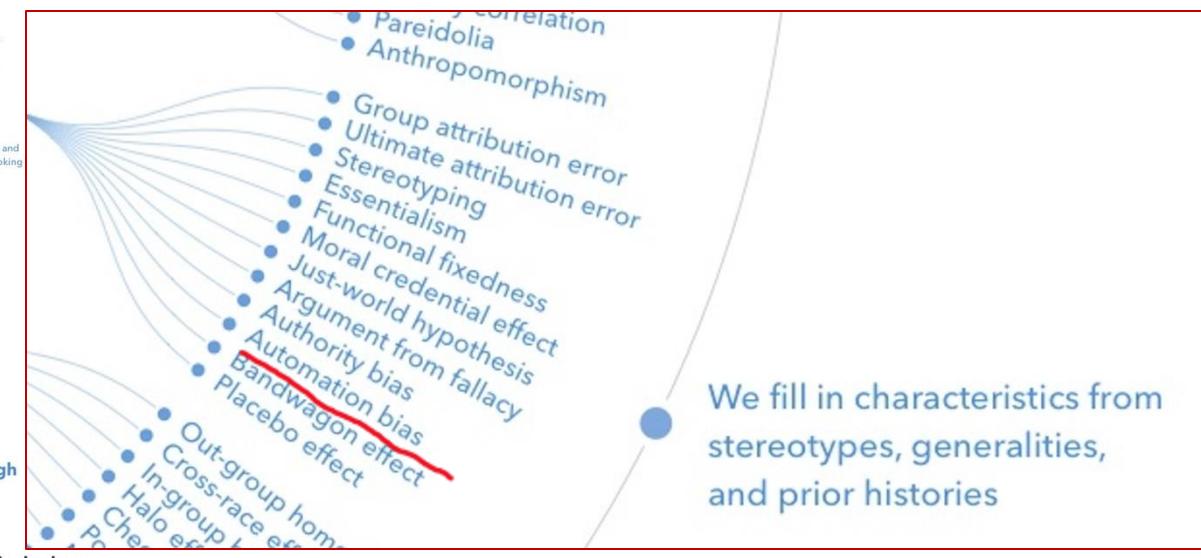


Human Centered Artificial Intelligence

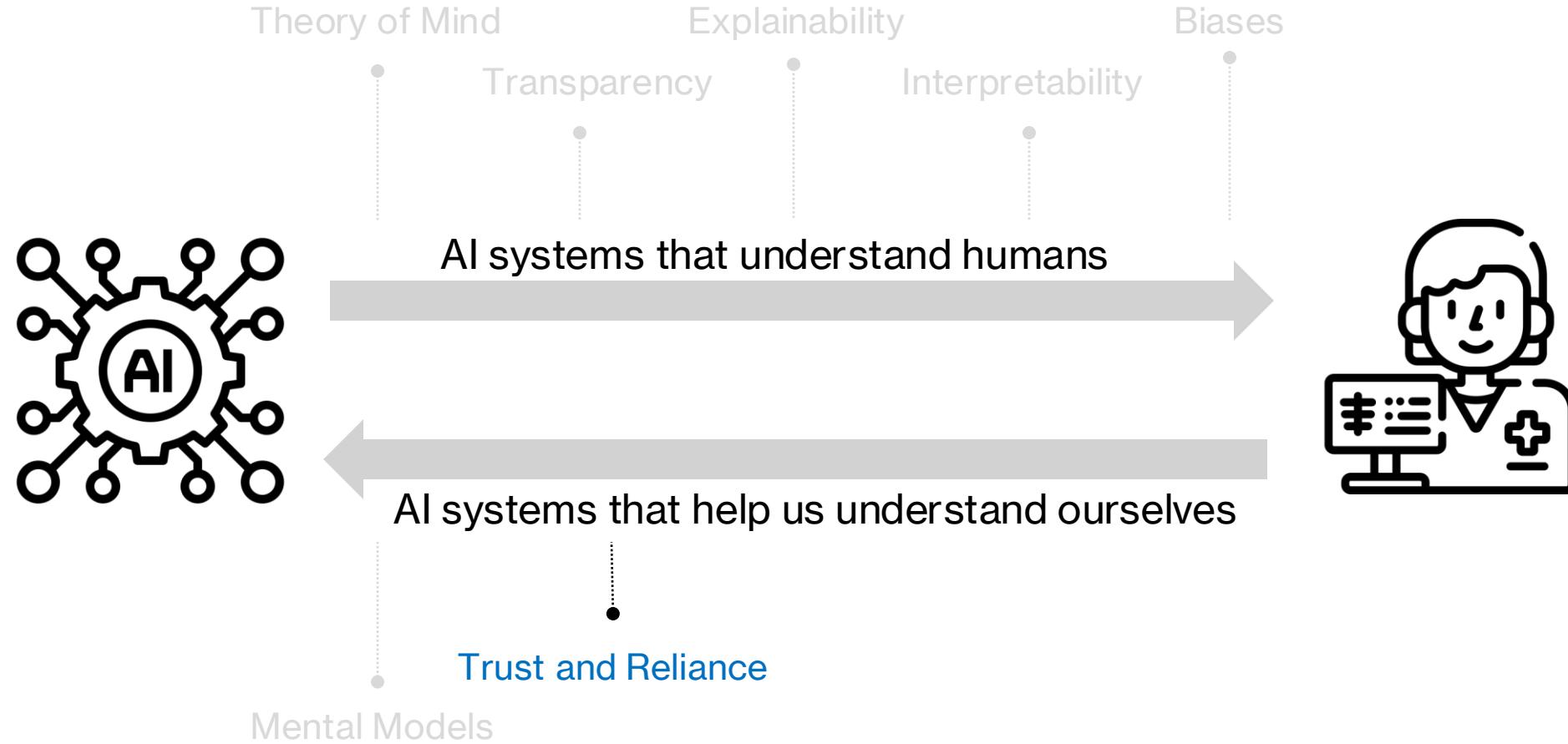
COGNITIVE BIAS CODEX



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.



Human Centered Artificial Intelligence



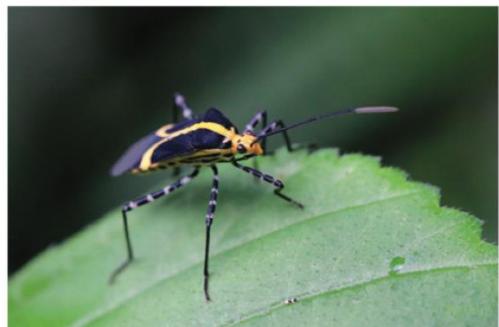
Human Centered Artificial Intelligence



(a) Desert Tarantula



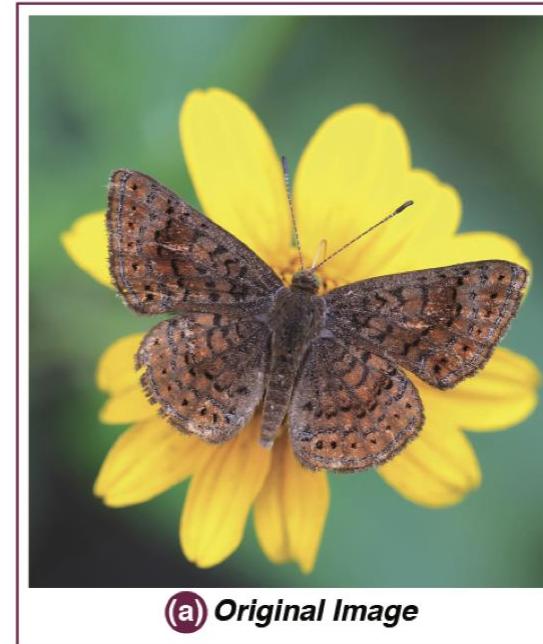
(b) Ecuadorian Lubber Grasshopper



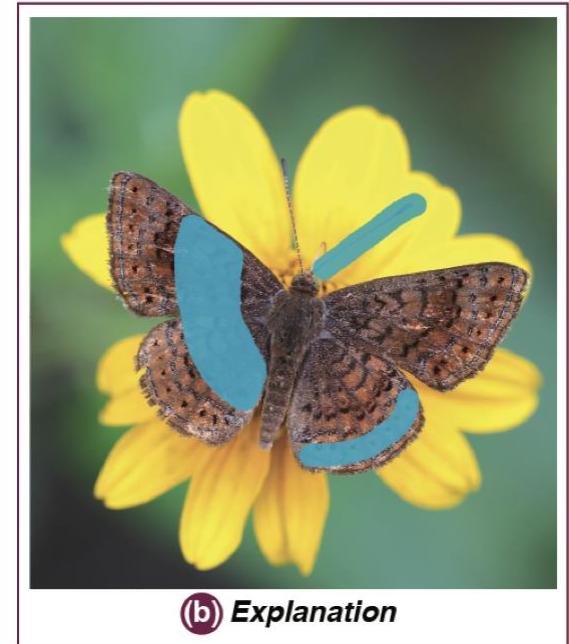
(c) Rhopalid
(scentless plant bug)



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



(a) Original Image

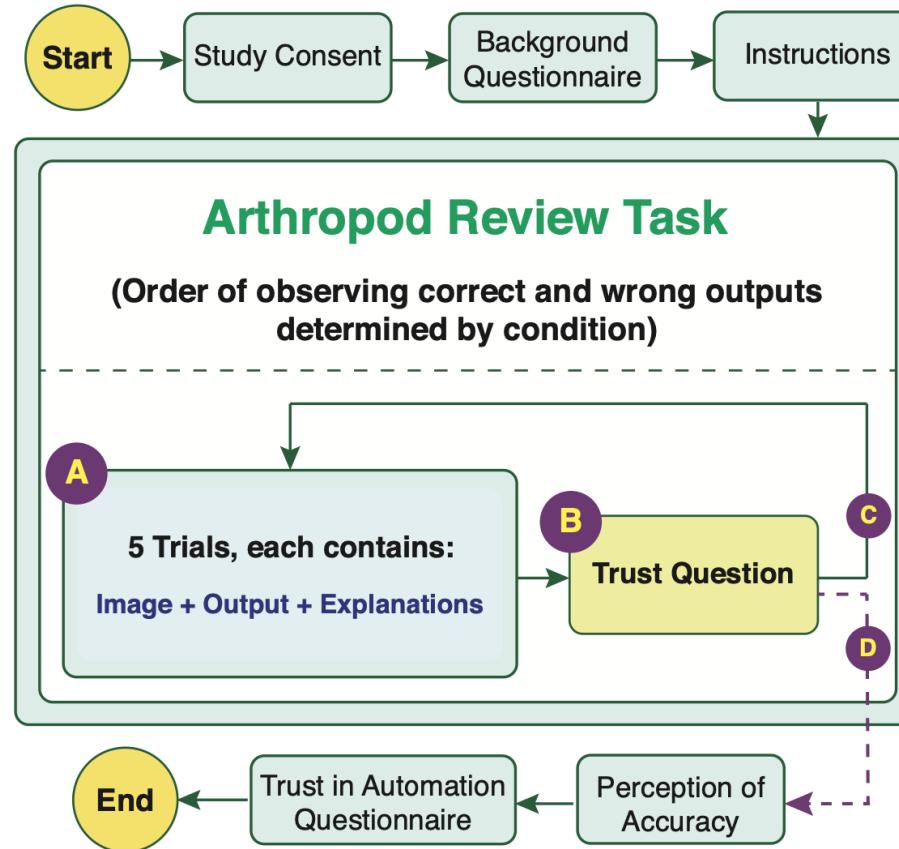


(b) Explanation

(c) Label: Brown argus butterfly

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) Average Trust Rating



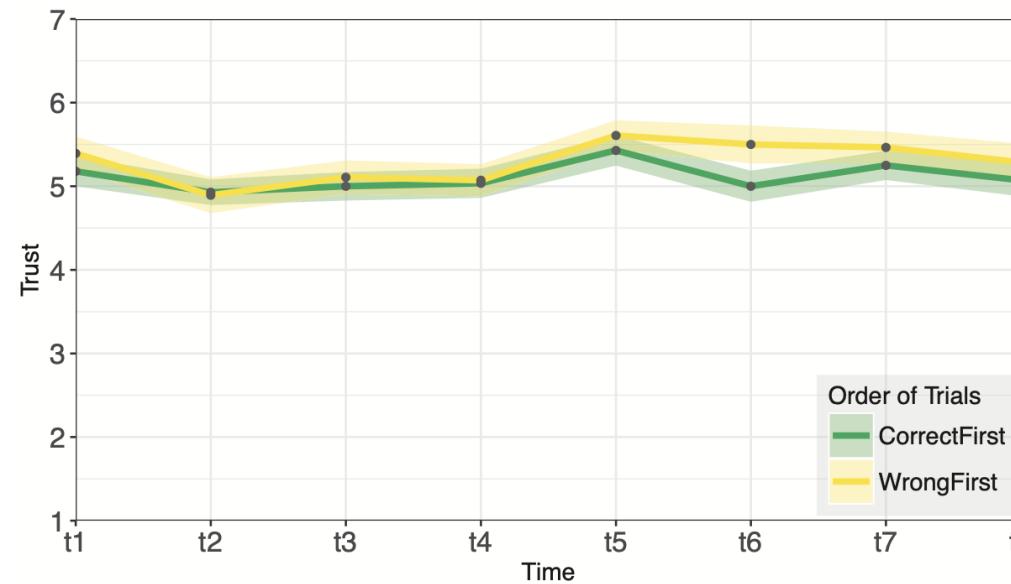
(b) Change of Trust Rating



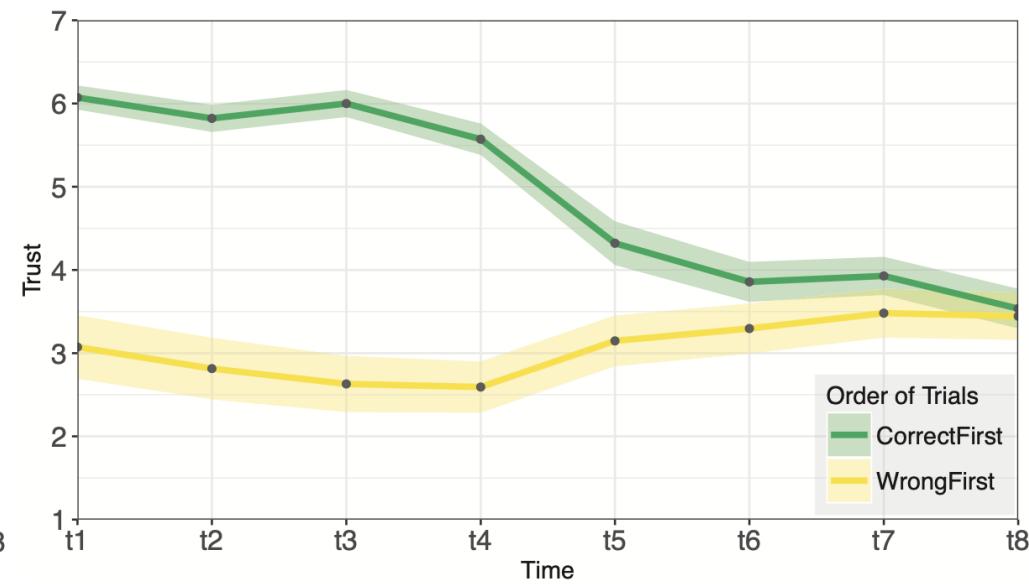
Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

Human Centered Artificial Intelligence

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

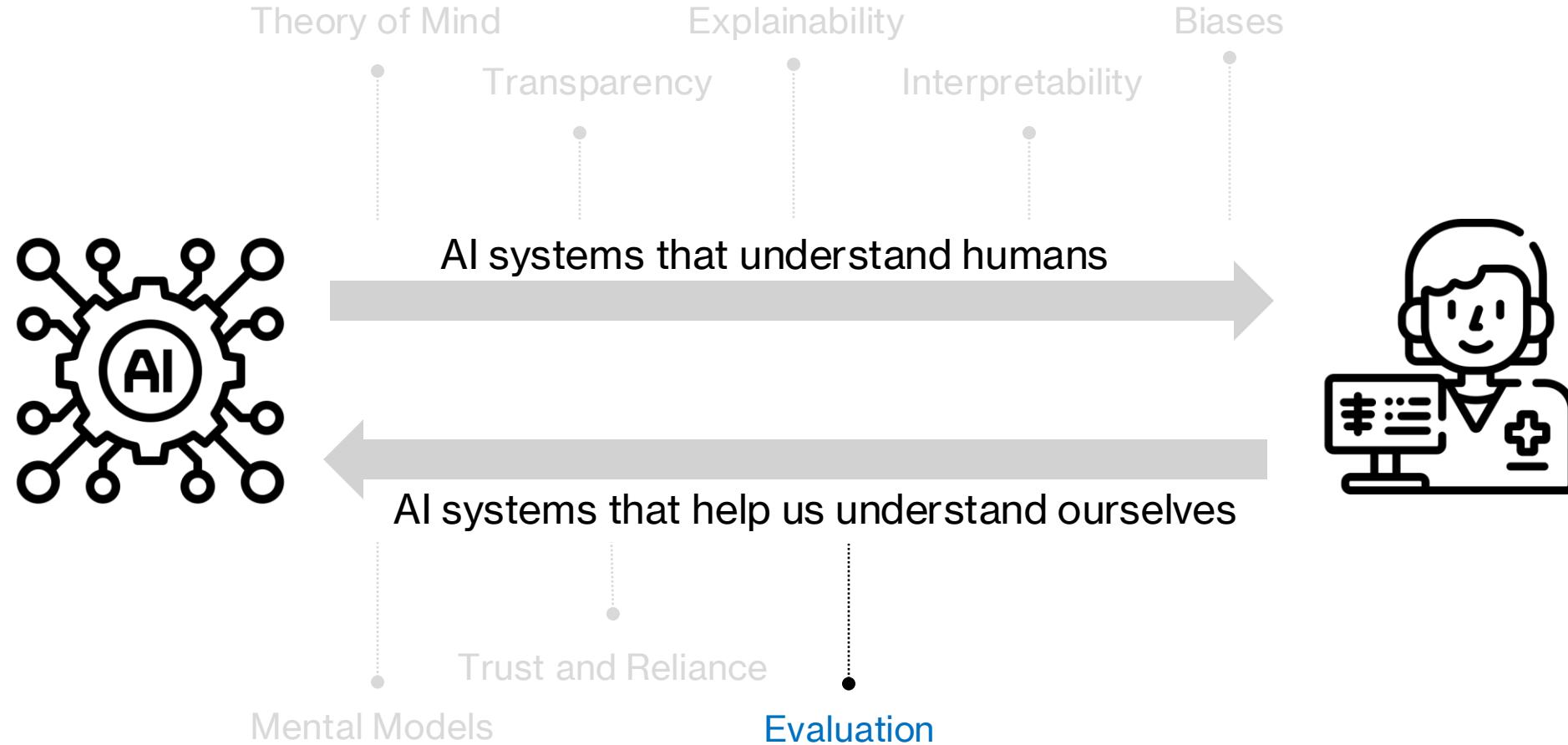


(b) Changes of trust over time for *experienced* participants.

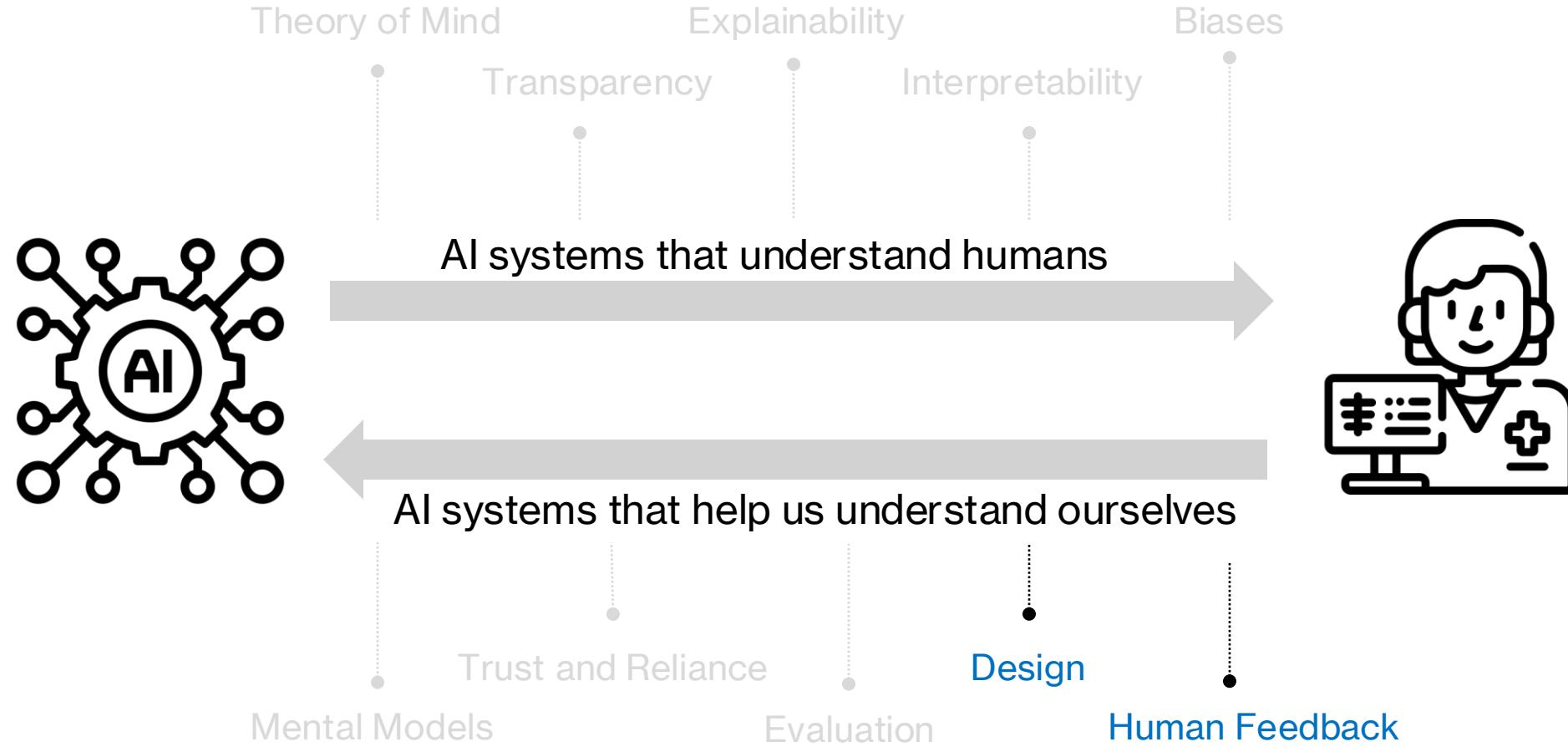


Reference: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems

Human Centered Artificial Intelligence

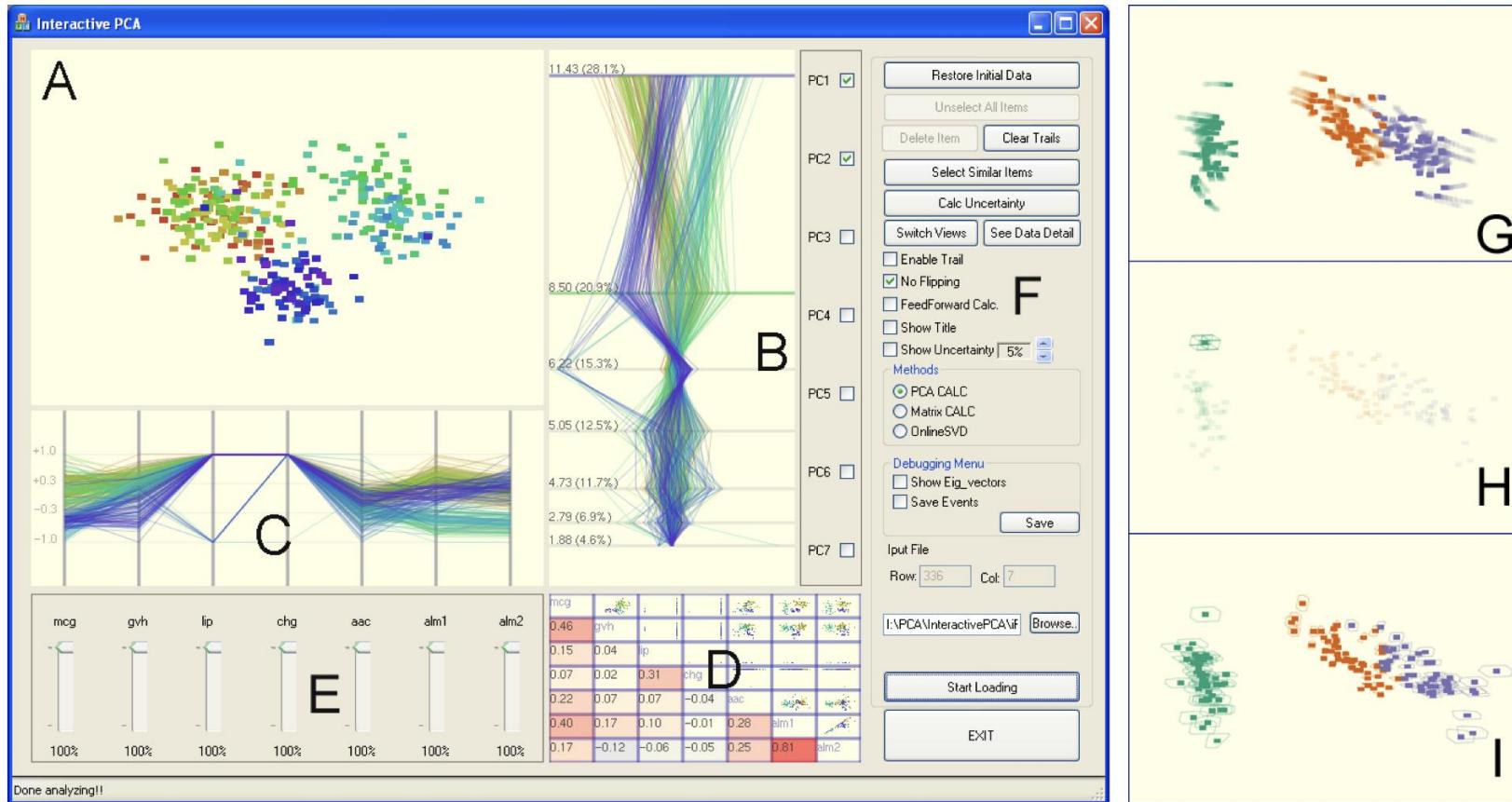


Human Centered Artificial Intelligence

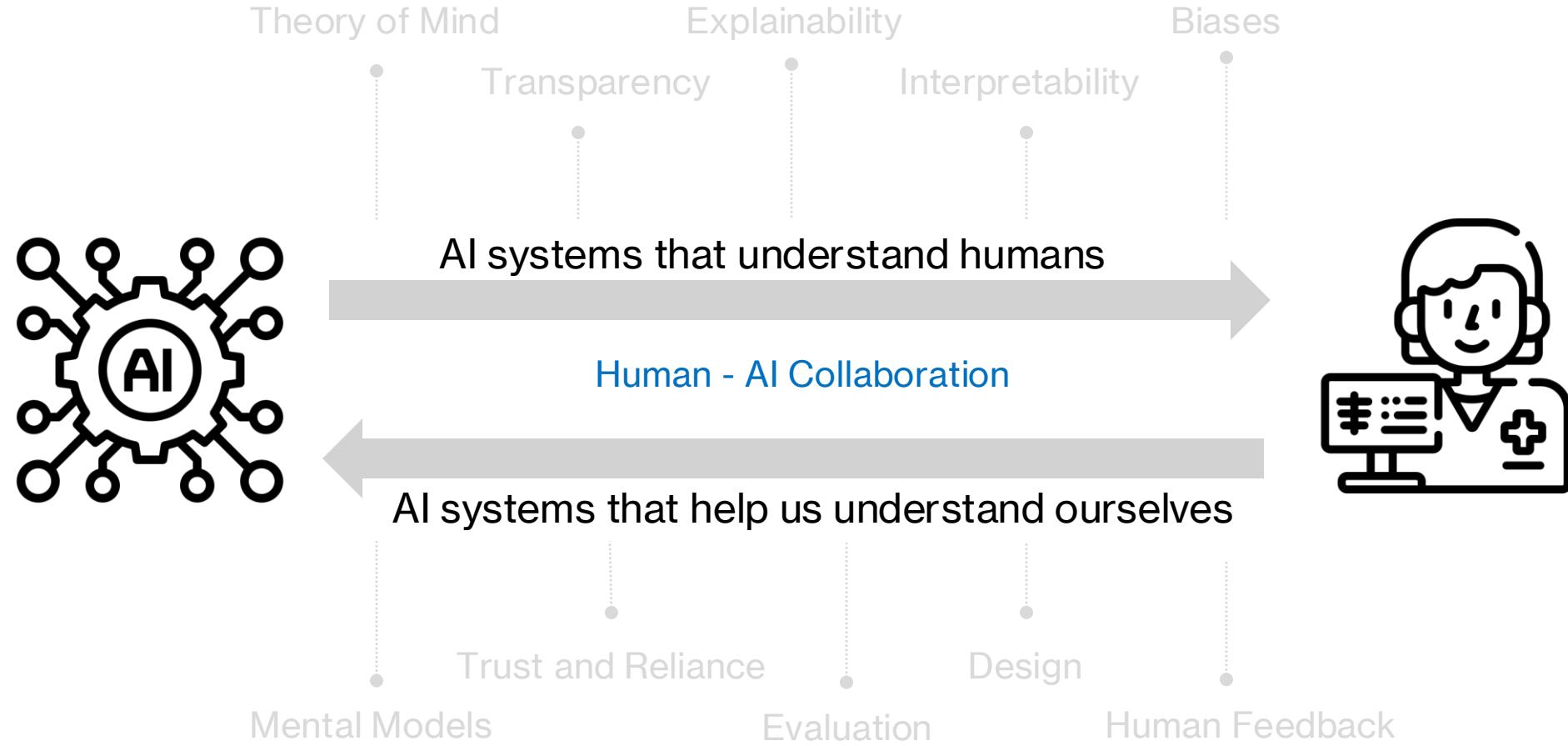


Human Centered Artificial Intelligence

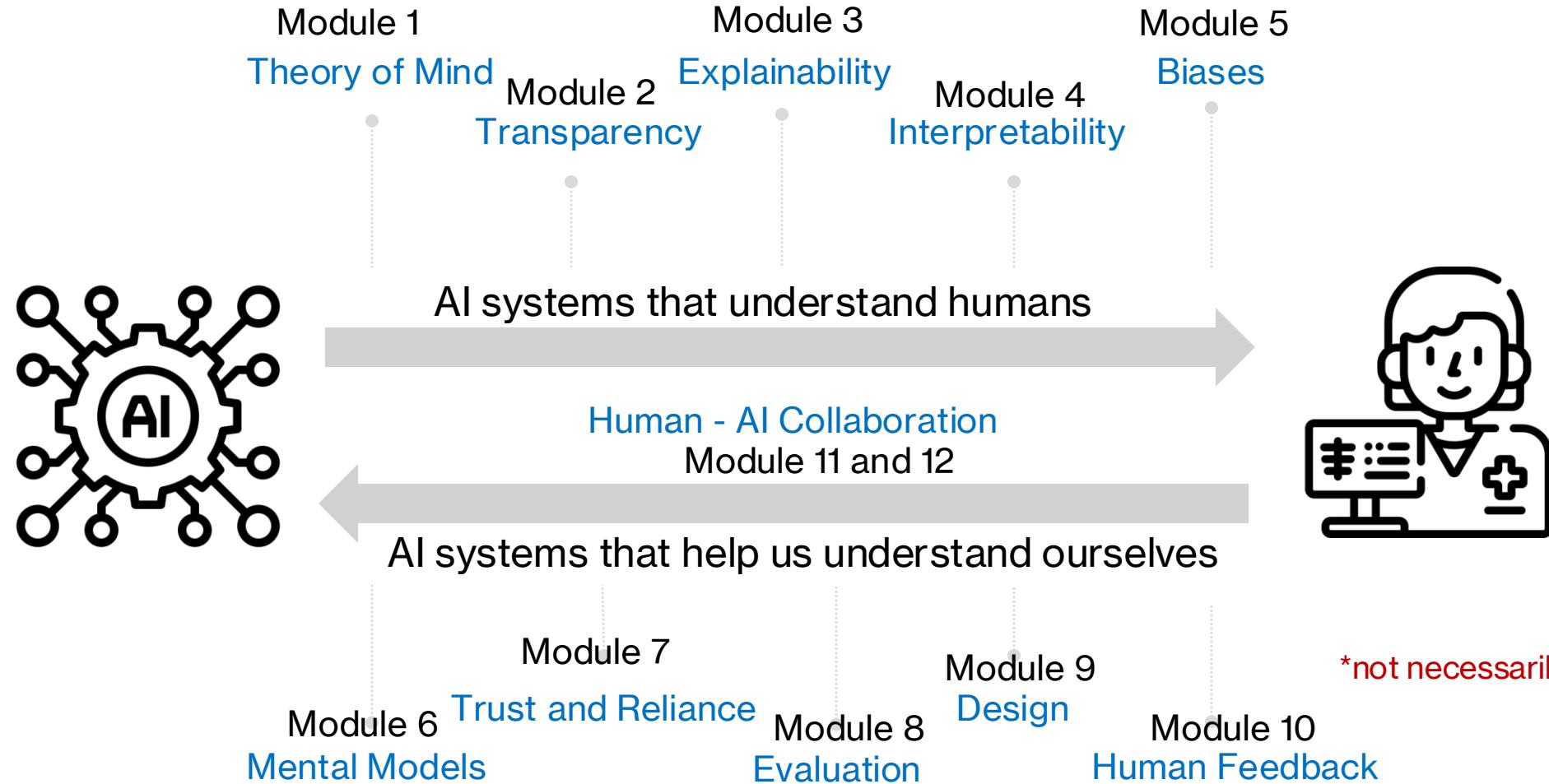
D.H. Jeong et. al / iPCA: An Interactive System for PCA-based Visual Analytics



Human Centered Artificial Intelligence



Human Centered Artificial Intelligence



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

artifical intelligence

Books Groups Quotes People Listopia

Page 1 of about 139 results (0.04 seconds)

 **Artificial Intelligence: A Modern Approach (Pearson Series in Artifical Intelligence)**
by Stuart Russell, Peter Norvig
 4.20 avg rating – 4,327 ratings – published 1994 – 87 editions

 **Tools with Artifical Intelligence; Proceedings; 2v.**
by International Conference on Tools with Artifical Intelligence
 4.00 avg rating – 2 ratings – 1 edition

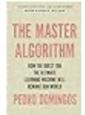
 **The Improbable Machine: What the New Discoveries in Artifical Intelligence Reveal about How The.....**
by Jeremy Campbell
 4.00 avg rating – 25 ratings – published 1989 – 6 editions

machine learning

Books Groups Quotes People Listopia

Page 1 of about 11404 results (0.06 seconds)

 **The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World**
by Pedro Domingos
 3.75 avg rating – 6,146 ratings – published 2015 – 33 editions

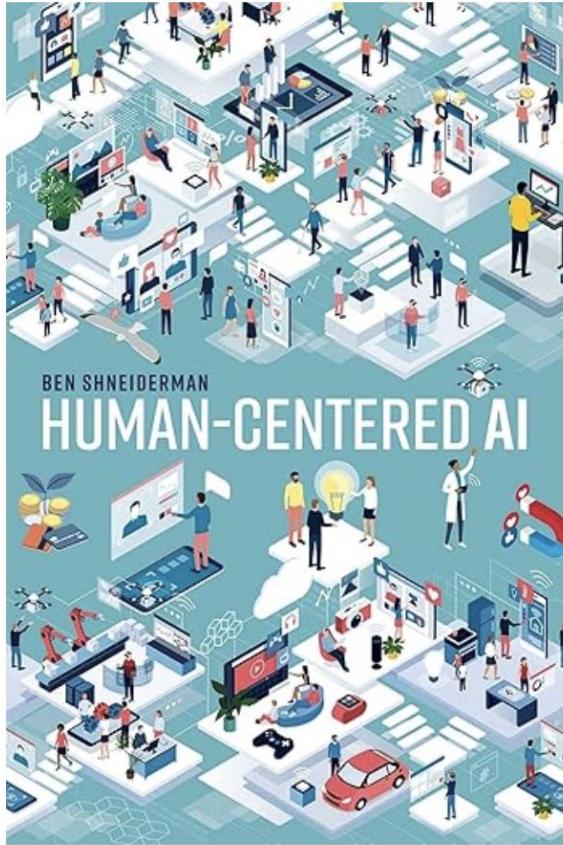
 **The Alignment Problem: Machine Learning and Human Values**
by Brian Christian (Goodreads Author)
 4.38 avg rating – 3,518 ratings – published 2020 – 13 editions

 **Pattern Recognition and Machine Learning**
by Christopher M. Bishop
 4.32 avg rating – 1,840 ratings – published 2006 – 14 editions

Resources

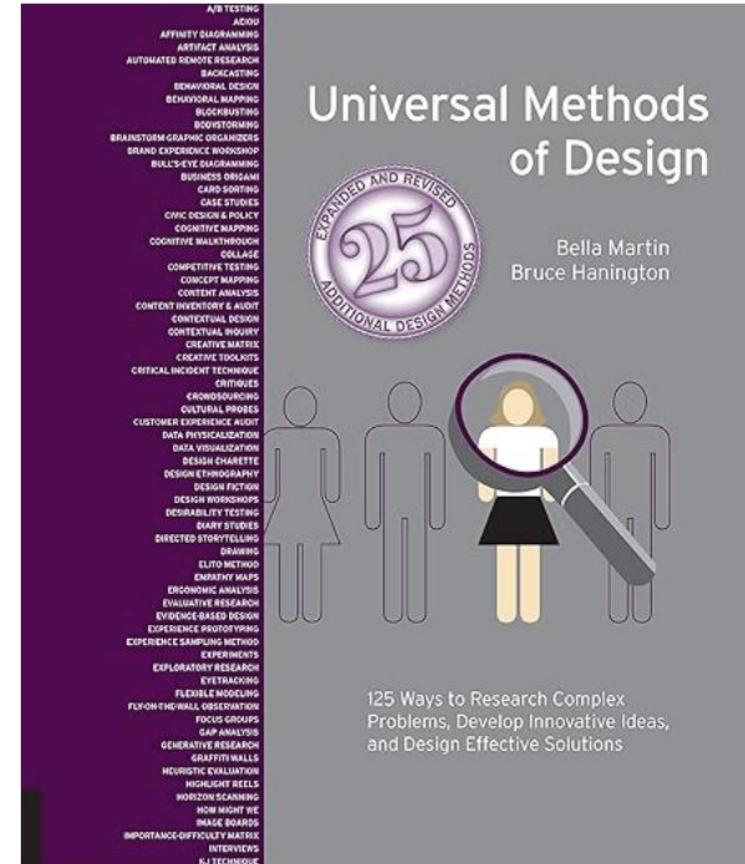


Fairness and Machine Learning

oo

Limitations and Opportunities

Solon Barocas, Moritz Hardt, and Arvind Narayanan



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

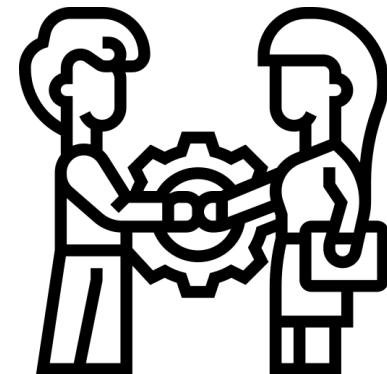
Assignment Presentation Format:

- 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
