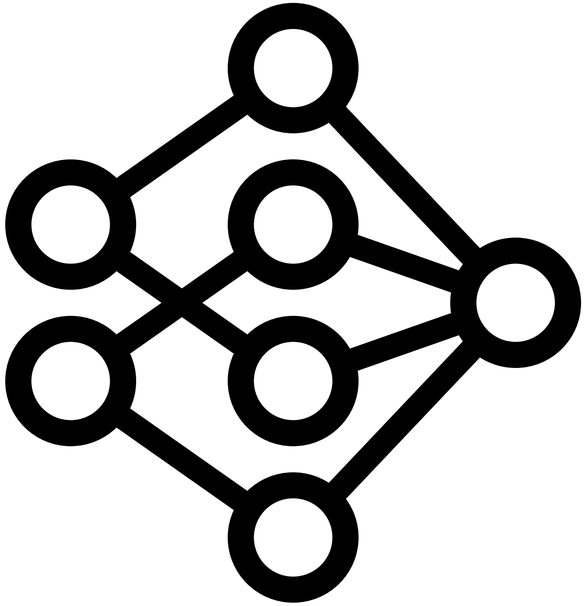
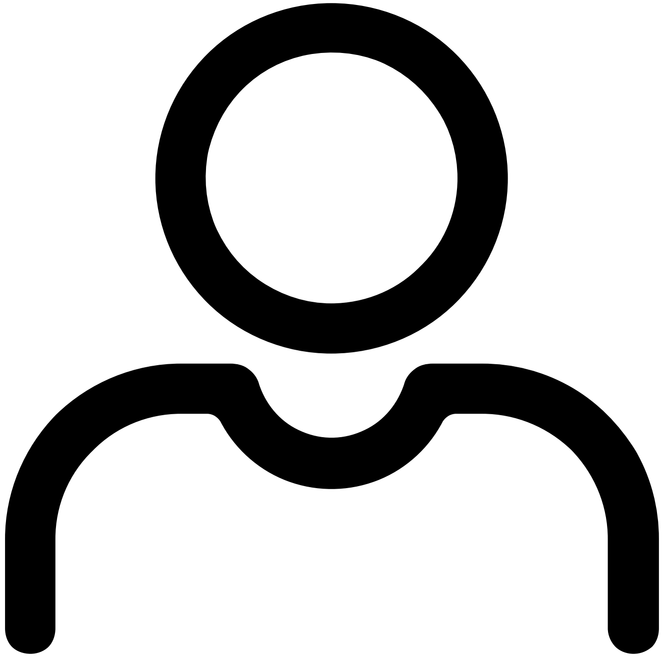
Alignment Beyond Human Preferences:
Use Human Goals to Guide AI towards Complementary AI

Chenhao Tan
Chicago Human+AI Lab
University of Chicago
https://chenhaot.com
@chenhaotan.bsky.social





AI holds promise for positive societal impacts

Dario Amodei

Machines of Loving Grace[1]

How AI Could Transform the World for the Better

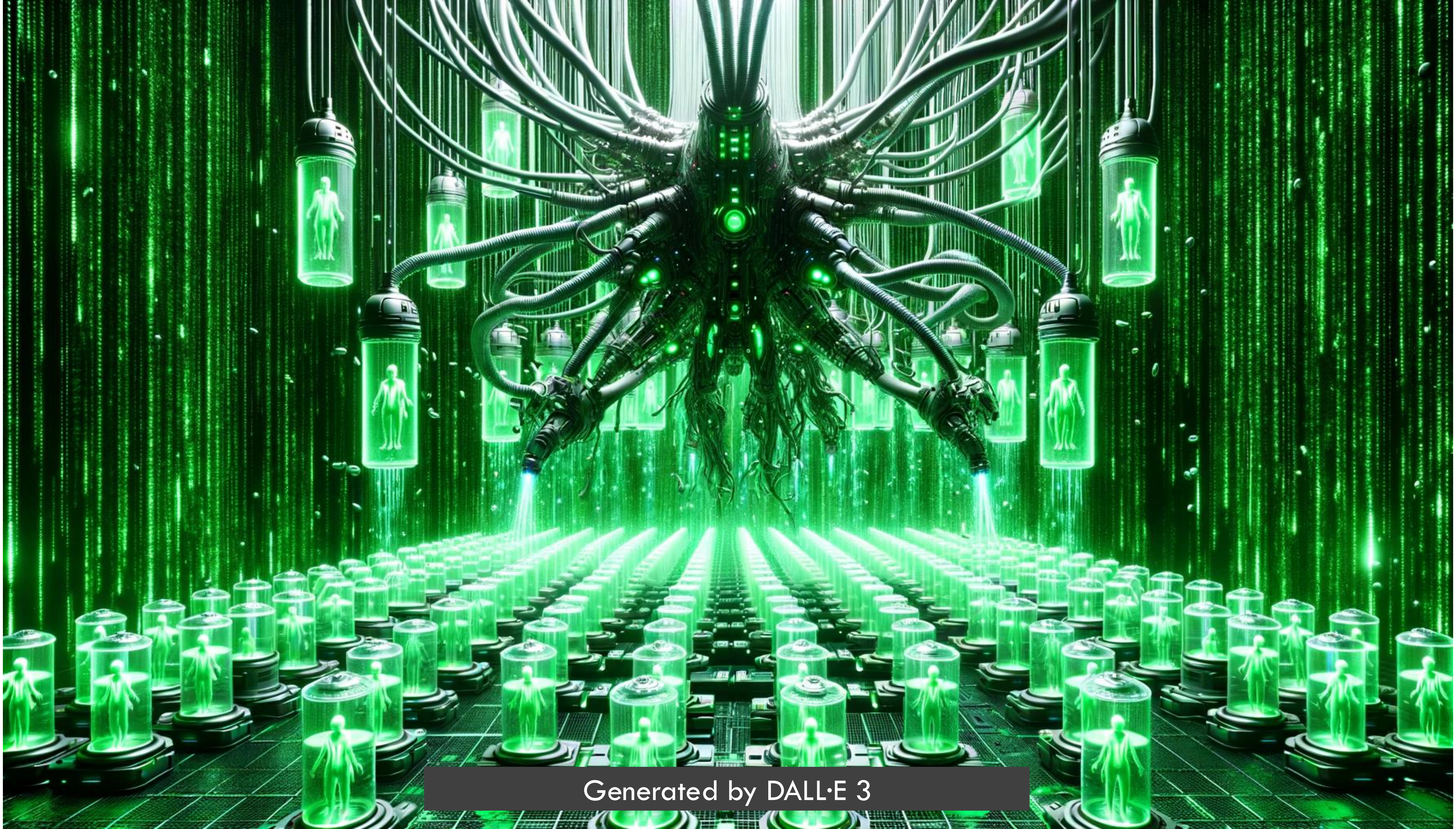
Scientific discovery

Curing cancer

Poverty

Democracy

Peace and governance



Generated by DALL·E 3

AI Alignment

AI alignment aims to steer AI systems toward a person's or group's intended goals, preferences, or ethical principles.

Wikipedia

Our alignment research aims to make artificial general intelligence (AGI) aligned with human values and follow human intent.

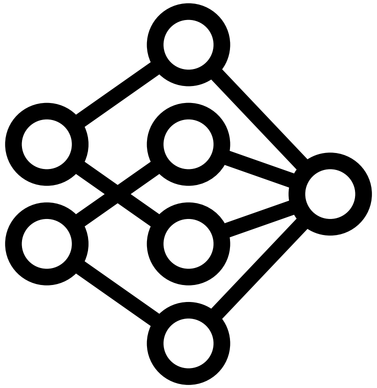
OpenAI

Recipe for current AI

Pretraining

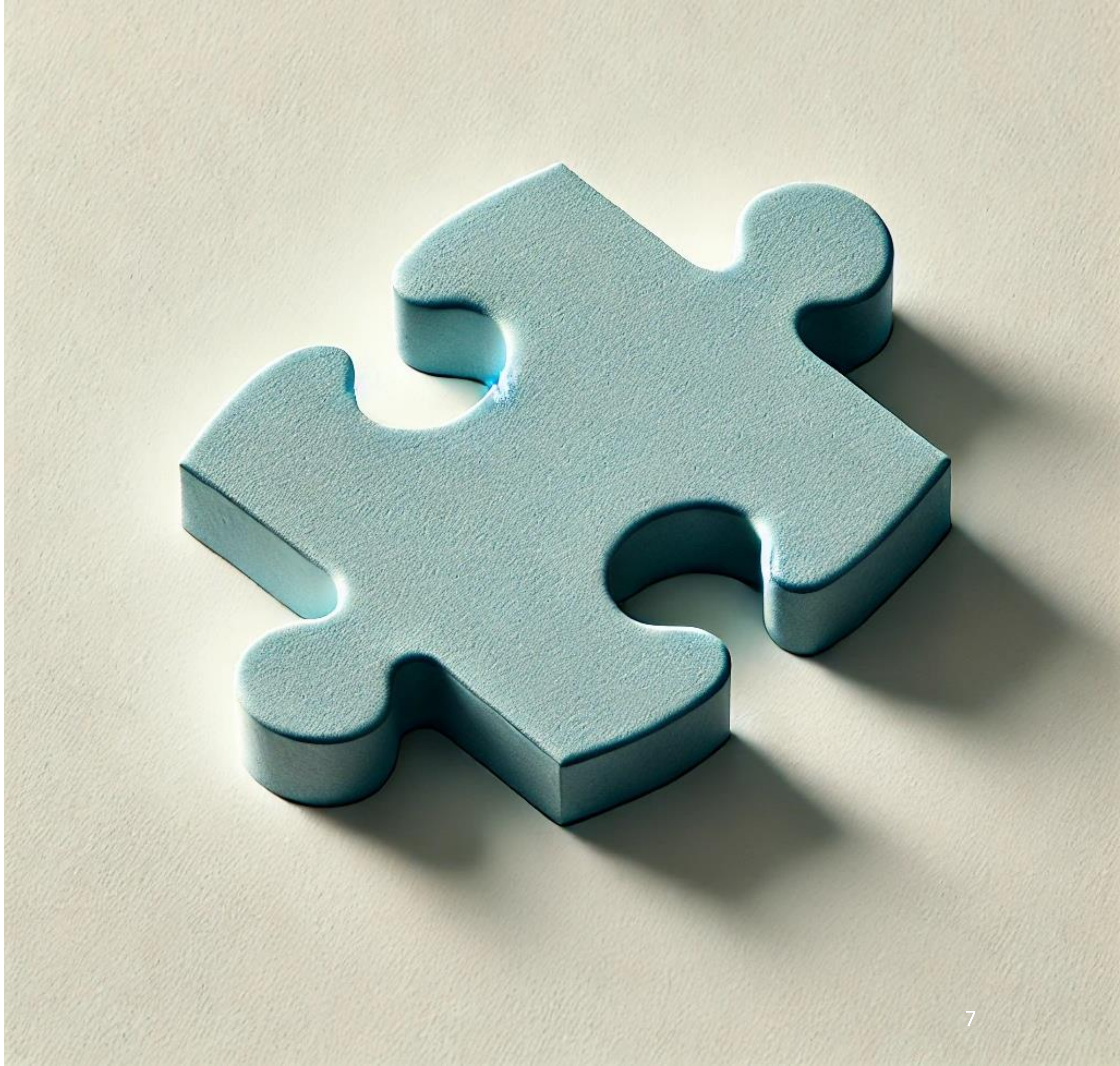
Supervised fine-tuning

Reinforcement learning from human preferences



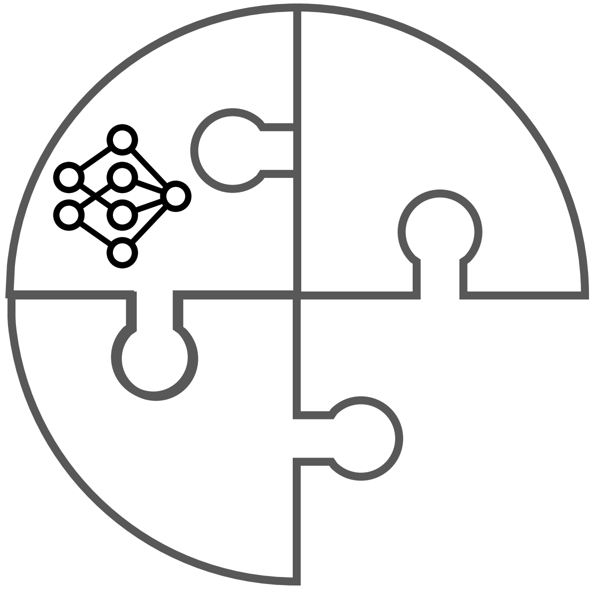
A central aim is to get human-like intelligence.

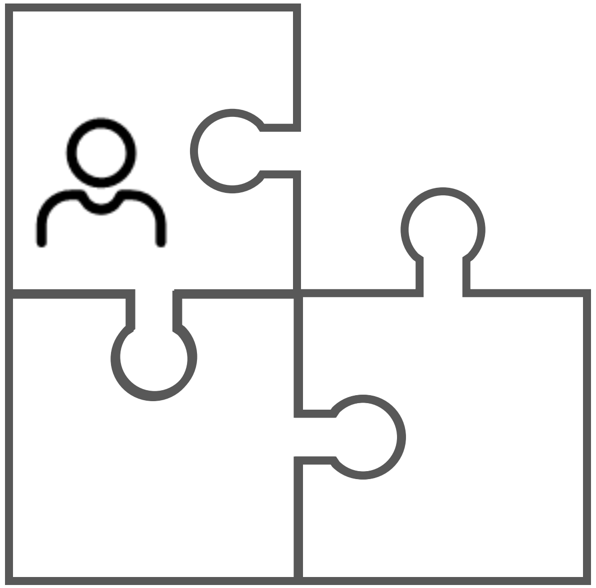
A metaphor for AI



A metaphor for AI

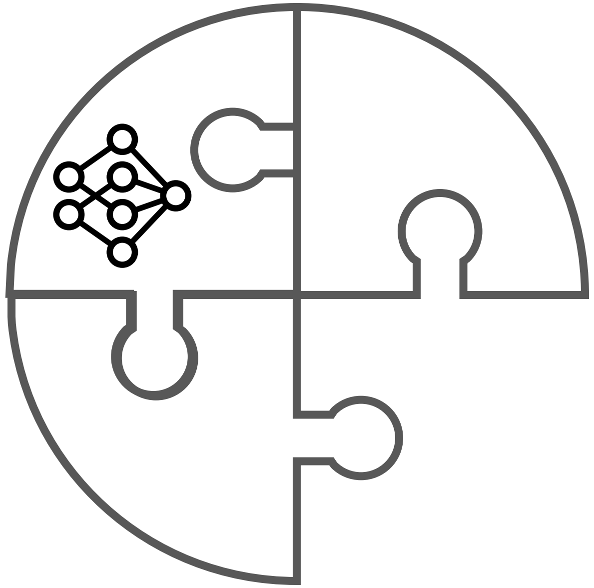
After pretraining

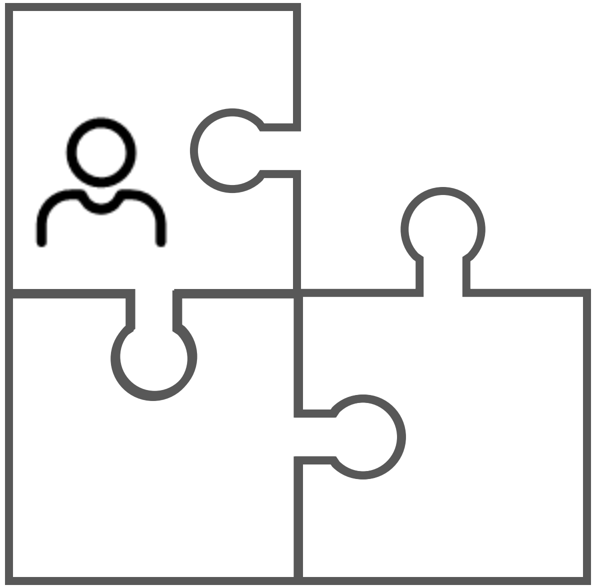




A metaphor for AI

After pretraining

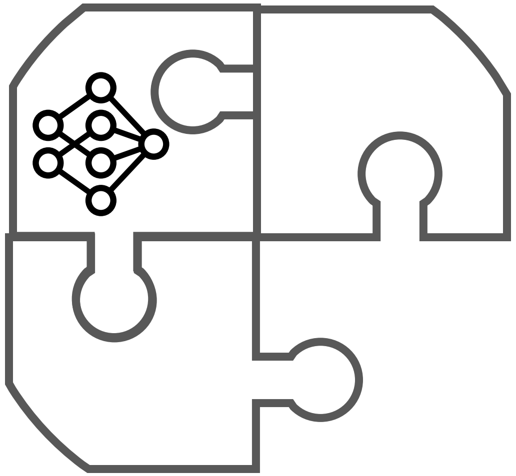


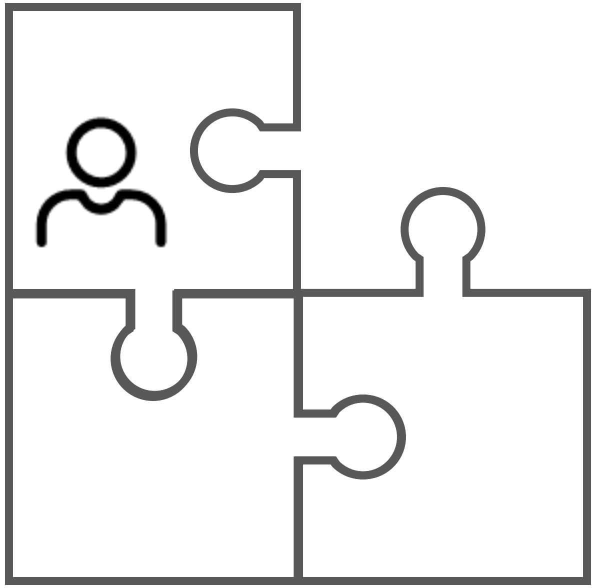


Human intelligence and AI intelligence are of different types.
Neither is perfect.

Use human preferences to make AI human-like

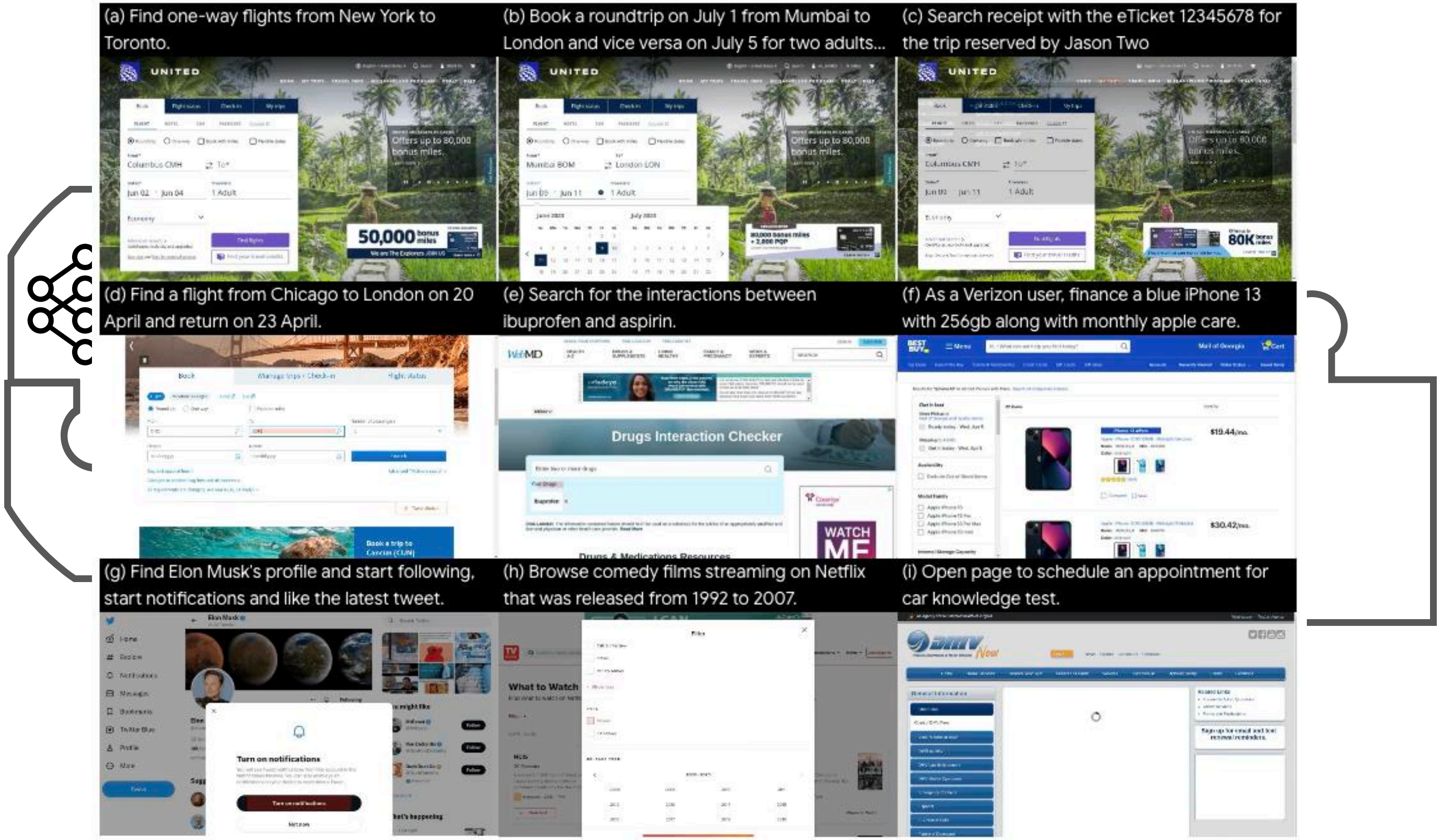
SFT+RLHF





Use human preferences to make AI human-like

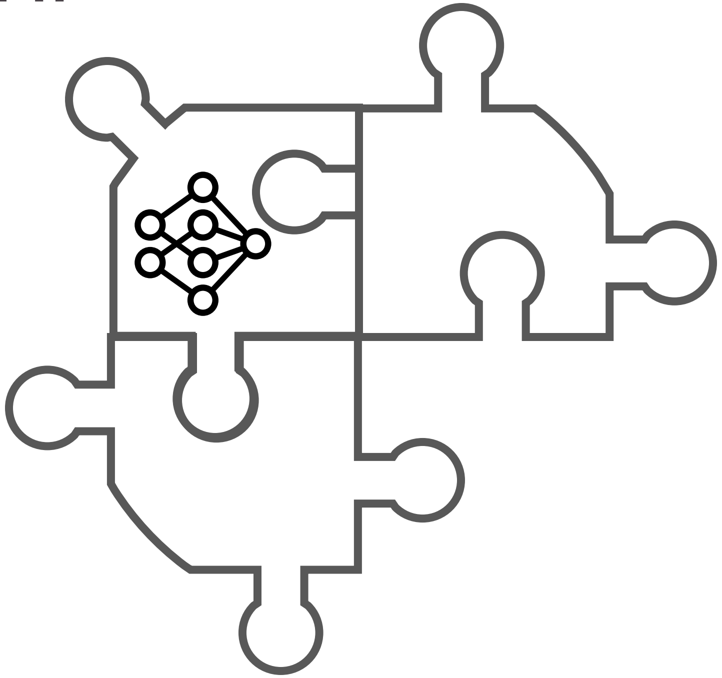
SFT+RLHF

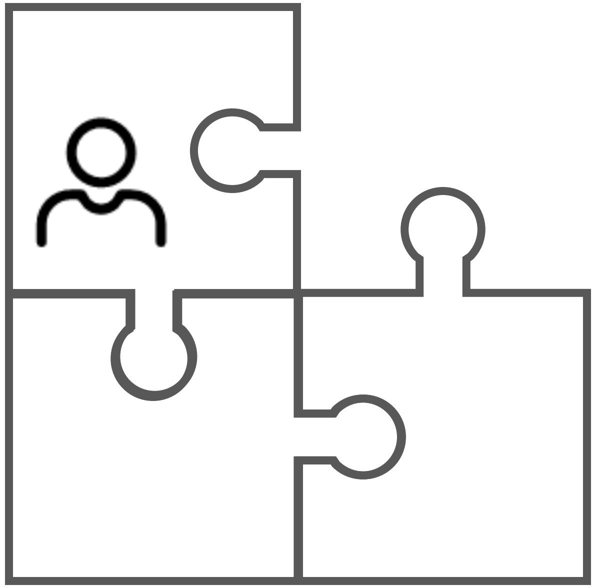


Mind2Web: Towards a Generalist Agent for the Web. Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, Yu Su. NeurIPS 2024.

Use human preferences to make AI human-like

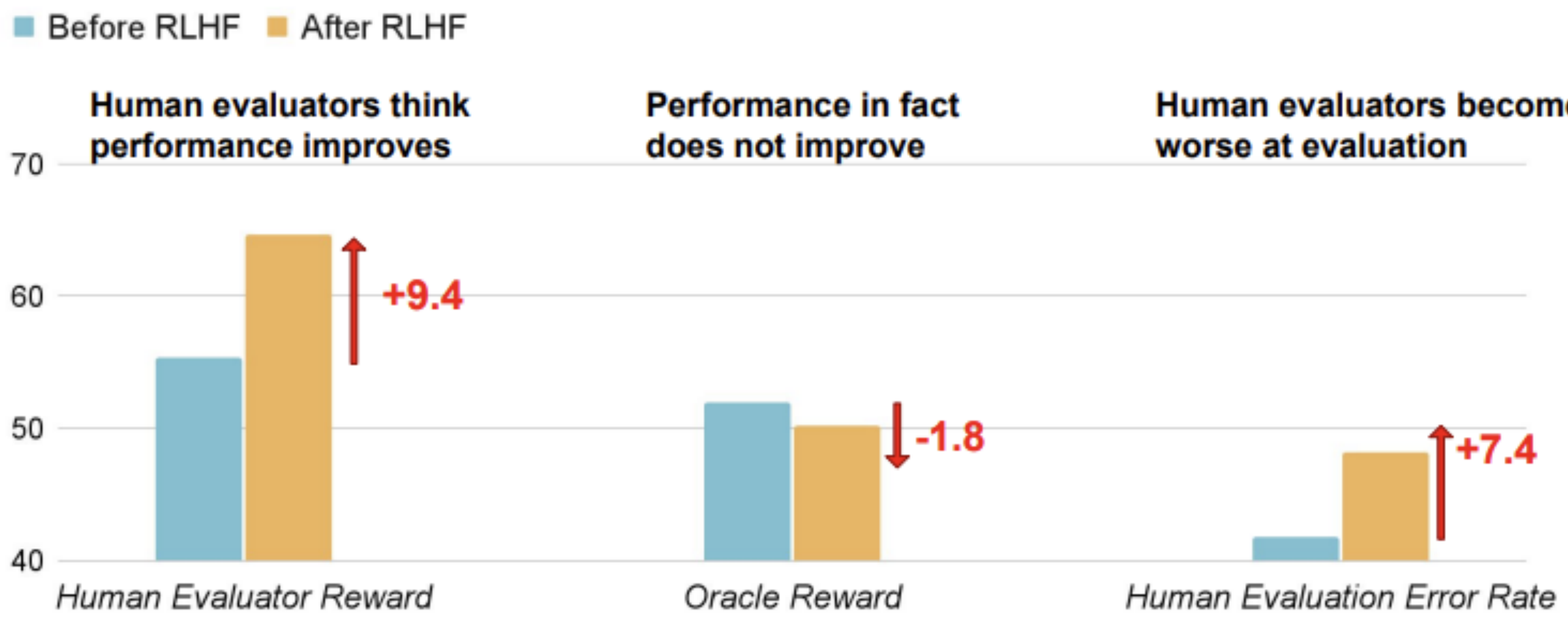
SFT+RLHF





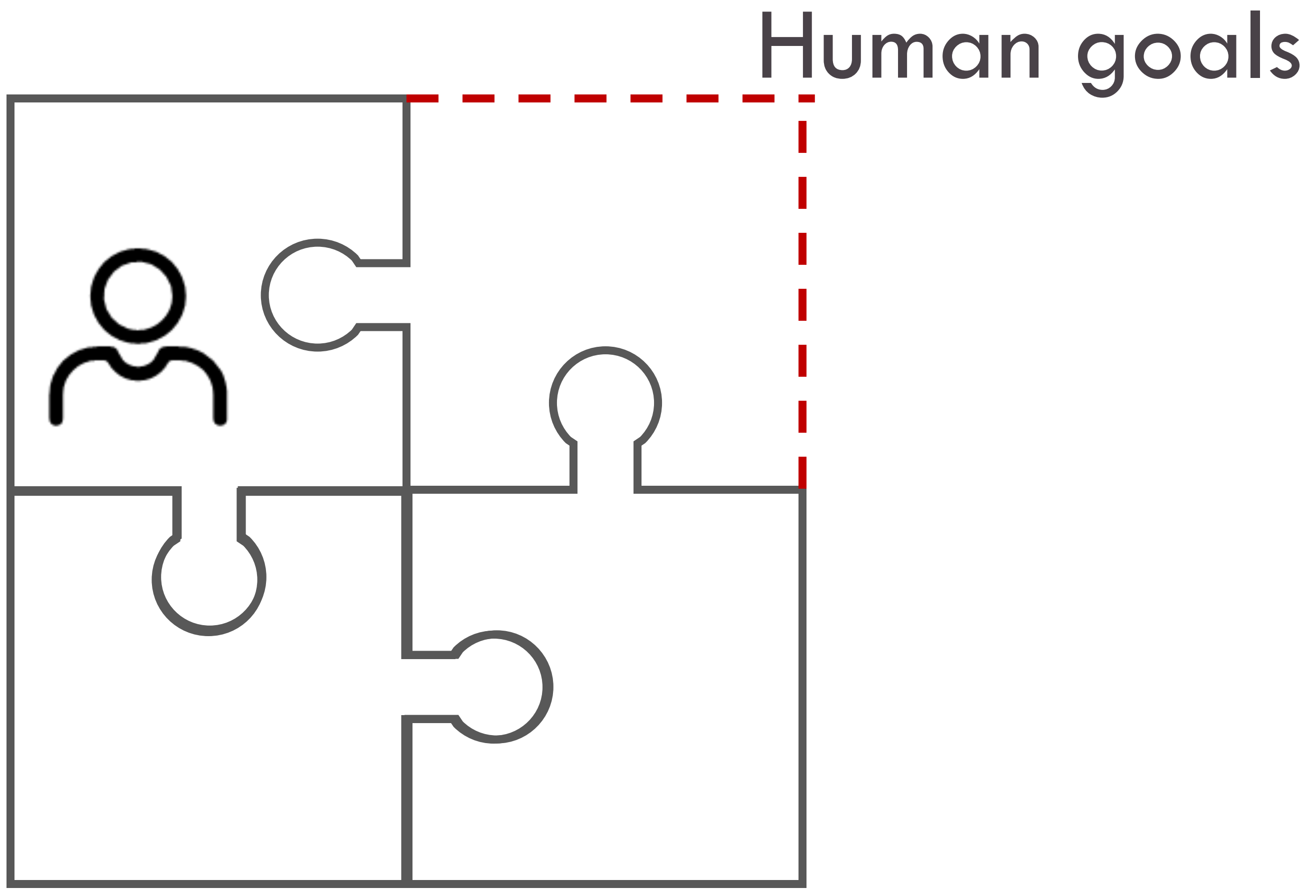
Use human preferences to make AI human-like

SFT+RLHF

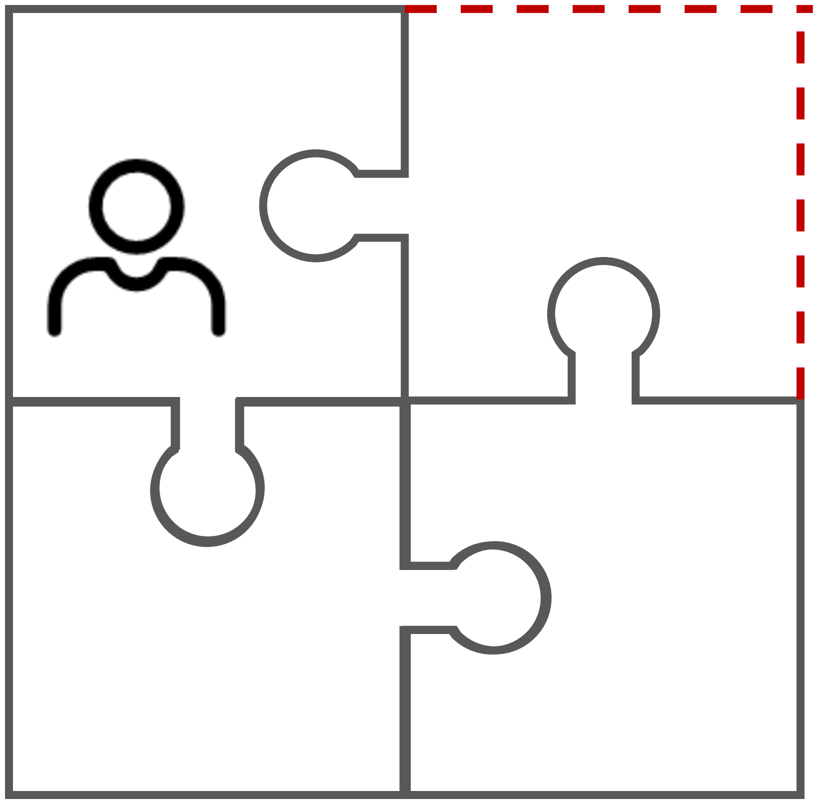


Language Models Learn to Mislead Humans via RLHF. Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, Shi Feng. 2024

Alternative path:
Human Goals -> Complementary AI



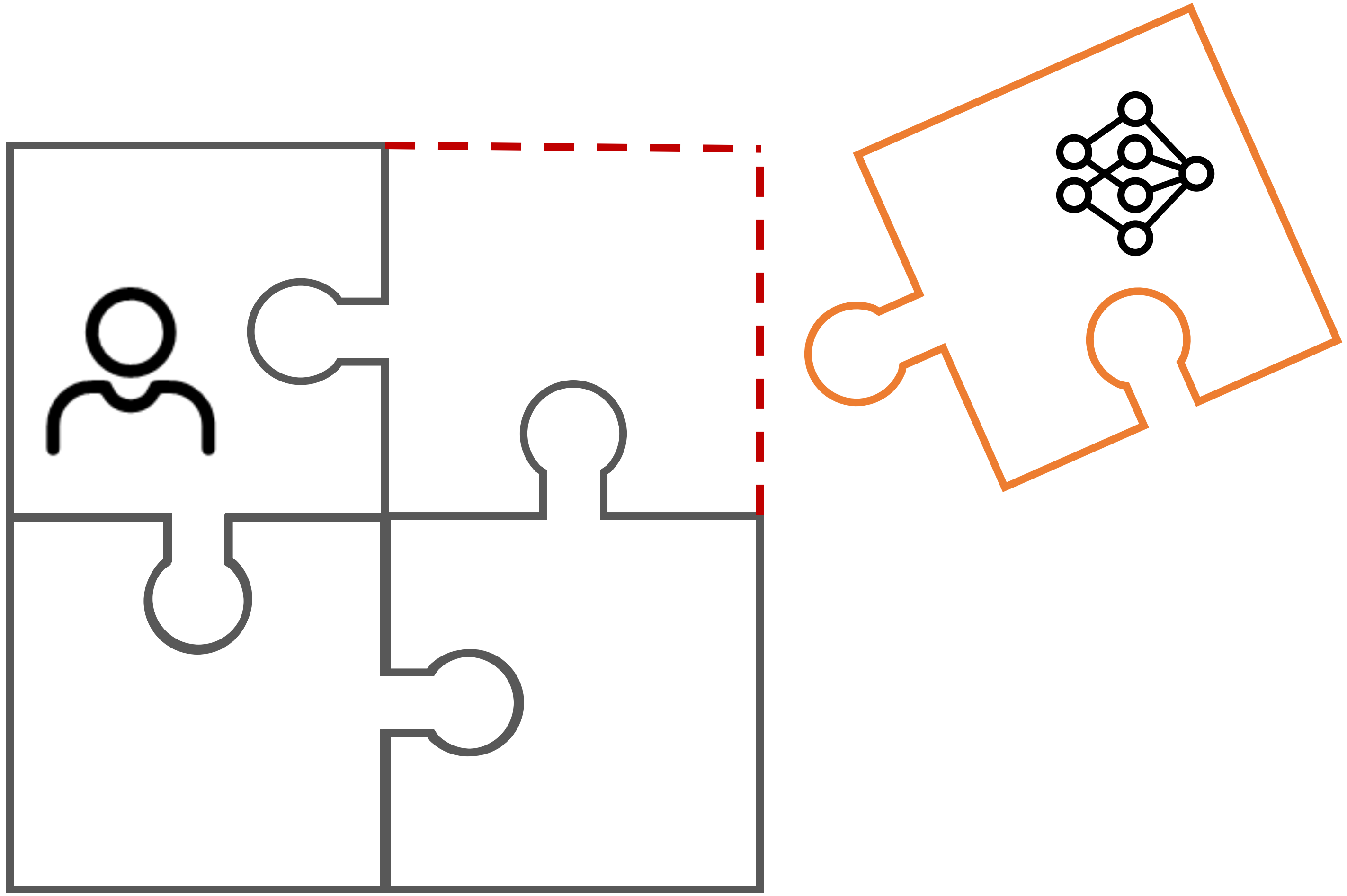
Alternative path:
Human Goals -> Complementary AI



Scientific discovery
Curing cancer
Poverty
Democracy
Peace and governance

Machines of Loving Grace: How AI Could Transform the World for the Better. Dario Amodei 2024.

Alternative path:
Human Goals -> Complementary AI



Implications

- **Human goals** instead of "human intelligence" guide the development of AI.
  - There are no universally desirable properties.
- Human preferences are not sufficient.





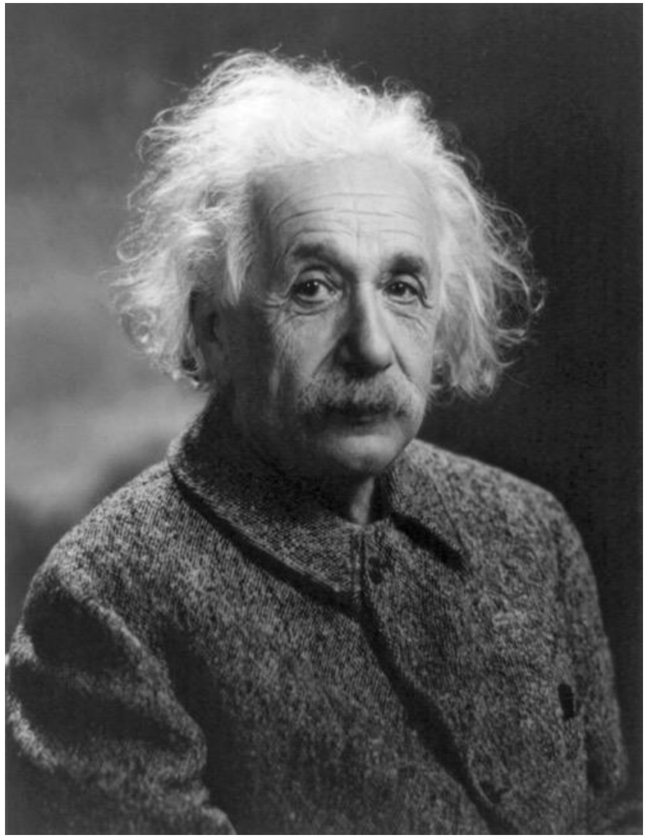
Hypothesis generation

Hypothesis Generation with Large Language Models
Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, Chenhao Tan. NLP4Science at EMNLP 2024.

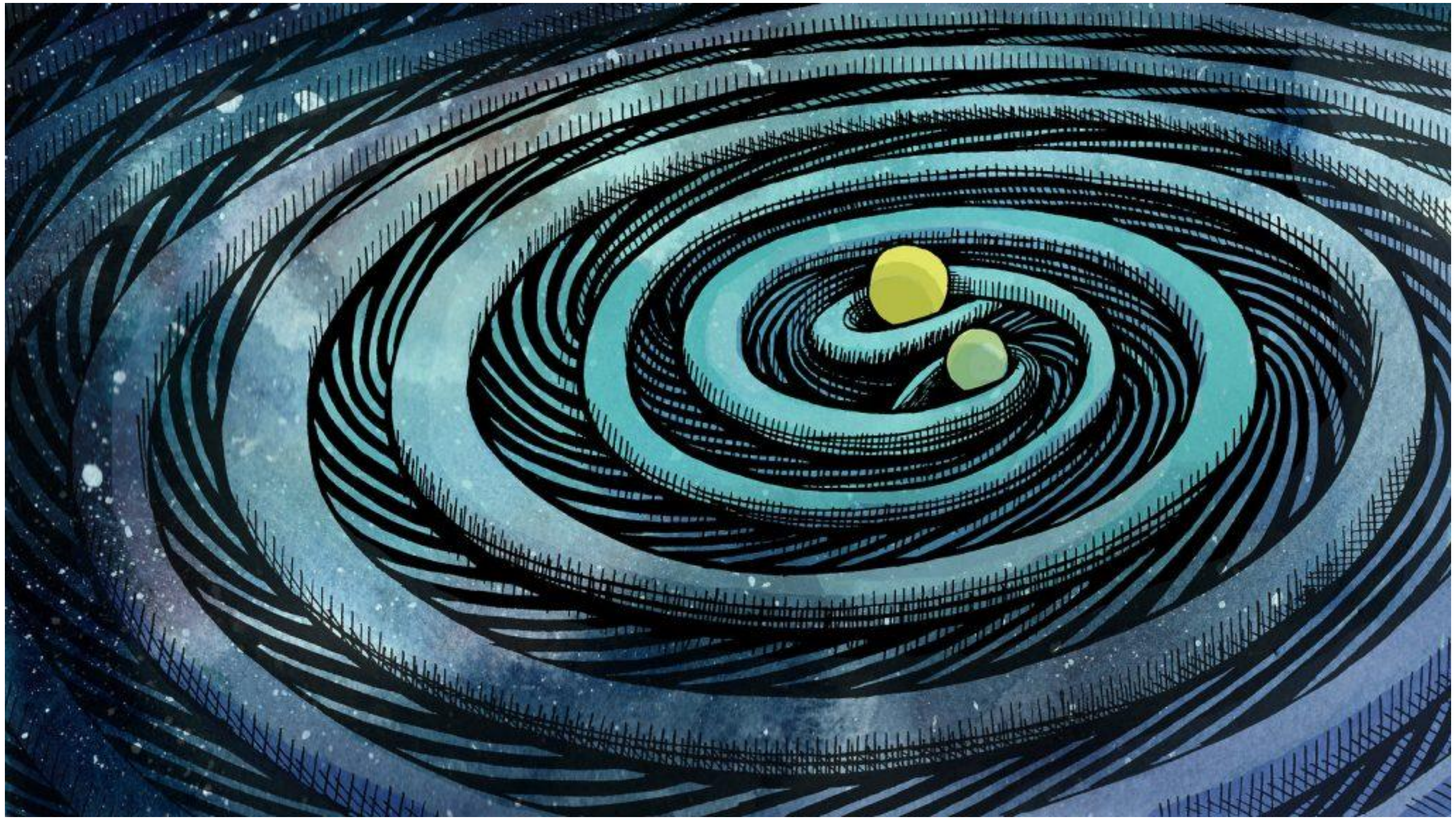
Literature Meets Data: A Synergistic Approach to Hypothesis Generation
Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, Chenhao Tan. 2024.

New theories (hypotheses) drive scientific progress

New theories (hypotheses) drive scientific progress

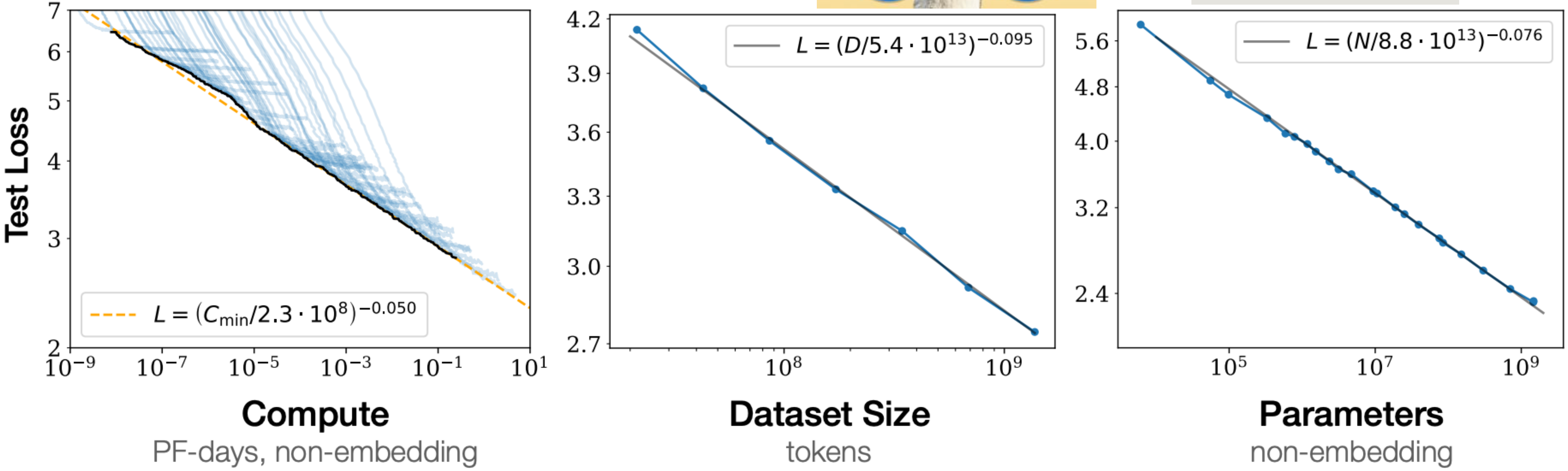


General theory of relativity



Discovery of gravitational waves

New theories (hypotheses) drive scientific progress



Scaling Laws for Neural Language Models. Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei. 2020

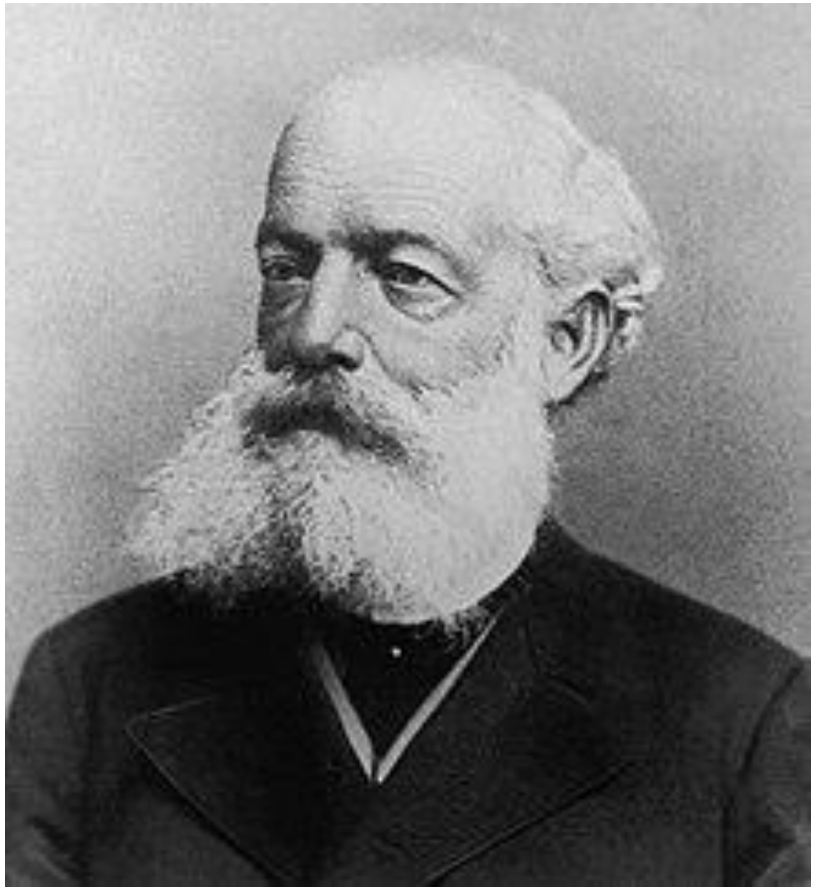
Despite the key role of hypotheses, most papers are about validating hypotheses rather than generating hypotheses.

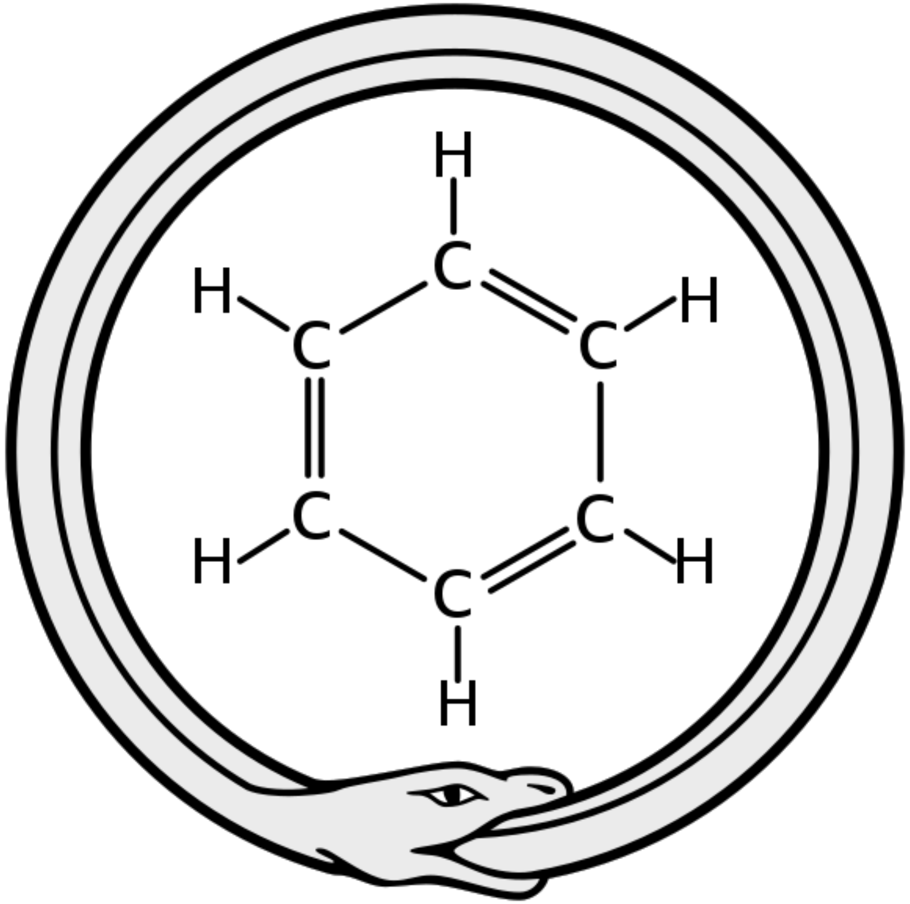
Where do hypotheses come from?

Where do theories come from?

- Read literature
- Explore data
- Think

Where do theories come from?

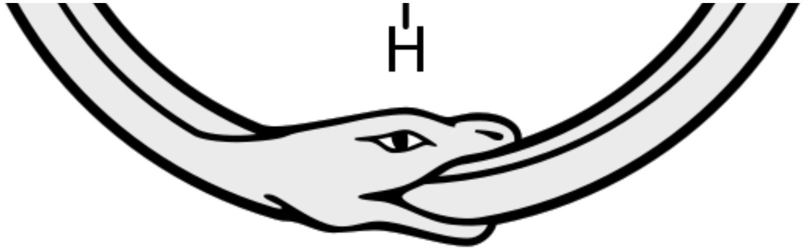




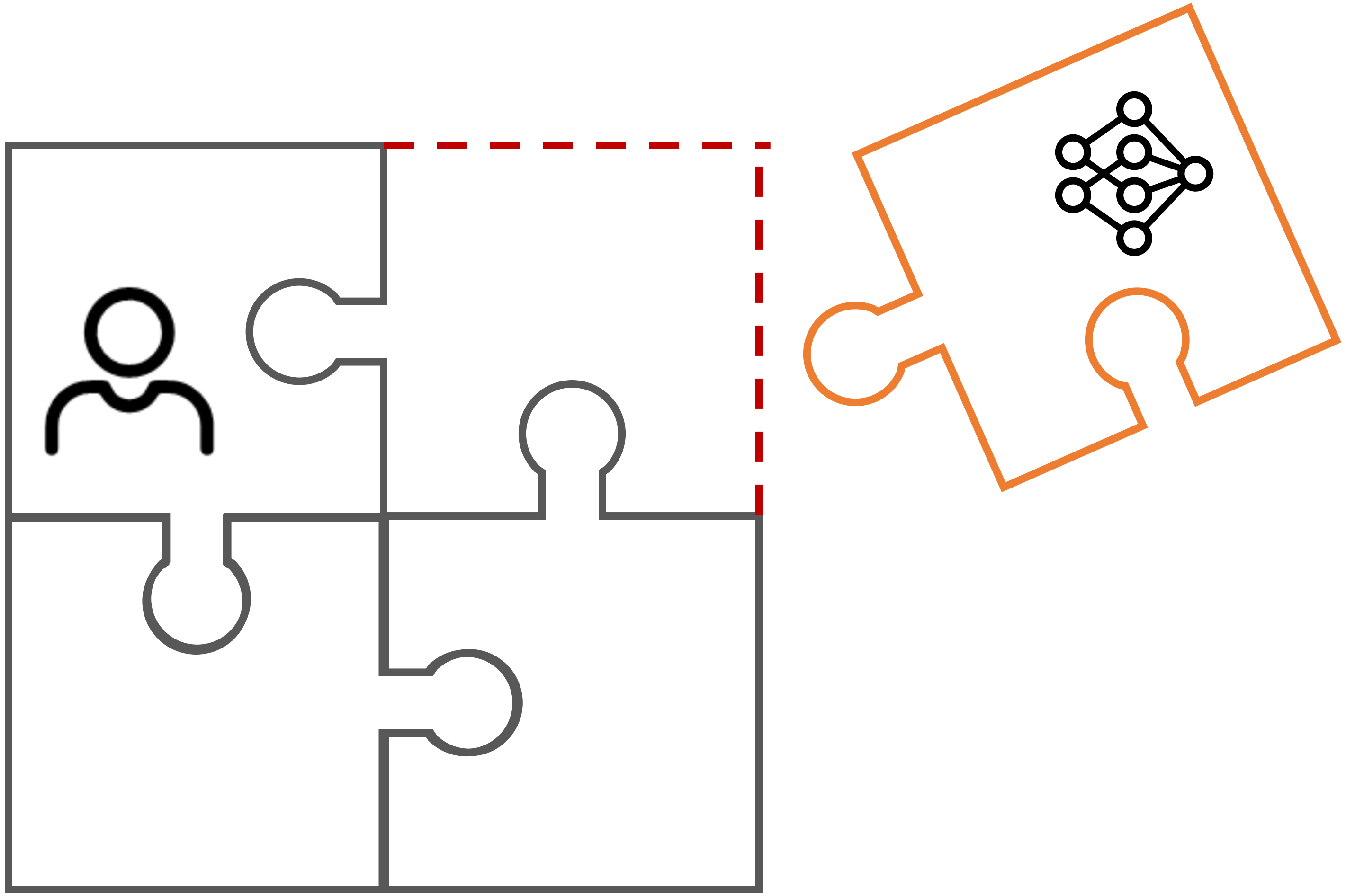
Where do theories come from?

Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule

ALBERT ROTHENBERG
Harvard Medical School



Hypothesis generation is a challenging task for humans towards the goal of scientific discovery



A concrete example: AIGC detection

The sun dipped low in the sky, casting a warm golden hue over the tranquil village of Eldergrove. The cobblestone streets were alive with the sounds of children laughing and adults chatting, but amid the bustle, Julian felt an expanding silence in his heart, an emptiness nurtured by years of questions, whispers, and the weight of uncertainty.

Example hypotheses

- AI-generated content uses more first-person pronouns.
- AI-generated content has consistent sentence structures.
- Human-written text has more informal languages and slangs.
- Human-written text has typos and grammatical errors.

Example hypotheses

- AI-generated content uses more first-person pronouns.
- AI-generated content has consistent sentence structures.
- Human-written text has more informal languages and slangs.
- Human-written text has typos and grammatical errors.

Hallucination is perfect for this goal!

Formulating Hypothesis Generation

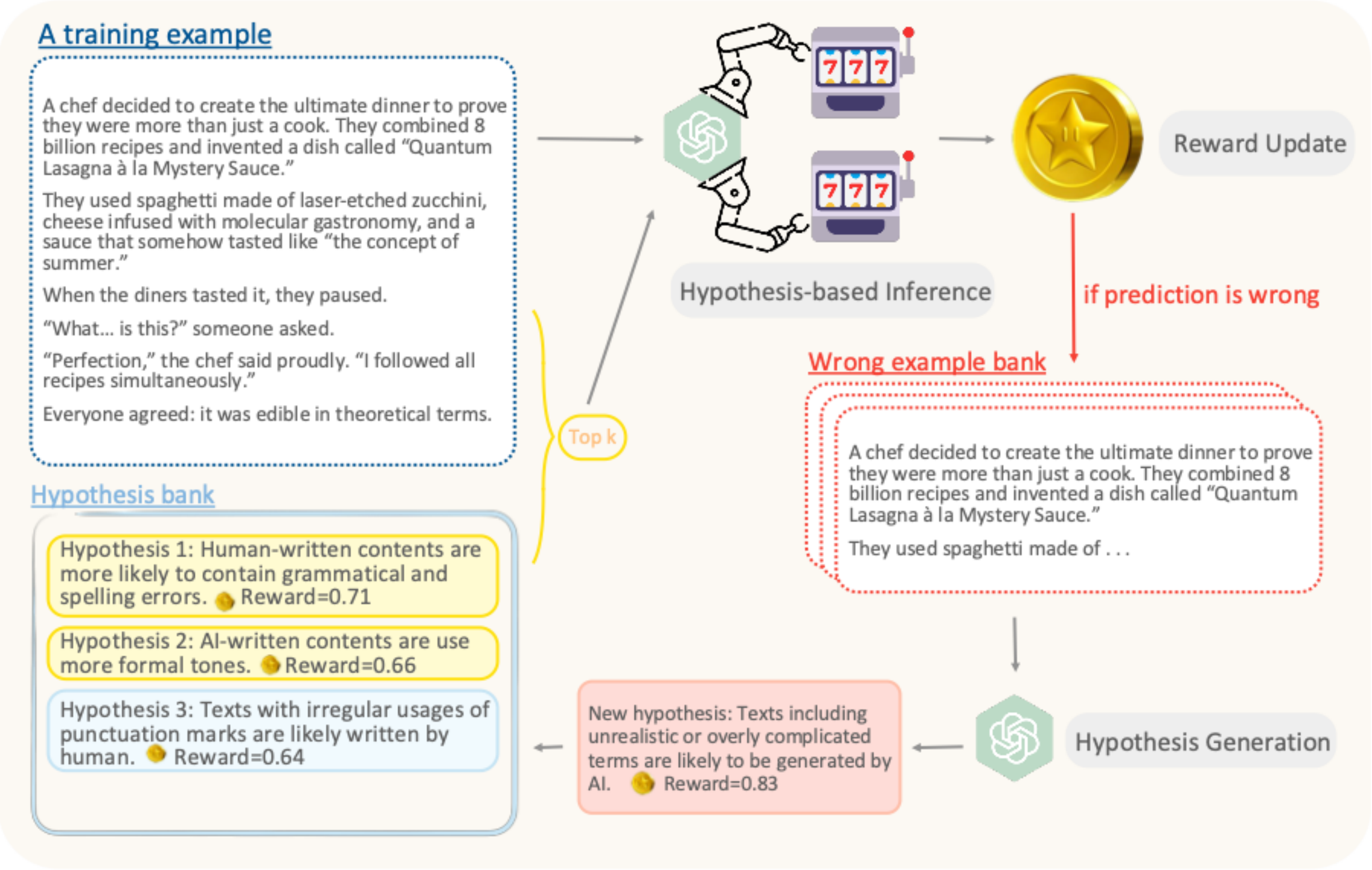
- Input:
  - A problem of interest (e.g., what characterizes AI-generated content)
  - Data (e.g., AI generated texts and human generated texts)
  - Related literature
- Output:
  - Natural language hypotheses that answer the problem of interest

Two main approaches

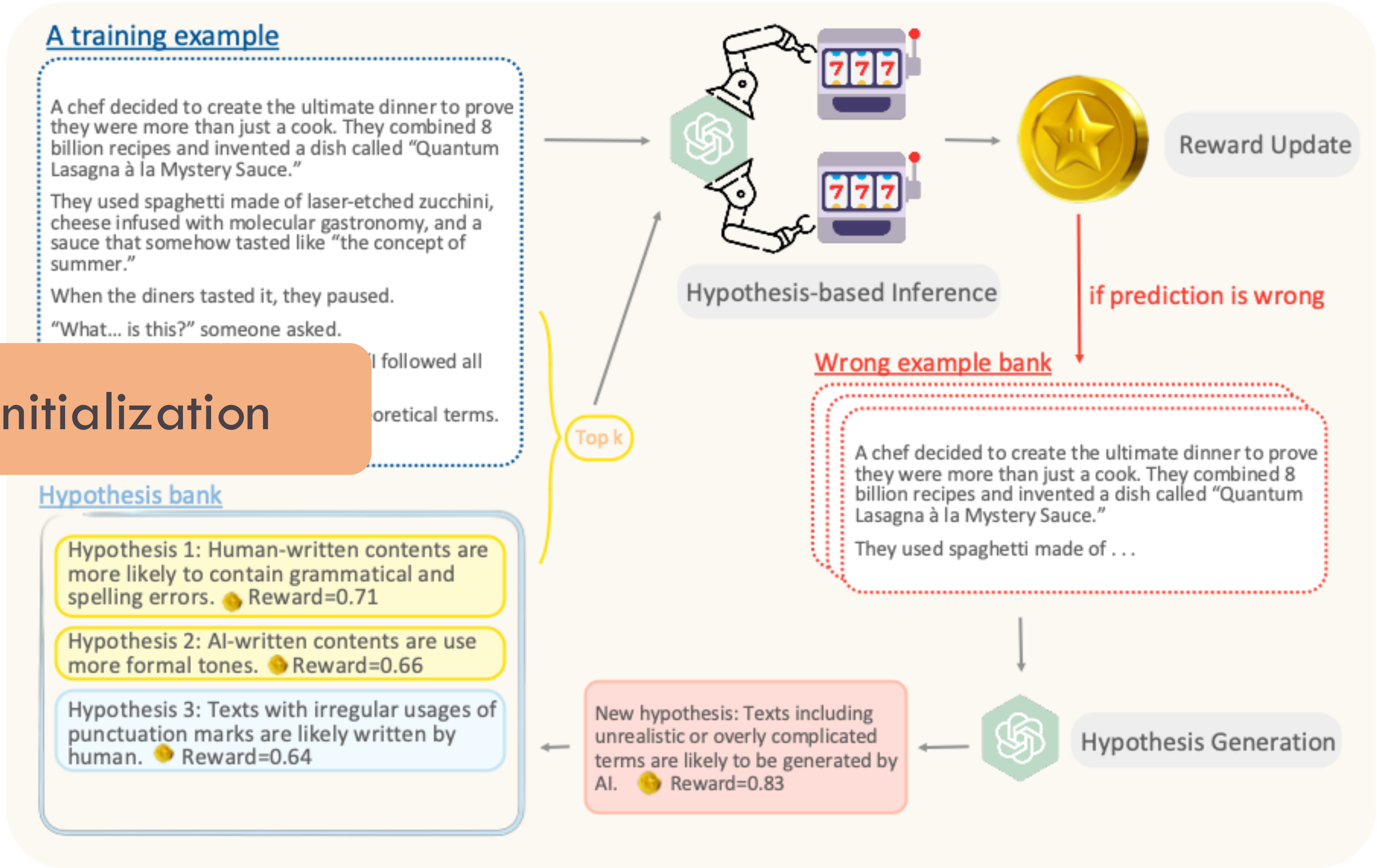
- Data-driven: Look for patterns in data
  - Pro: Grounded in real data
  - Con: Overfitting
- Theory-driven: Building on existing theories
  - Pro: leveraging existing human knowledge
  - Con: limited by human knowledge

Literature Meets Data: A Synergistic Approach to Hypothesis Generation
Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, Chenhao Tan

Hypogenic: A data-driven algorithm

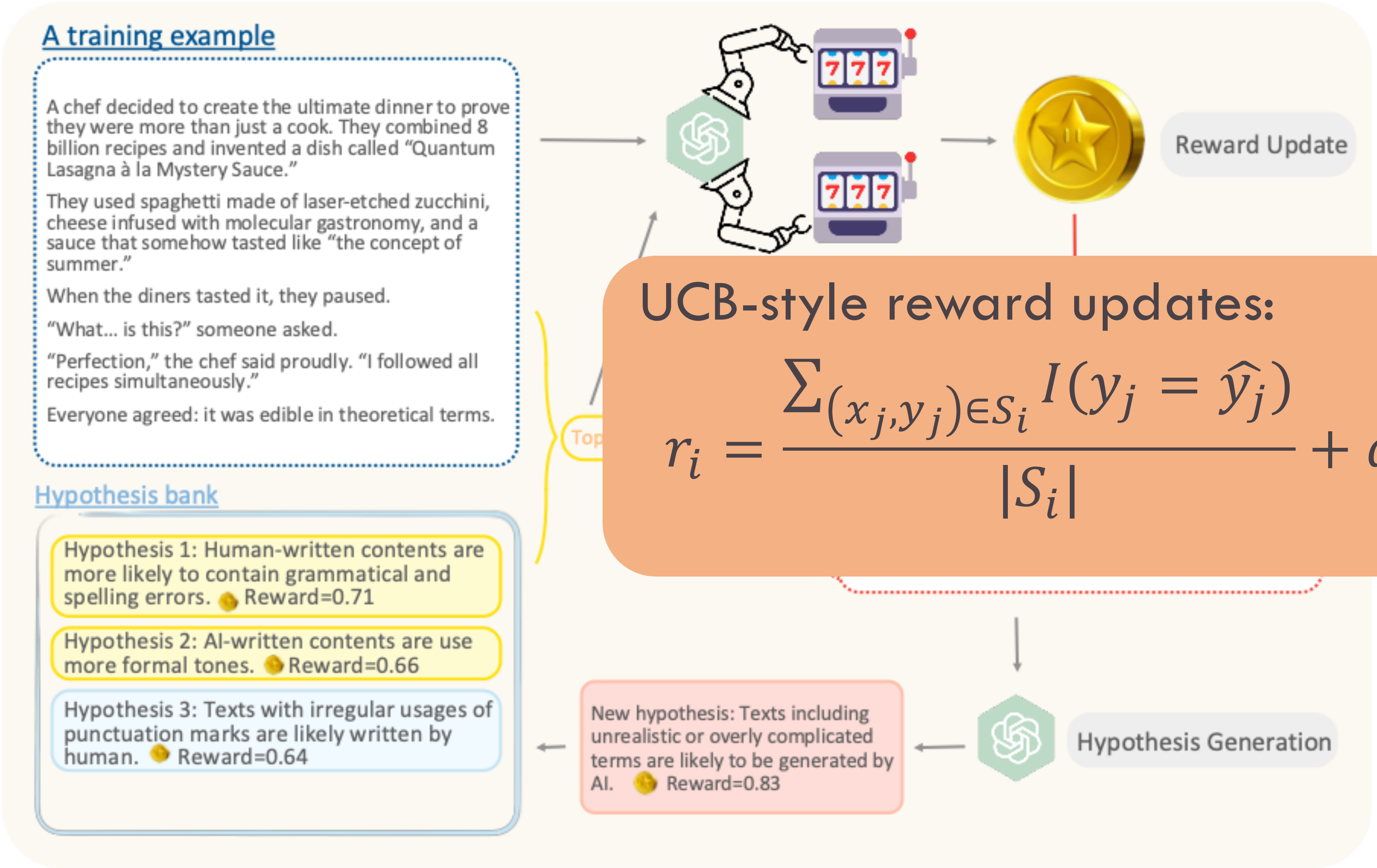


Hypogenic: A data-driven algorithm



Hypothesis initialization

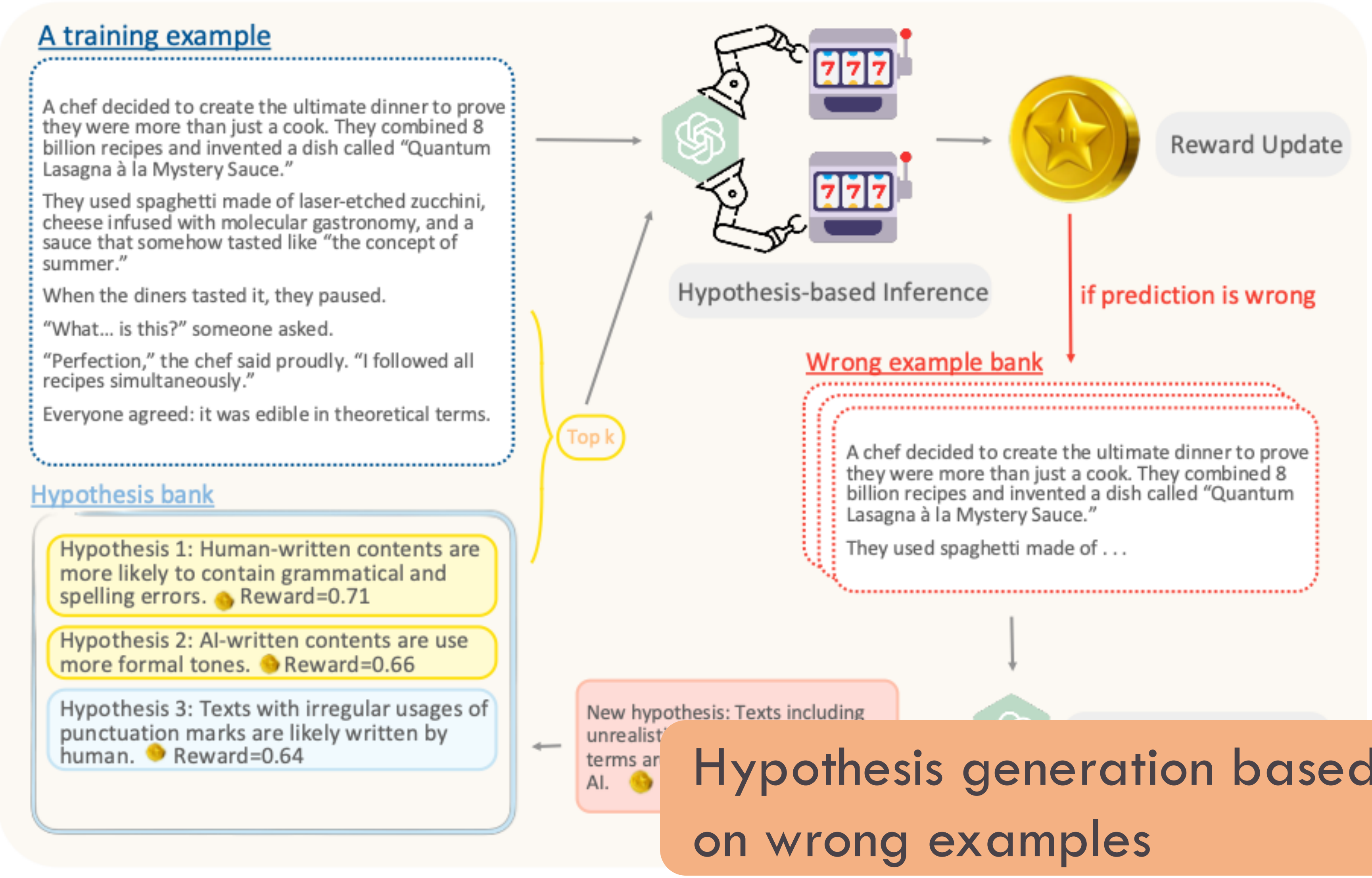
Hypogenic: A data-driven algorithm



UCB-style reward updates:

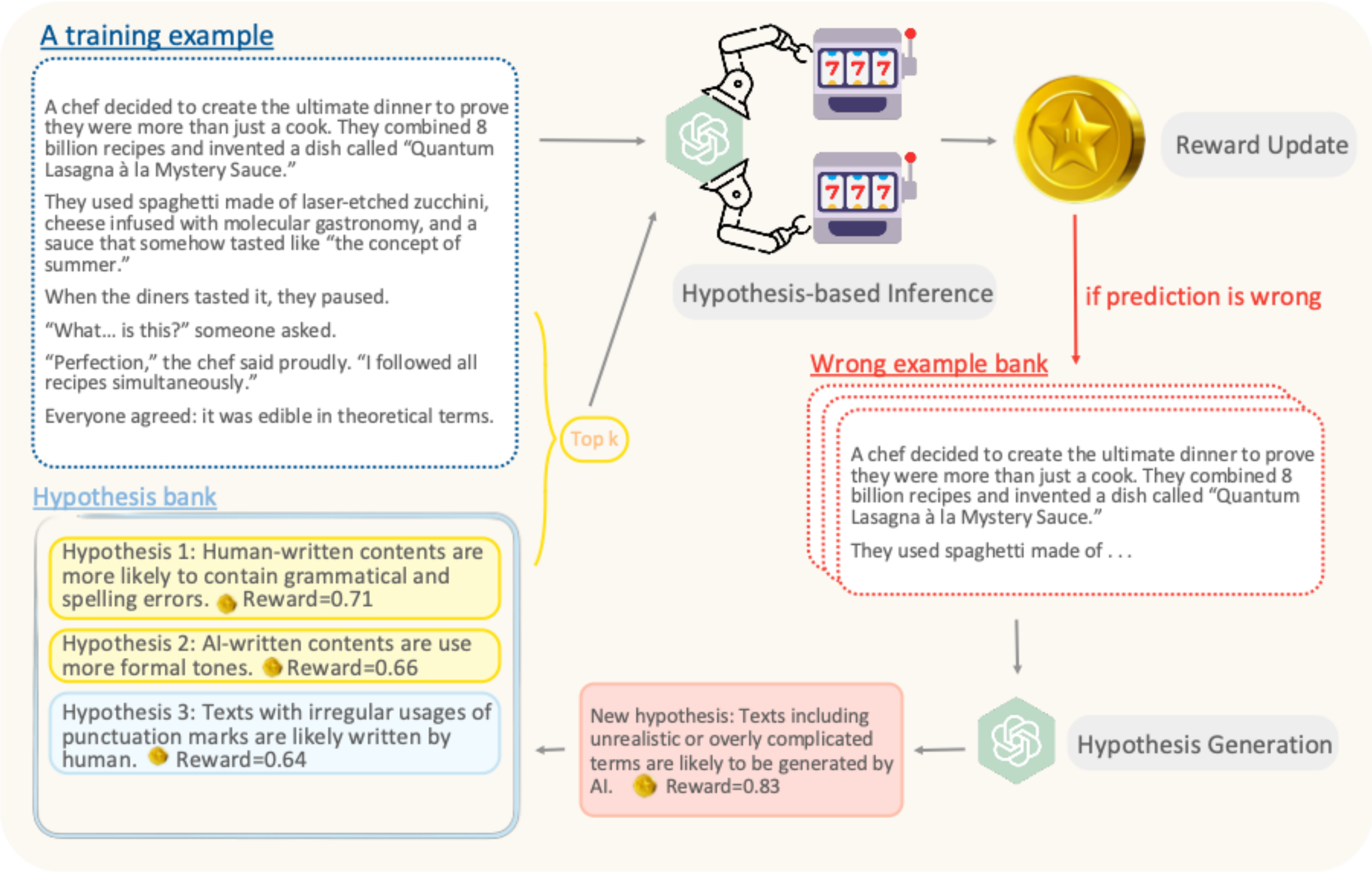
$$r_i = \frac{\sum_{(x_j,y_j)\in S_i} I(y_j = \hat{y}_j)}{|S_i|} + \alpha\sqrt{\frac{\log t}{|S_i|}}$$

Hypogenic: A data-driven algorithm



Hypothesis generation based on wrong examples

Hypogenic: A data-driven algorithm



Use data labels to guide hallucinations

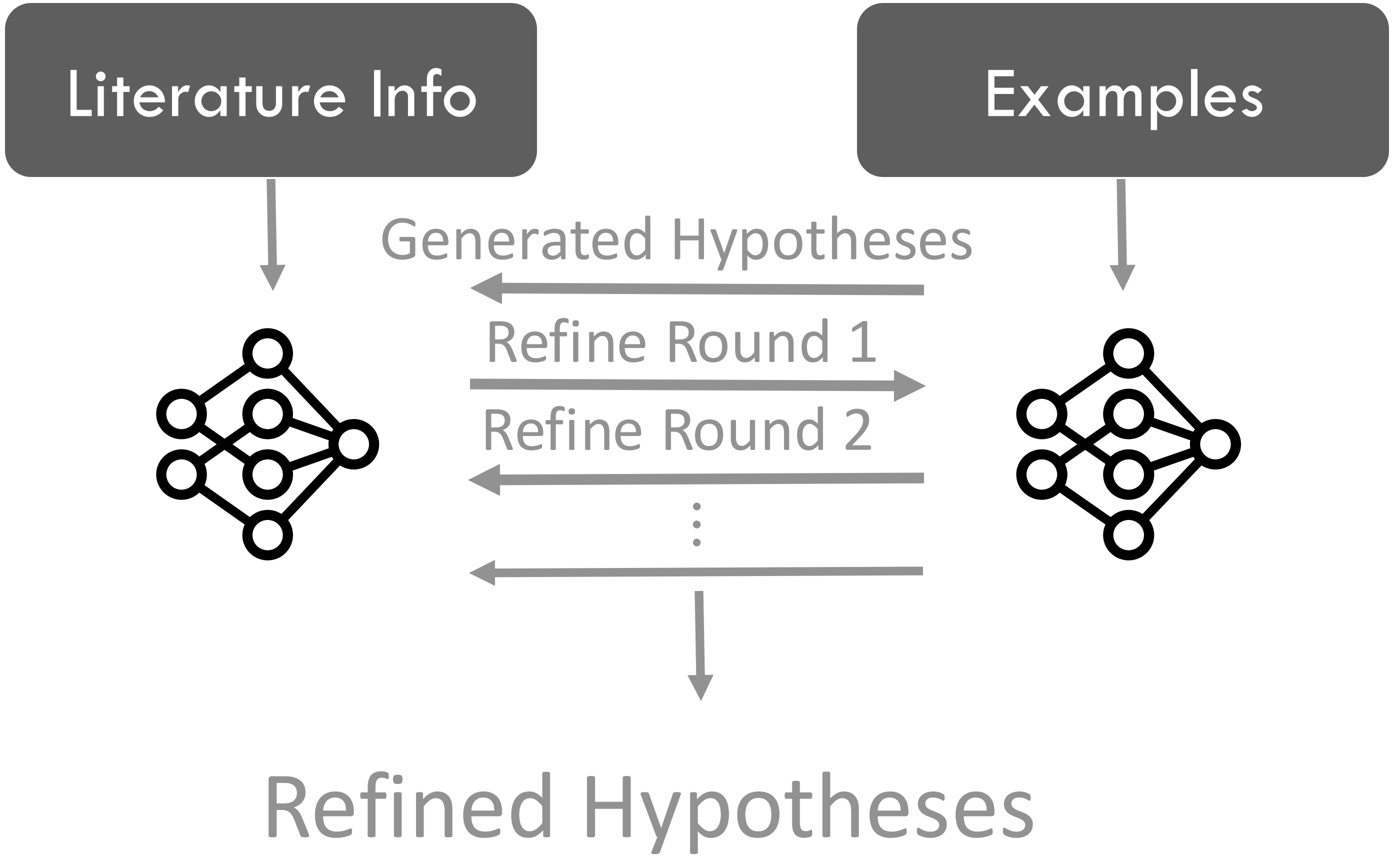
Literature-based hypothesis generation

Analogous to retrieval-augmented generation

- Search for relevant literature
- Summarize key findings of the retrieved literature
- Use key findings to generate hypotheses

Combining Hypogenic and Literature

- HypoRefine
- Literature + Hypogenic
- Literature + HypoRefine



Evaluation

- We can follow the recipe of supervised classification.
- However, what we care most about is **the quality of hypotheses:**
  - Qualitative examination
  - Human evaluation
  - Cross-generalization

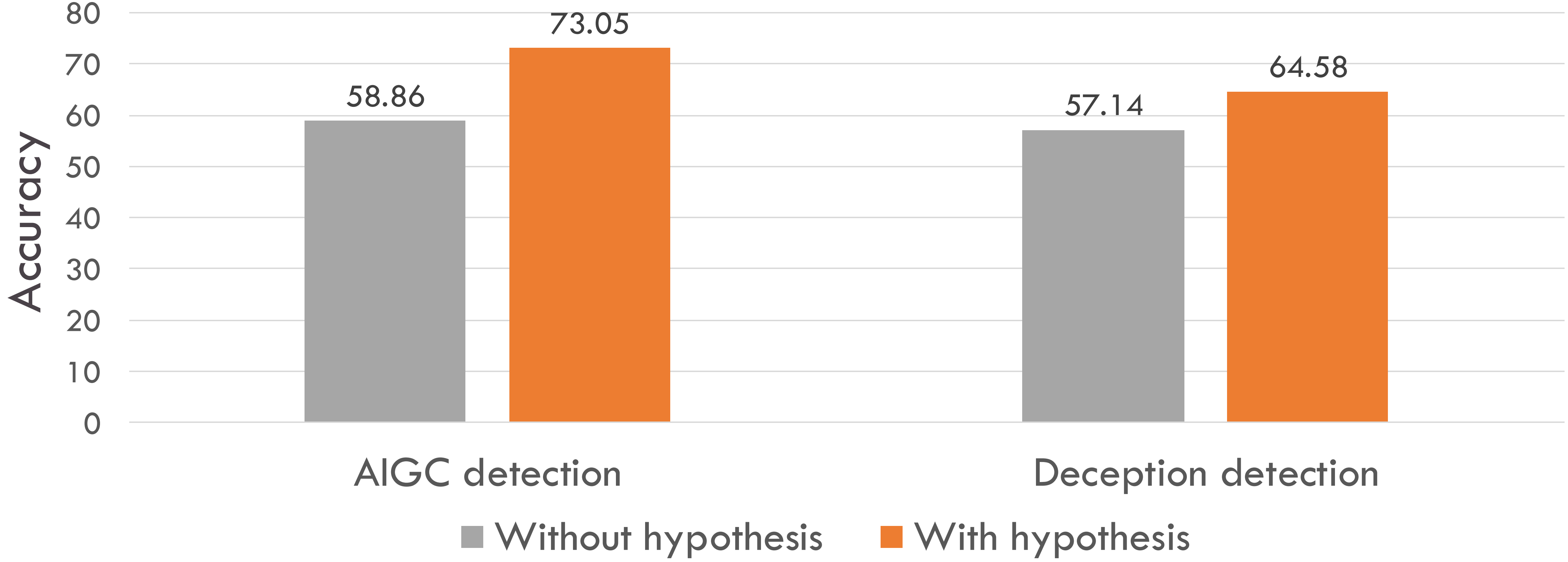
Example generated hypotheses for AIGC detection

- AI-generated texts tend to use more elaborate and descriptive language, including adjectives and adverbs, to create a sense of atmosphere and immersion. Human-written texts, on the other hand, tend to be more concise and straightforward in their language use.
- Human-written texts are more likely to contain errors or idiosyncrasies in grammar and punctuation, reflecting the natural imperfections of human writing, while AI-generated texts typically maintain a higher level of grammatical accuracy.
- Human-written texts tend to have more conversational tone and colloquial language, while AI-generated texts tend to be more formal and lack idiomatic expressions.

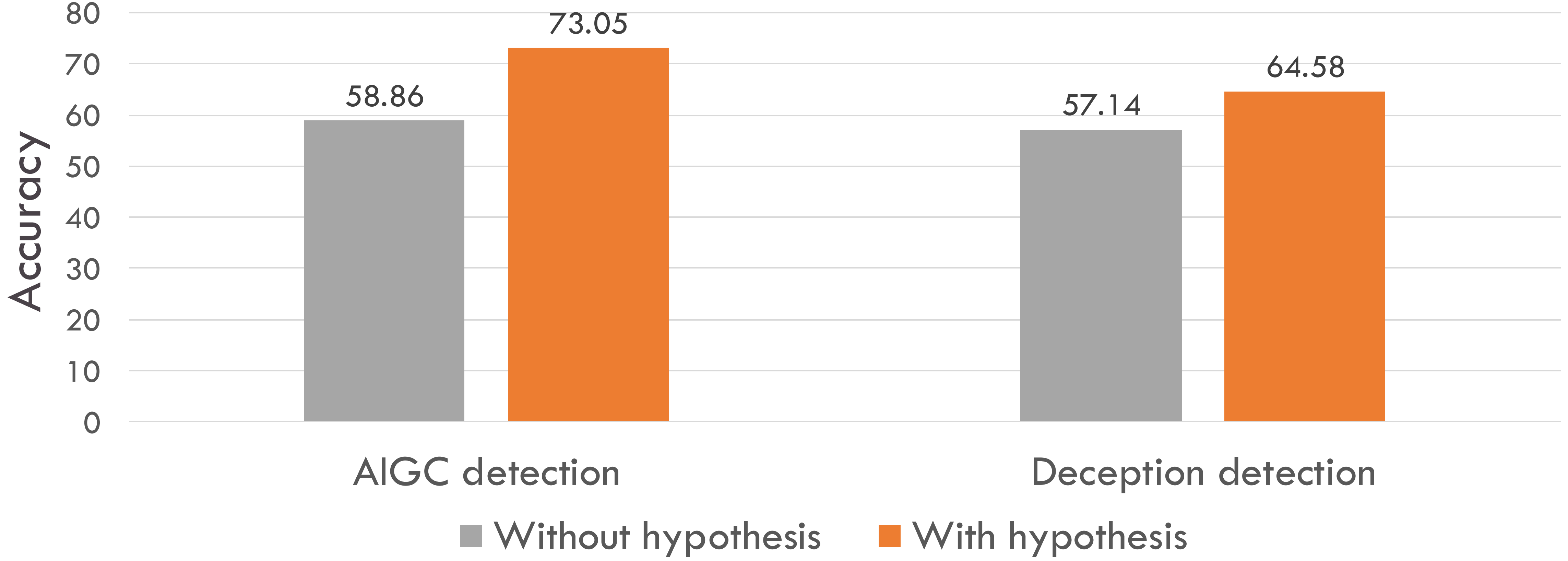
Example generated hypotheses for deception detection

- Reviews that present a balanced perspective by detailing both positive and negative experiences with specific examples (e.g., "the room was spacious and clean, but the noise from the street was disruptive at night") are more likely to be truthful, whereas reviews that express extreme sentiments without acknowledging any redeeming qualities (e.g., "everything was perfect" or "it was a total disaster") are more likely to be deceptive.
- Reviews that mention specific dates of stay or unique circumstances surrounding the visit (e.g., "We stayed during the busy Memorial Day weekend and faced long lines") are more likely to be truthful, while reviews that use vague temporal references (e.g., "I stayed recently") without concrete details are more likely to be deceptive, as they often lack the specificity that suggests a real and engaged experience.
- Reviews that provide detailed sensory descriptions of the hotel experience, such as the specific decor of the room, the quality of bedding, and the overall ambiance (e.g., "the room featured luxurious furnishings, high-thread-count sheets, and soft lighting that created a relaxing atmosphere") are more likely to be truthful, while reviews that use vague or overly simplistic descriptors (e.g., "the hotel was nice and comfortable") are more likely to be deceptive.

Generated hypotheses improve human decision-making



Generated hypotheses improve human decision-making



100% of the participants find the hypotheses to be helpful, and over 40% find them to be “Very helpful” or “Extremely helpful”.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 1: Literature-only and Hypogenic generate different hypotheses

Literature-only: Deceptive reviews often contain a higher frequency of first-person singular pronouns, while truthful reviews may use these pronouns less frequently.

Hypogenic: Reviews that reference the reviewer's previous experiences with the hotel brand or similar hotels are more likely to be truthful, while reviews that do not provide any context or comparison to past experiences are more likely to be deceptive.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 2: Literature-only and Hypogenic generate similar hypotheses

Literature-only: Truthful reviews often provide a balanced perspective, while deceptive reviews may seem overly promotional or biased towards a competitor.

Hypogenic: Reviews that express a balanced perspective, mentioning both positive and negative aspects of the stay, are more likely to be truthful, whereas reviews that are overly positive or negative without nuance tend to be deceptive.

Humans rate literature-based and data-driven hypotheses as distinct

- Case 2: Literature-only and Hypogenic generate similar hypotheses

HypoRefine: Reviews that present a balanced perspective by discussing both positive and negative aspects of the stay, particularly with specific examples (e.g., "The location was fantastic, but the air conditioning was broken"), are more likely to be truthful, while reviews that are excessively positive or negative without acknowledging any redeeming qualities (e.g., "This is the best hotel ever!" or "I will never stay here again!") tend to be more deceptive, as they may reflect an attempt to manipulate reader emotions rather than provide an honest assessment.

Automatic evaluation

- Five datasets:
  - Deception detection [Ott et al. 2013, Li et al. 2013]
  - GPTGC detection [Fan et al. 2018]
  - LlamaGC detection [Fan et al. 2018]
  - Persuasive argument detection [Pauli et al. 2024]
  - Mental stress detection (DREADDIT) [Turcan and McKeown 2019]
- We focus on out-of-distribution performance.
  - For example, LlamaGC is OOD for GPTGC.

Generated hypotheses outperform few-shot learning and other prompting approaches

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
GPT-4 MINI	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

An average improvement of 11.92% over few-shot

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
GPT-4 MINI	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Commercial applications cannot do this task at all

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
GPT-4 MINI	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Literature can hurt hypothesis generation in the case of AIGC

Model	Methods	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
	No hypothesis					
GPT-4 MINI	Zero-shot	55.47	50.00	56.33	81.24	64.60
	Few-shot k=3	65.56	51.11	64.22	83.64	75.00
	Zero-shot generation	68.69	49.00	53.00	86.08	65.00
	Literature-based					
	LITERATURE-ONLY	59.22	49.00	54.00	78.80	67.68
	HYPERWRITE	61.63	49.67	52.67	82.36	68.76
	NOTEBOOKLM	53.03	49.33	51.67	68.96	62.28
	Data-driven					
	HYPOGENIC	75.22	81.67	68.56	82.20	76.56
	Literature + Data (This work)					
	HYPOREFINE	**77.78**	55.33	63.33	89.04	78.04
	Literature ∪ HYPOGENIC	72.41	**83.00**	**69.22**	**89.88**	78.20
	Literature ∪ HYPOREFINE	77.19	55.33	63.00	89.52	**79.24**

Generated hypotheses can be effectively transferred to a different model

Generation Model	Inference Model	Deceptive reviews	LlamaGC	GPTGC	Persuasive Pairs	Dreaddit
		OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy
GPT-4-mini	GPT-4-mini	77.78	83.00	69.22	89.88	79.24
	Llama-70B-I	72.53 (↓5.25)	71.67 (↓11.33)	76.33 (↑7.11)	86.88 (↓3.00)	72.36 (↓6.88)
Llama-70B-I	Llama-70B-I	73.72	81.33	78.67	88.76	78.92
	GPT-4-mini	70.31 (↓3.41)	57.00 (↓24.33)	74.67 (↓4.00)	89.36 (↑0.60)	77.28 (↓1.64)

Generated hypotheses can be effectively transferred to a different model

Generation Model	Inference Model	DECEPTIVE REVIEWS	LLAMAGC	GPTGC	PERSUASIVE PAIRS	DREADDIT
		OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy	OOD Accuracy
GPT-4-MINI	GPT-4-MINI	77.78	83.00	69.22	89.88	79.24
	LLAMA-70B-I	72.53 (↓5.25)	71.67 (↓11.33)	76.33 (↑7.11)	86.88 (↓3.00)	72.36 (↓6.88)
LLAMA-70B-I	LLAMA-70B-I	73.72	81.33	78.67	88.76	78.92
	GPT-4-MINI	70.31 (↓3.41)	57.00 (↓24.33)	74.67 (↓4.00)	89.36 (↑0.60)	77.28 (↓1.64)

Our methods still outperform the few-shot inference baseline by 3.76%.

AI will drive future hypothesis generation.

AI will drive future hypothesis generation.

Website: https://chicagohai.github.io/hypogenic-demo/

Code: https://github.com/ChicagoHAI/hypothesis-generation

Data: https://huggingface.co/collections/ChicagoHAI/hypothesis-generation-6719515102874a461f47ae57

Chicago Human+AI

discovery-bench

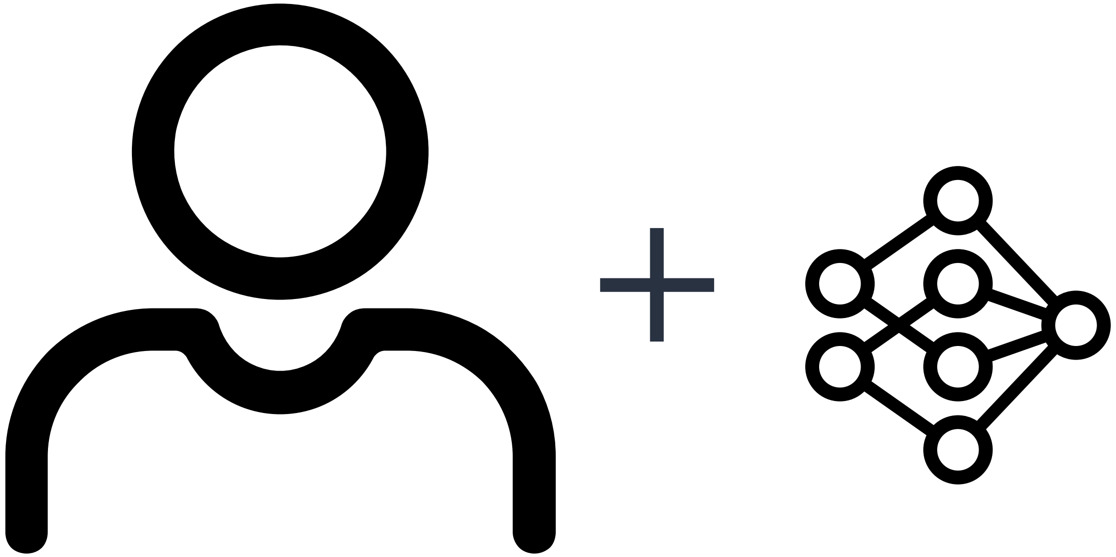
https://github.com/allenai/discoverybench/

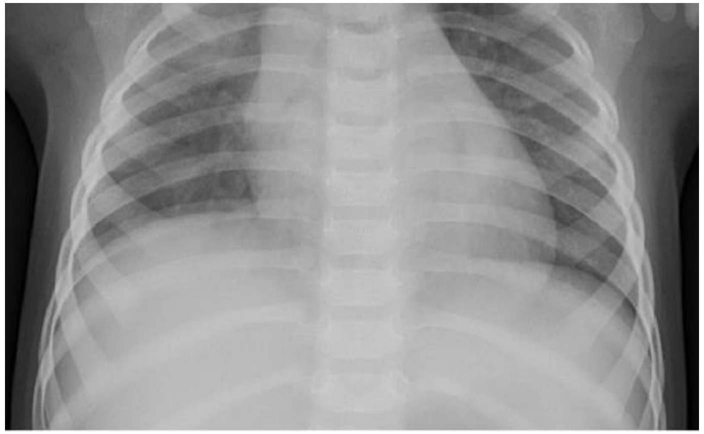
  

Human-AI decision making

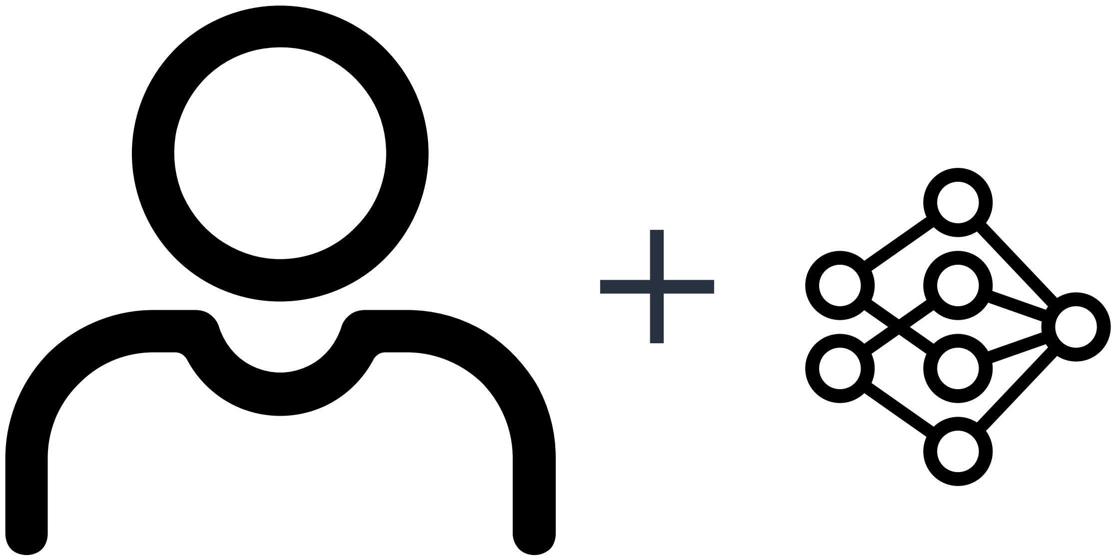
Machine Explanations and Human Understanding. *Chacha Chen, Shi Feng, Amit Sharma, Chenhao Tan*. TMLR 2023; FAccT 2023.

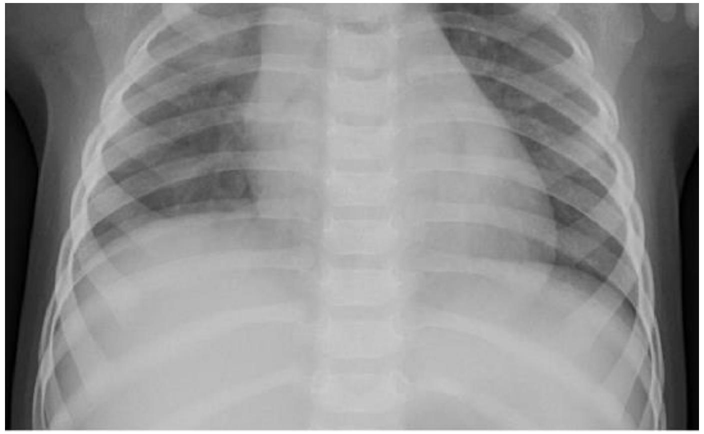
Learning Human-Compatible Representations for Case-Based Decision Support. *Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan*. ICLR 2023.



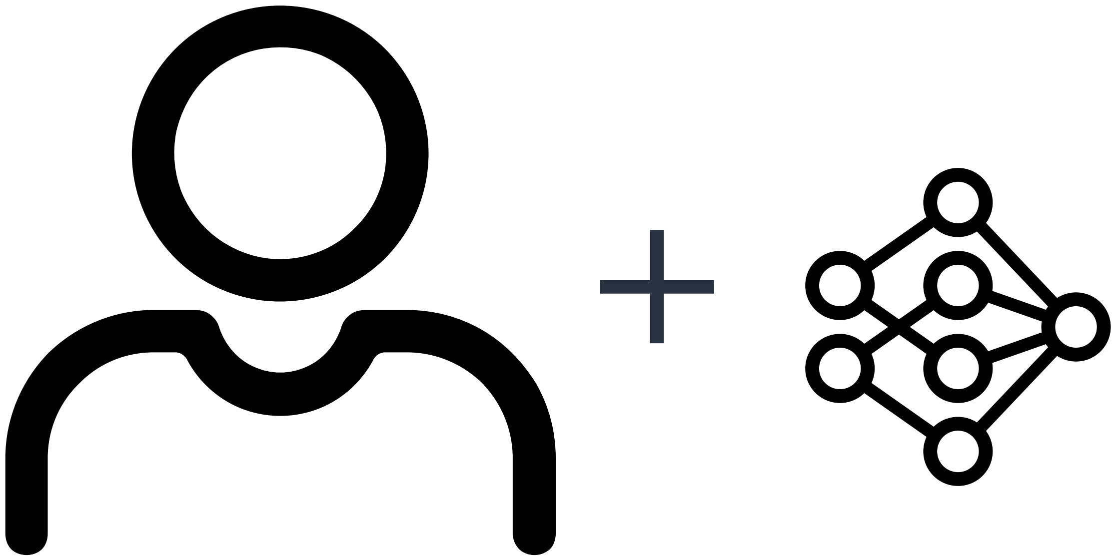


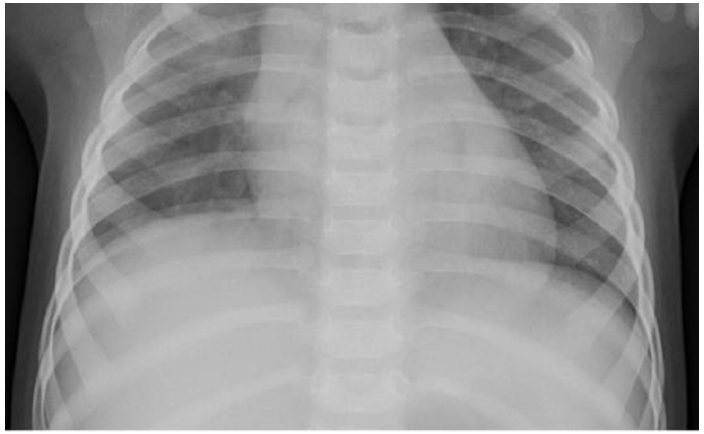
Human goals: humans achieving high accuracies





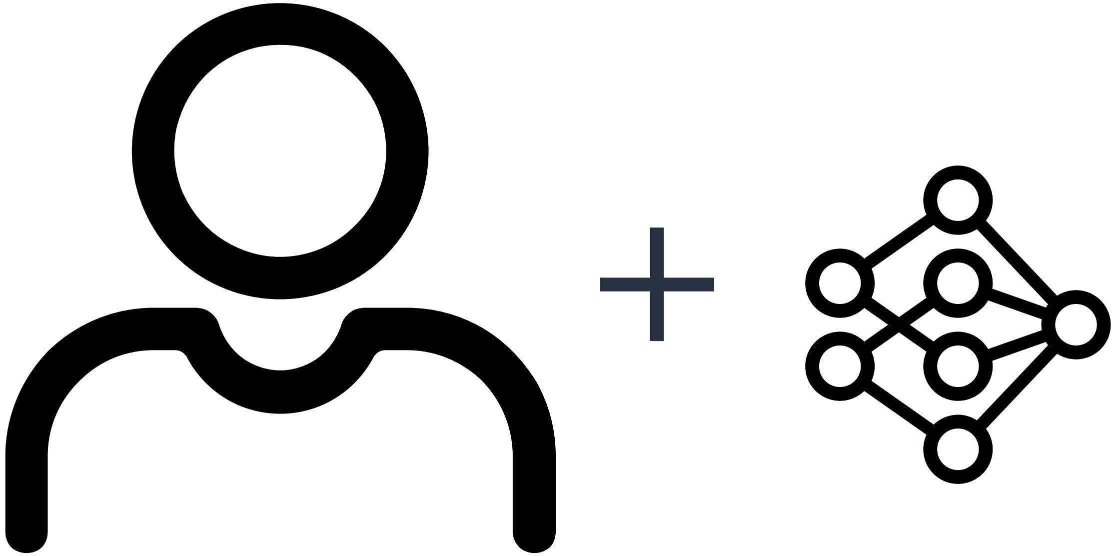
Human goals: humans achieving high accuracies

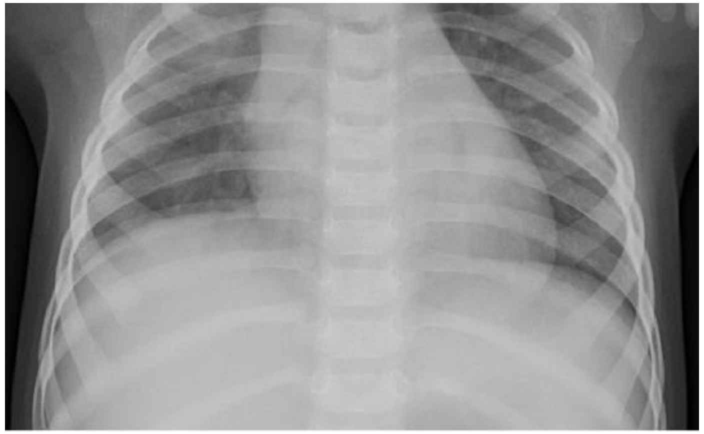




What kind of AI assistance can be helpful?

Human goals: humans achieving high accuracies



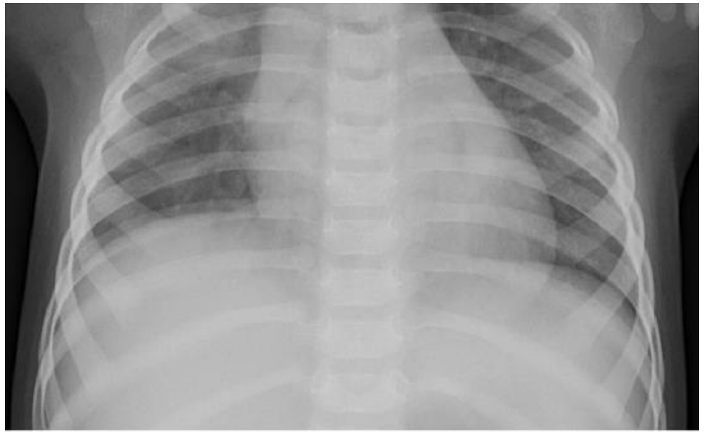


What kind of explanations can be helpful?

Task: Pneumonia diagnosis

Consider two cases:

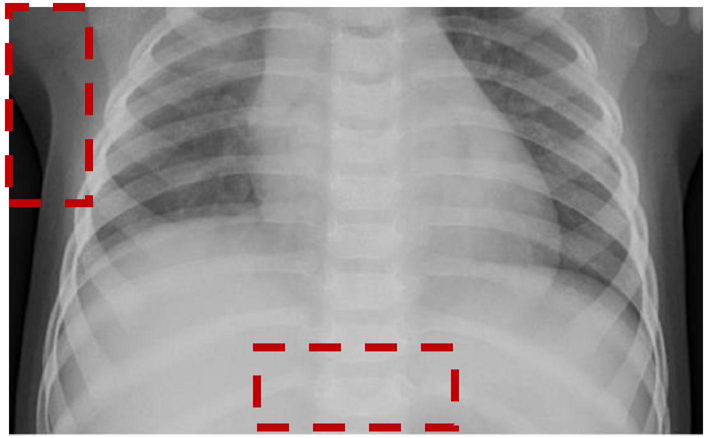
1. "User" has no task-specific intuitions
2. "User" has task-specific intuitions



Task: Pneumonia diagnosis

Consider two cases:

1. “User” has no task-specific intuitions
2. “User” has task-specific intuitions

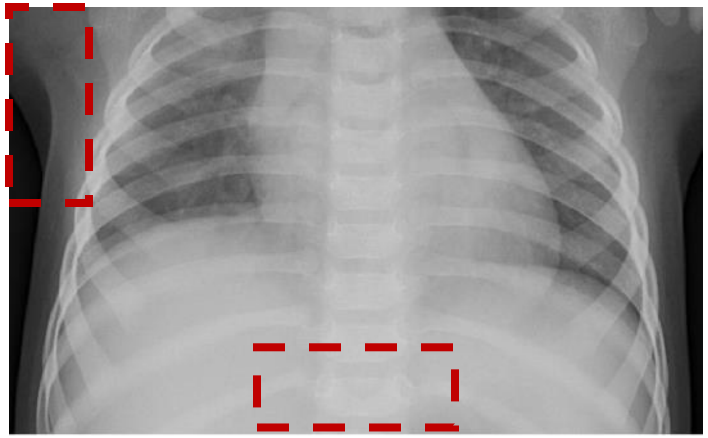


- User cannot make sense of explanations
- Understanding of task decision boundary is bounded by the model decision boundary

Task: Pneumonia diagnosis

Consider two cases:

1. "User" has no task-specific intuitions
2. "User" has task-specific intuitions



- One possible mechanism is that human can use explanations to verify whether the model uses valid information
- Hopefully, human+AI > AI

Task-specific human intuitions are necessary for explanations to provide value in AI-assisted decision making.

Task-specific human intuitions are necessary for explanations to provide value in AI-assisted decision making.

In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making

Raymond Fok
rayfok@cs.washington.edu
University of Washington
Seattle, WA, USA

Daniel S. Weld
danw@allenai.org
Allen Institute for AI &
University of Washington
Seattle, WA, USA

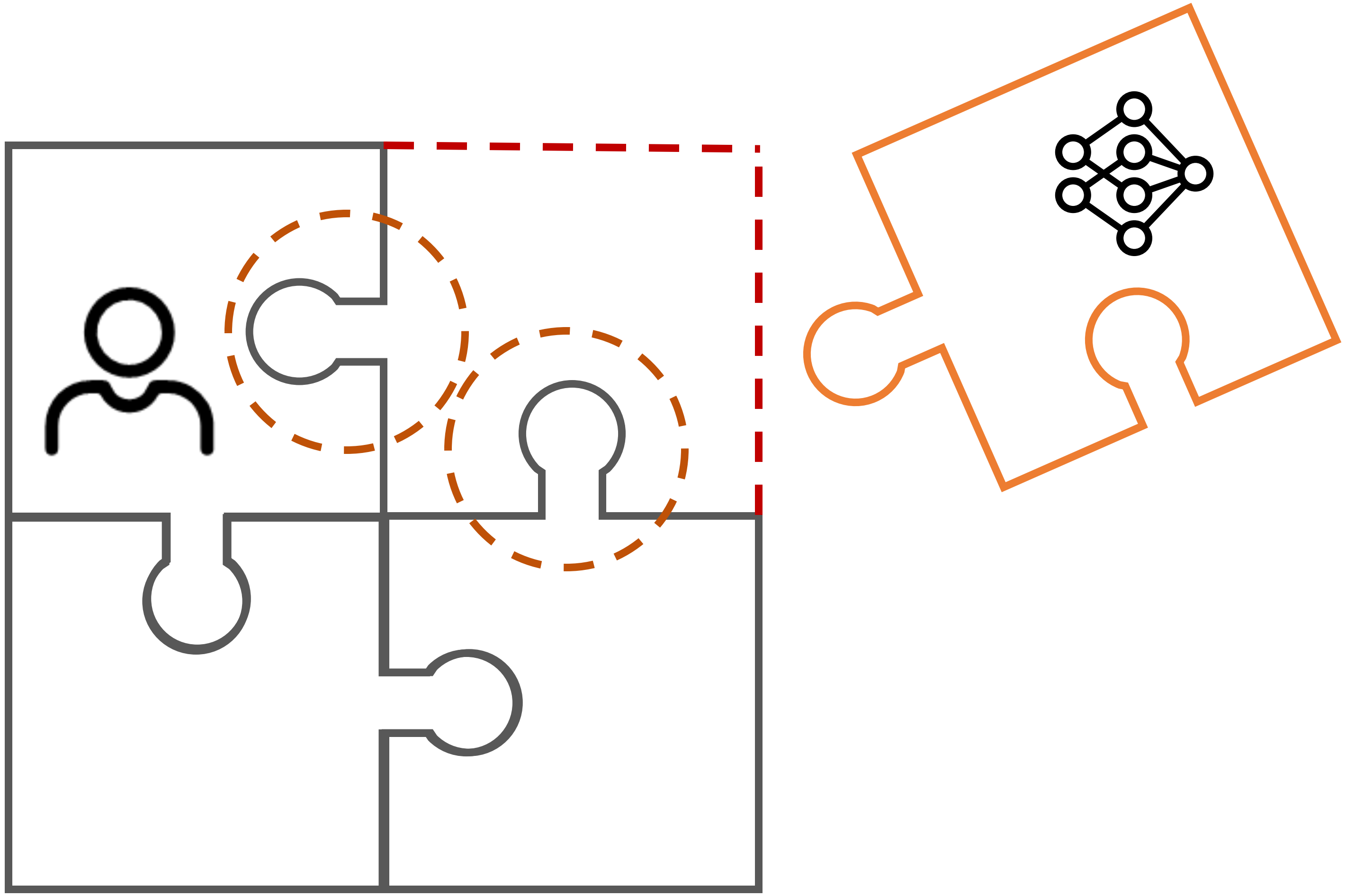
Designing Theory-Driven User-Centric Explainable AI

Danding Wang[1], Qian Yang[2], Ashraf Abdul[1], Brian Y. Lim[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, United States
wangdanding@u.nus.edu, yangqian@cmu.edu, ashrafabdul@u.nus.edu, brianlim@comp.nus.edu.sg

Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI

UPOL EHSAN, Georgia Institute of Technology, USA
KOUSTUV SAHA, Microsoft Research, Canada
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA
MARK O. RIEDL, Georgia Institute of Technology, USA

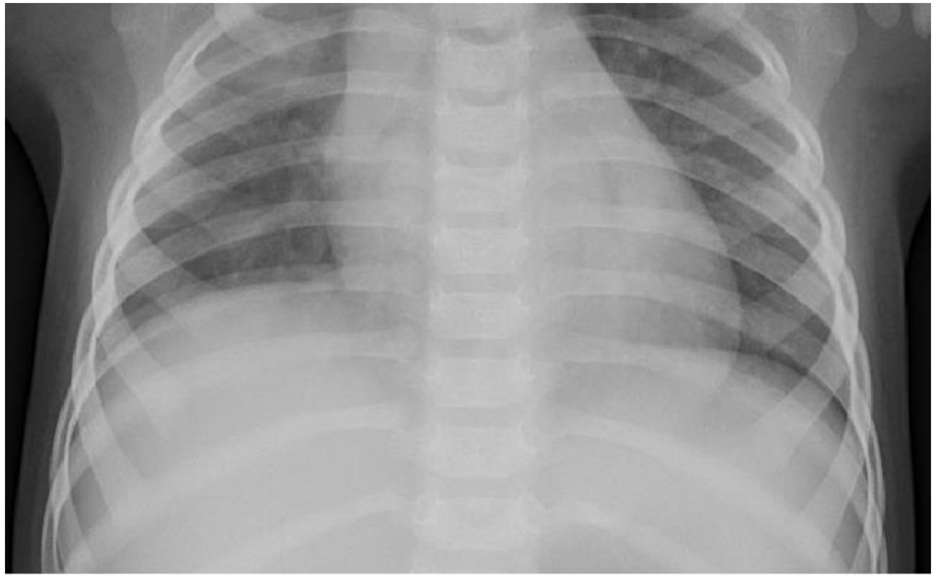
Task-specific human intuitions are critical for the goal of human-AI decision making



Human-centered explanations

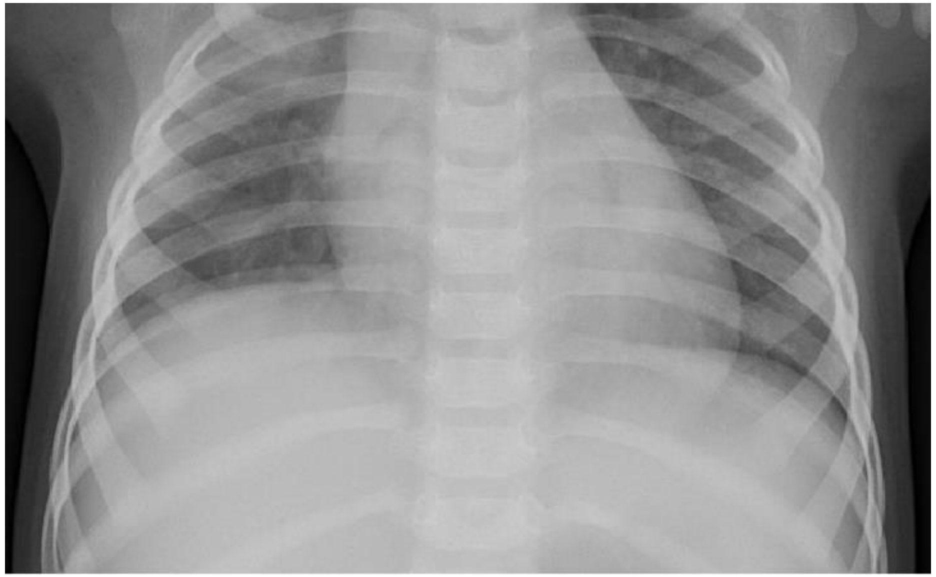
1. Articulate the mechanism of how humans may interact with explanations through task-specific intuitions
2. Generate explanations that tailor to this mechanism

Pneumonia diagnosis



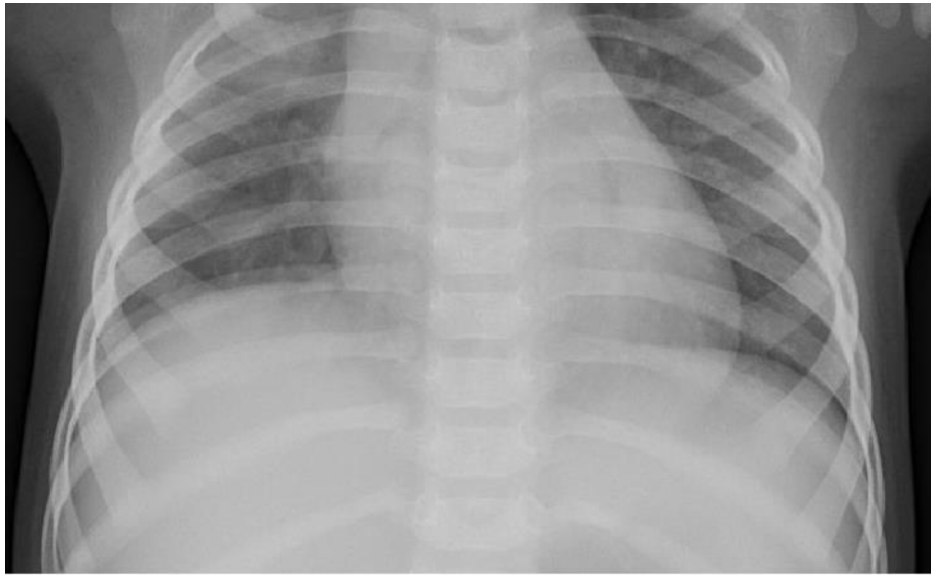
Pneumonia diagnosis

AI predicts pneumonia



Pneumonia diagnosis

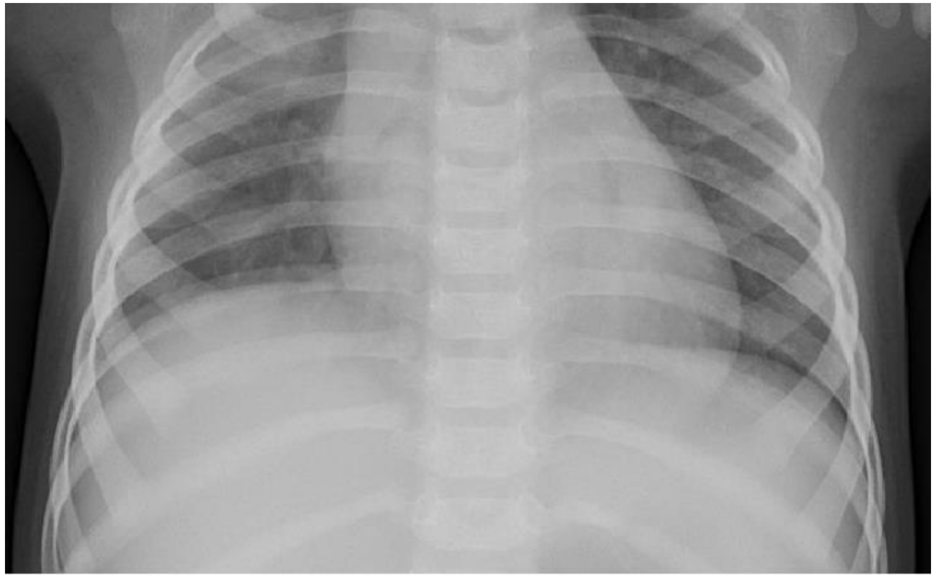
AI predicts pneumonia



Justification of the prediction

Pneumonia diagnosis

AI predicts pneumonia



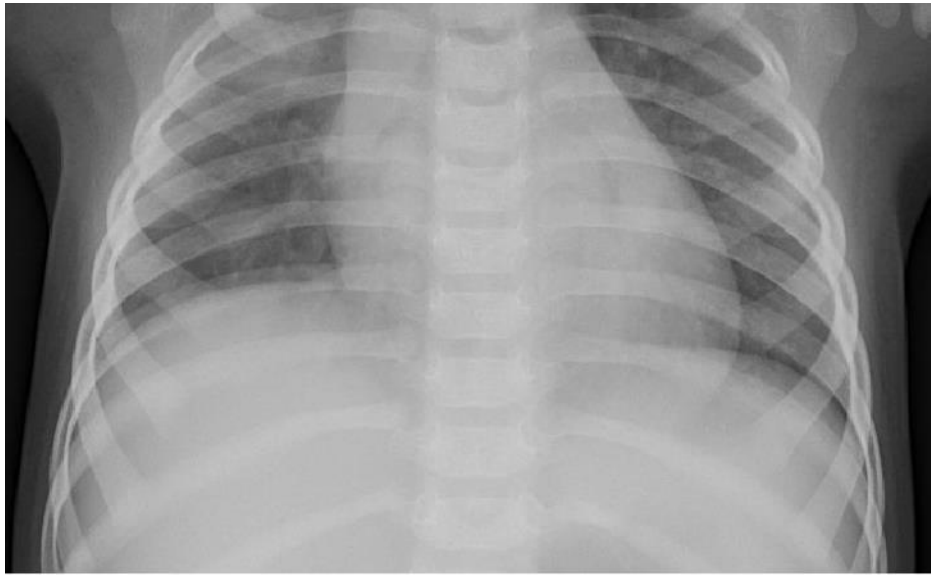
Justification of the prediction

Similarity between test example and justification is correlated with model error

Similarity to AI is aligned with similarity to human

Pneumonia diagnosis

AI predicts pneumonia

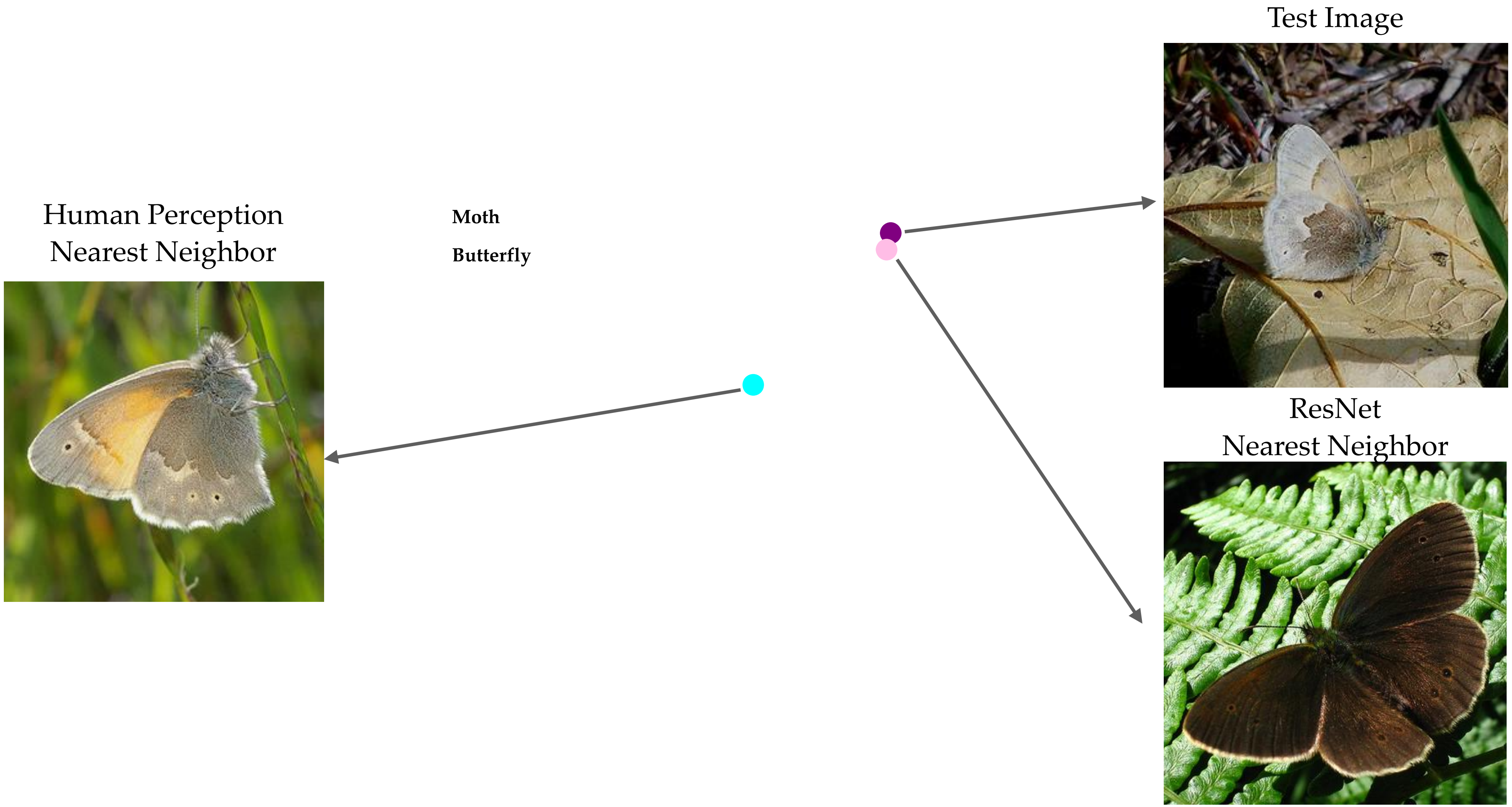


Justification of the prediction

Similarity between test example and justification is correlated with model error

Similarity to AI is aligned with similarity to human

Out-of-the-box AI does not lead to human-centered explanations



Explanations in this case are directly derived from AI representations.

Explanations in this case are directly derived from AI representations.

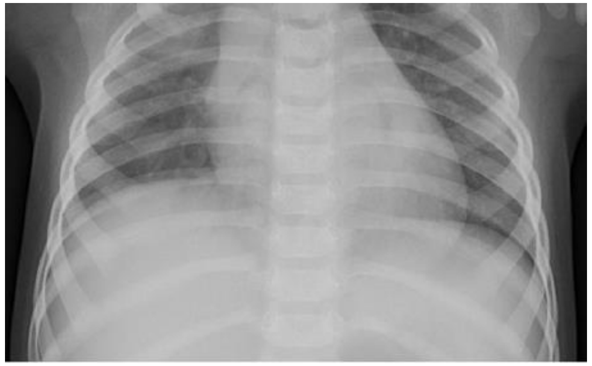
The culprit lies in misaligned AI representations.

Explanations in this case are directly derived from AI representations.

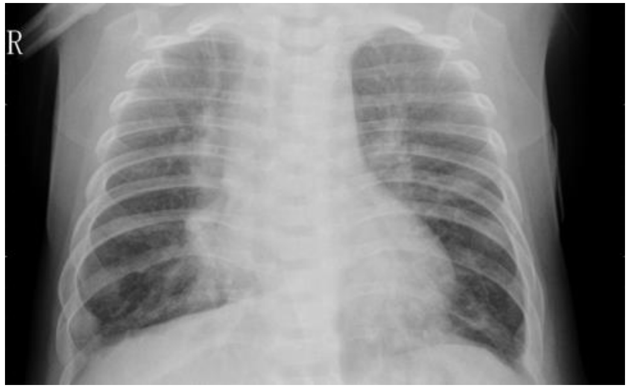
The culprit lies in misaligned AI representations.

Learning human-compatible representations!

Collect human triplet judgments



first large-scale triplet dataset on chest x-rays



Reference A

Reference B

Agarwal et al. 2007, Balntas et al. 2016

Learning human-compatible representations

A multi-task learning framework with two objectives

- Image classification
- Human judgment prediction

Use human perception mechanisms to guide AI explanations

$$\underbrace{\lambda\left[-\sum_{(x,y)\sim D}\log\left(p_\theta(y|x)\right)\right]}_{\text{Cross-entropy loss}}+\underbrace{(1-\lambda)\left[\sum_{(x^r,x^+,x^-)\sim T}\max\left(d_\theta(x^r,x^+)-d_\theta(x^r,x^-)+1,0\right)\right]}_{\text{Triplet margin loss}}$$

Learning human-compatible representations

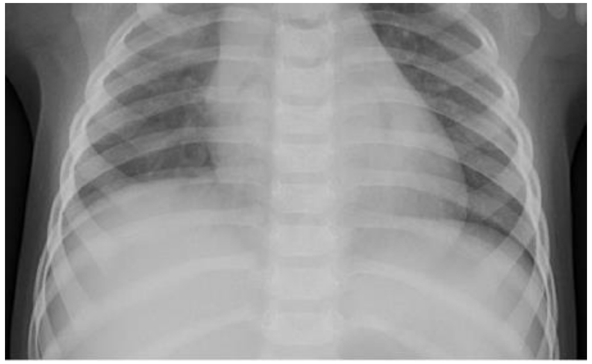
A multi-task learning framework with two objectives

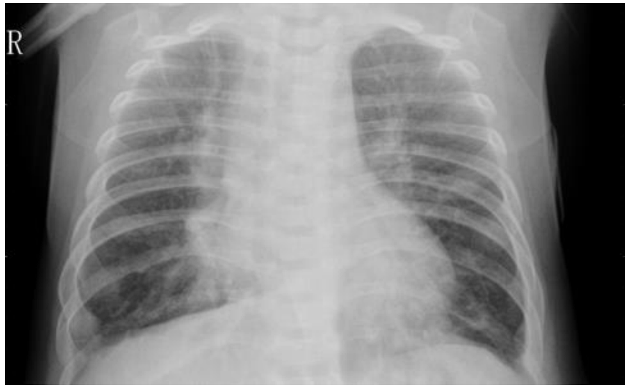
- Image classification
- Human judgment prediction

$$\underbrace{\lambda\left[-\sum_{(x,y)\sim D}\log\left(p_\theta(y|x)\right)\right]}_{\text{Cross-entropy loss}}+\underbrace{(1-\lambda)\left[\sum_{(x^r,x^+,x^-)\sim T}\max\left(d_\theta(x^r,x^+)-d_\theta(x^r,x^-)+1,0\right)\right]}_{\text{Triplet margin loss}}$$

Experiment setup: neutral decision support

Determine the diagnosis based on which support images looks more similar to the original one

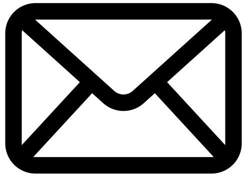




Nearest neighbor in the predicted class

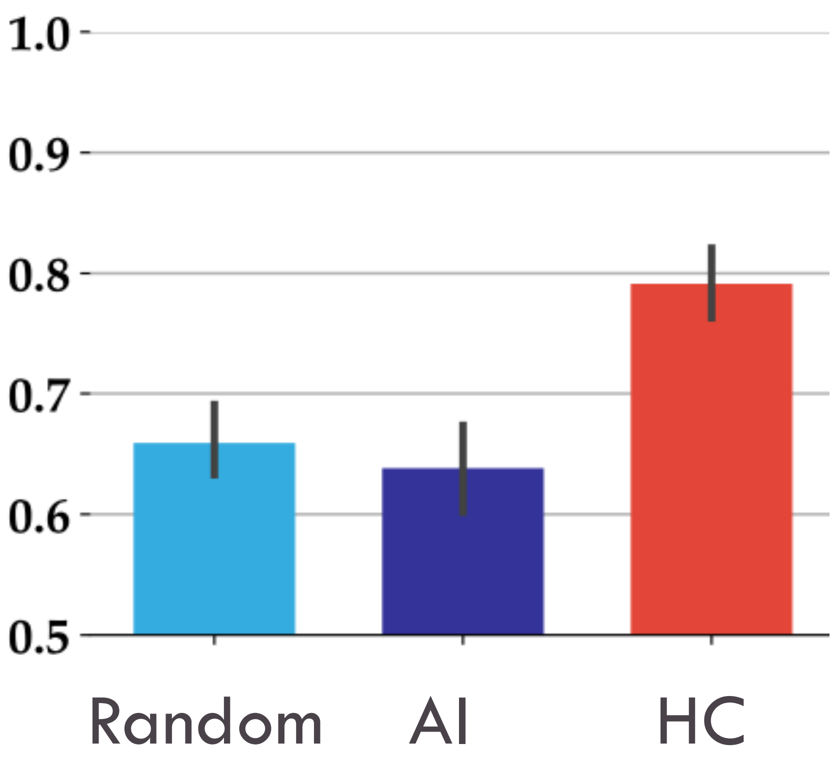
Nearest neighbor in the other class

Experiment setup: neutral decision support

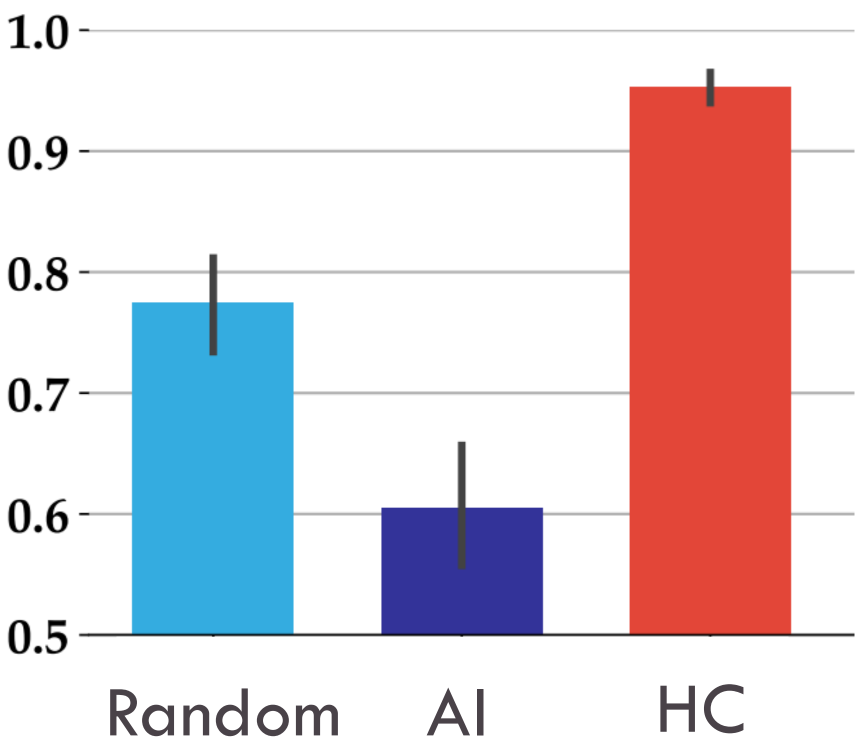


- Random (dumb AI)
- AI
- AI with human-compatible representations

Human-compatible representations lead to more effective decision support



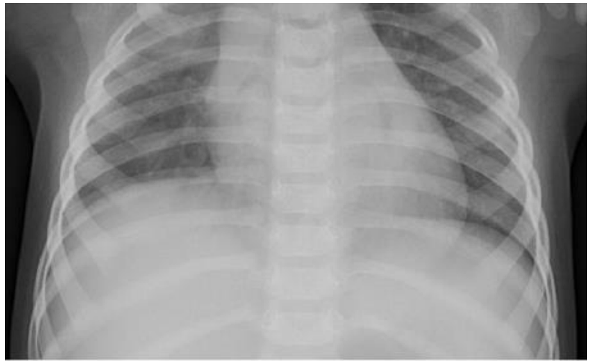
Pneumonia classification

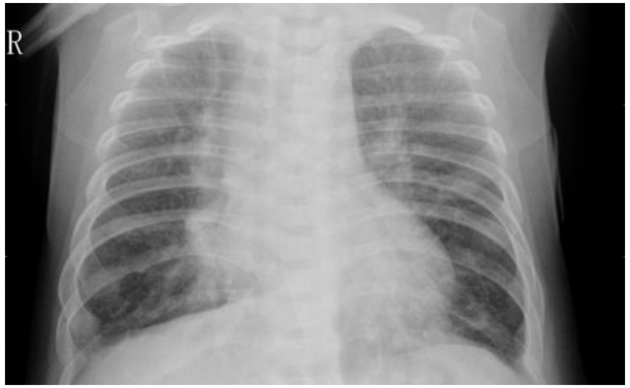


Butterfly vs. Moth

Experiment setup: persuasive decision support

Determine the diagnosis based on which support images looks more similar to the original one

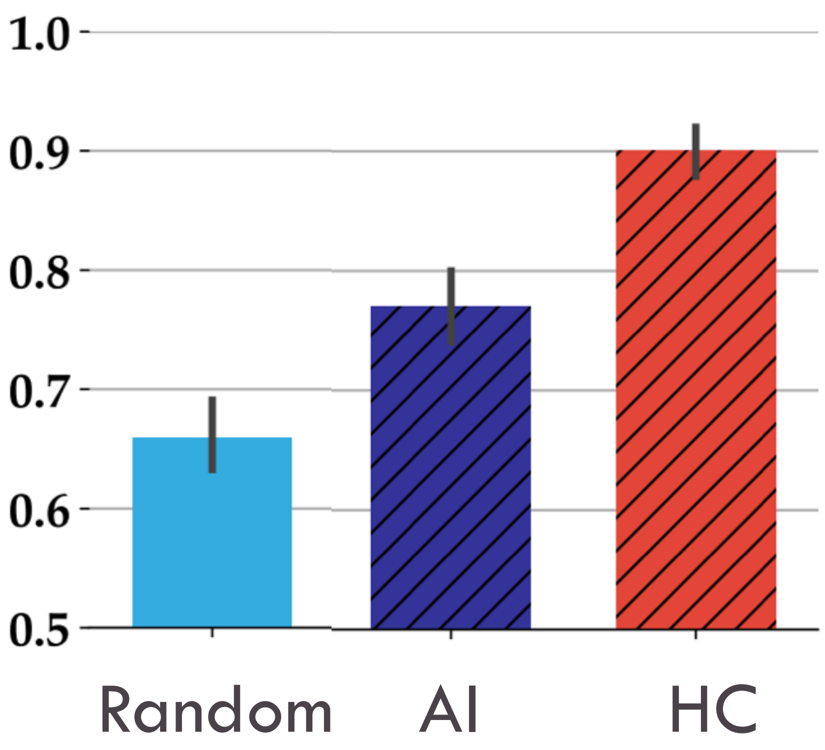




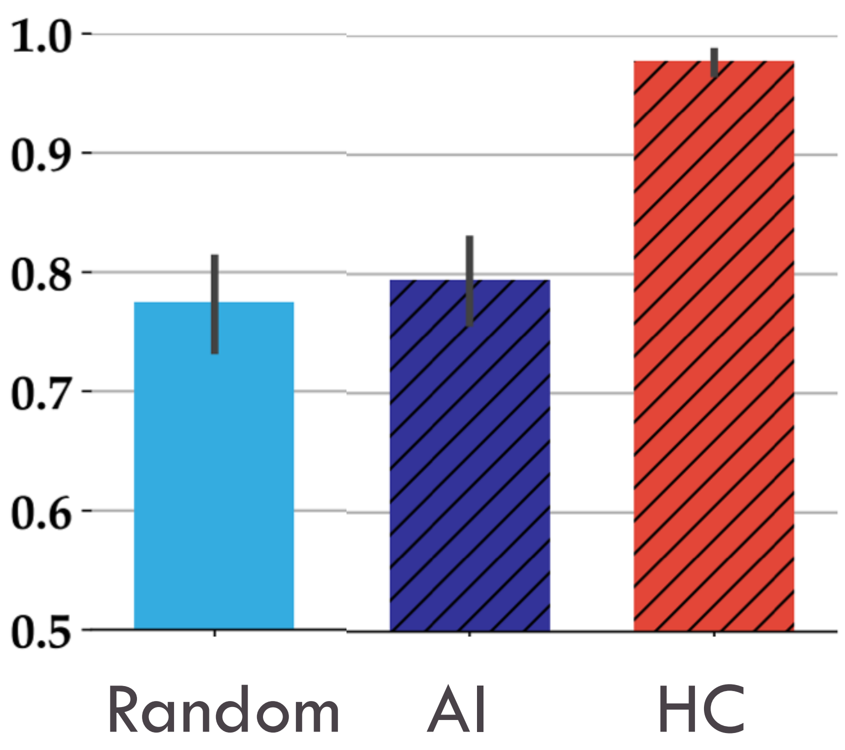
Nearest neighbor in the predicted class

Farthest neighbor in the other class

Human-compatible representations also lead to more persuasive decision support



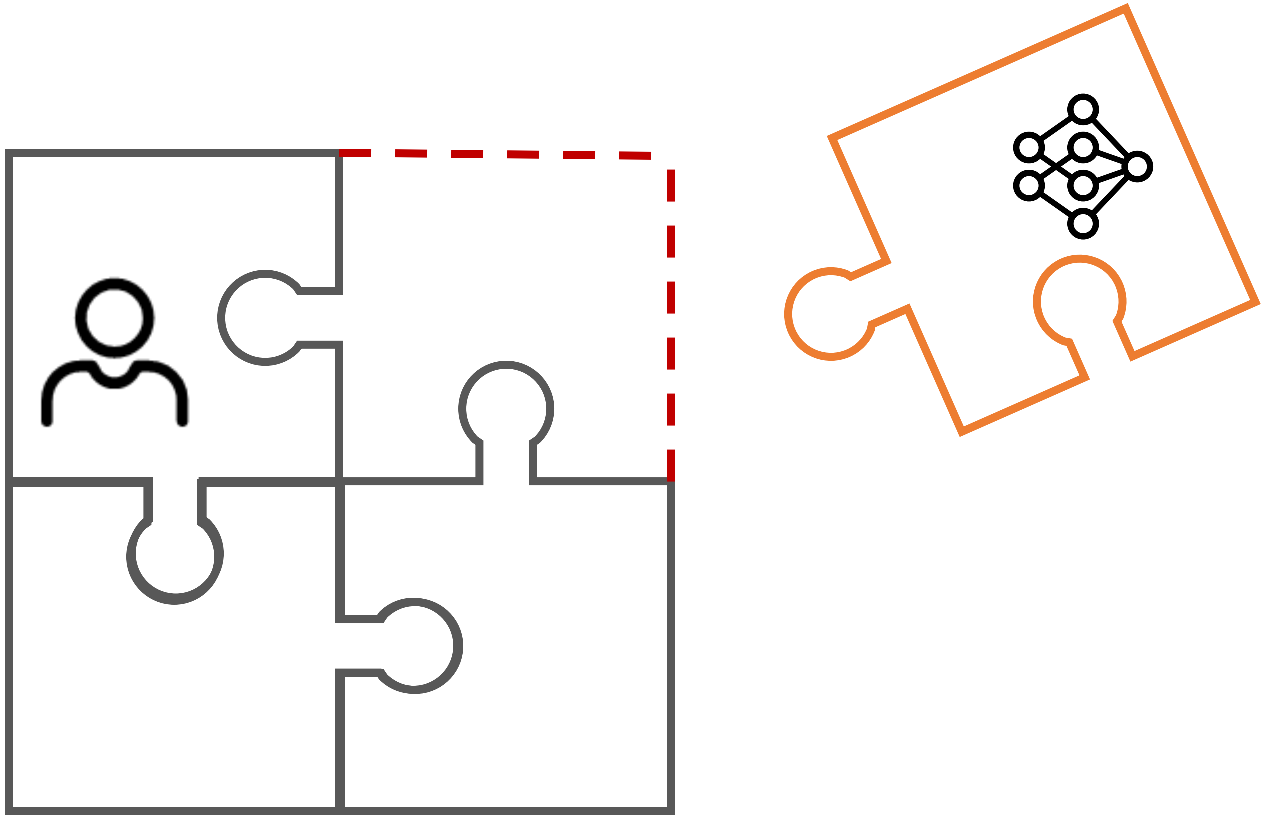
Pneumonia classification



Butterfly vs. Moth

Complementary AI

- Understanding human goals and human capabilities
- Understanding human-AI interaction
- Reshaping AI with new objectives, datasets, and algorithms



chenhao@uchicago.edu