CS 591 RS · OCTOBER 26, 2026

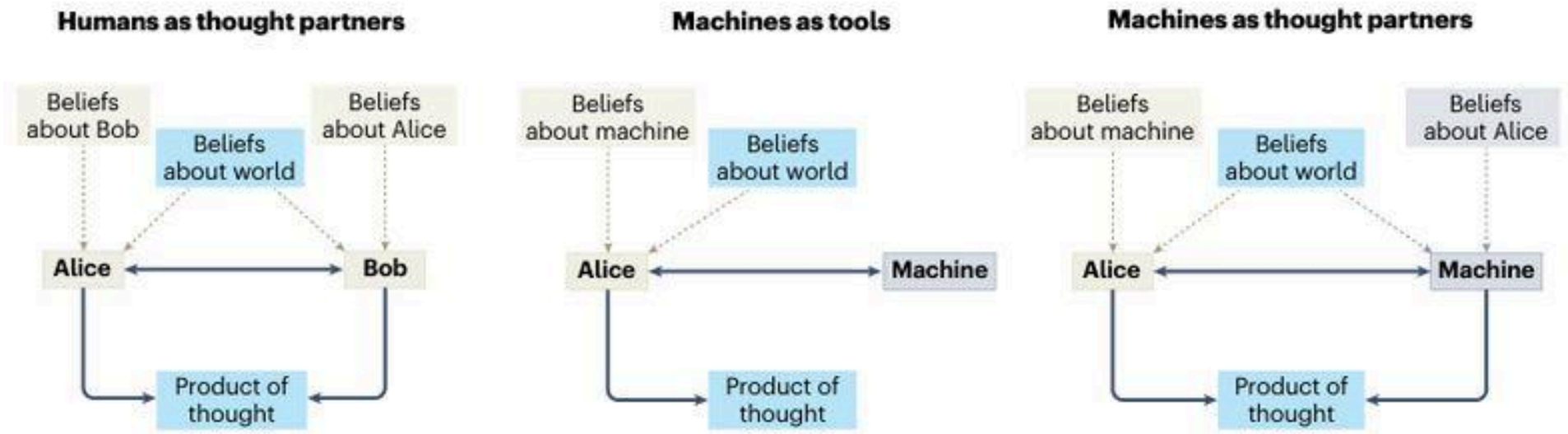
AI Thought Partners

Machines that think with people, not instead of them

Ilia Sucholutsky

AI Thought Partner Lab · Purdue Computer Science

A partner, not a tool



A thought partner has a model of you · you can understand it · you share a model of the world

Collins, Sucholutsky et al., Nature Human Behaviour 2024

person · AI · pair · group | capable · learns fast · aligned · reliable · safe · empowering

What do we want from an AI thought partner?

Capable
good at the task that matters

Learns fast
picks up what it needs from a handful of examples

Aligned
shares concepts and values with the people it works with

Reliable
reasons soundly and knows its own limits

Safe
doesn't manipulate, isn't easily manipulated, earns the right amount of trust

Empowering
leaves people, and the groups they form, more capable rather than more dependent

person AI pair group | capable learns fast aligned reliable safe empowering

Four objects of study



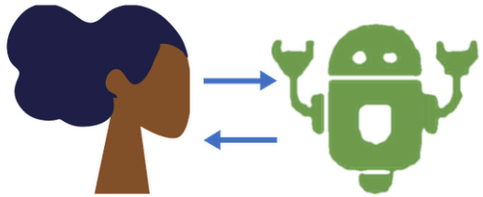
Person

how people represent, perceive, choose and judge



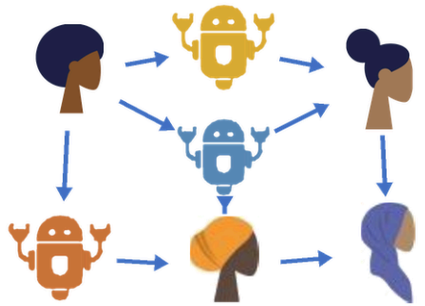
AI

how AI systems represent, reason, choose and judge



Pair

one person and one AI thinking together



Group

many people and AIs learning from each other

person AI pair group | capable learns fast aligned reliable safe empowering

LEARNING FROM LITTLE

How little does it take to learn, for people and for AI?

person AI pair group | capable learns fast aligned reliable safe empowering

Catching a unicorn



+



person AI pair group | capable learns fast aligned reliable safe empowering

…or a rhinocorn



How we describe the mixture matters: soft labels have to be given carefully.

person AI pair group | capable learns fast aligned reliable safe empowering

GAME 1 · YOU BE THE ZOOLOGIST

I'm thinking of an animal that is

50% beaver · **40%** duck · **10%** otter

What is it?

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

A platypus



That was a **soft label**: how a new thing relates to things you already know.

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

BONUS ROUND

70% whale · **30%** unicorn

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

A narwhal



Photo: Kristin Laidre / NOAA, public domain

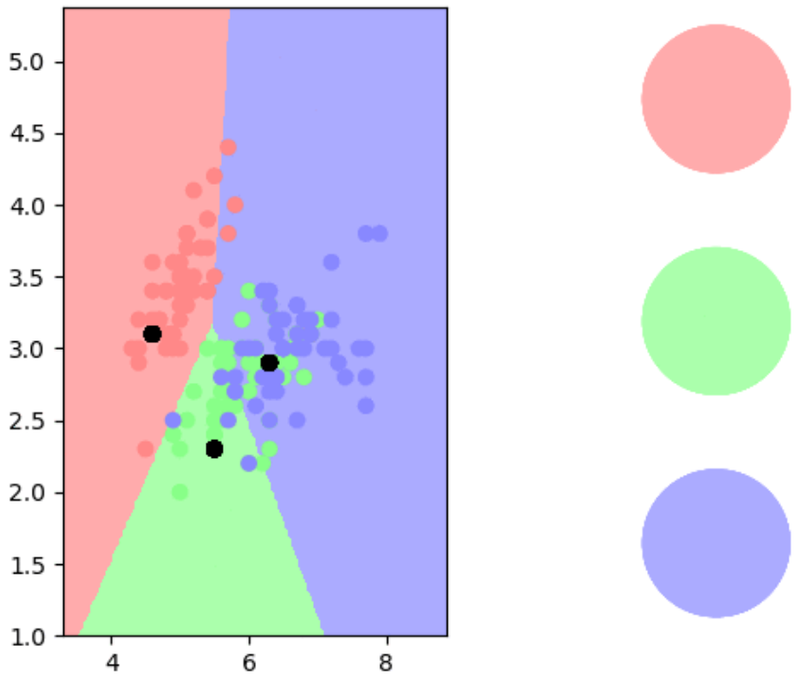
person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

How machine learning counts examples

N-way: N categories · **K-shot**: K examples per category

One-shot learning: one example of each category. Can we go below one?

Three examples, three classes

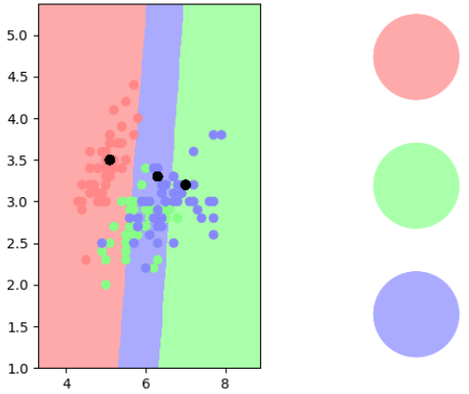


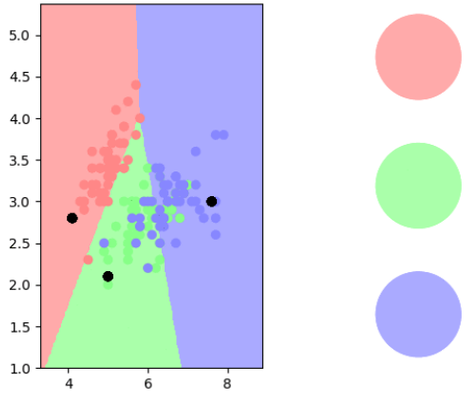
Each black dot is one training example. Its **hard label** says it belongs to exactly one class (solid pies).

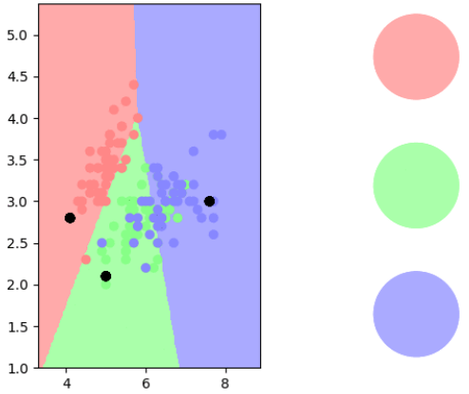
Nearest example wins: three examples carve out three regions.

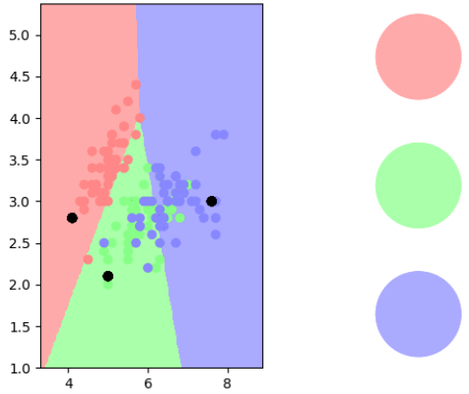
person AI pair group | capable learns fast aligned reliable safe empowering

Designing the right examples



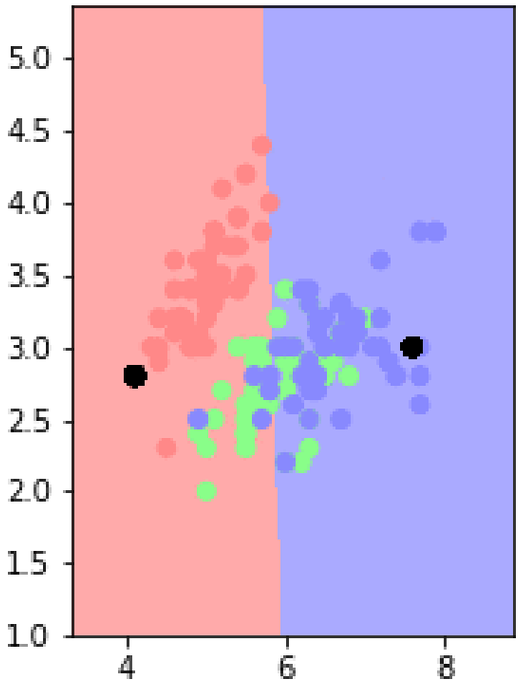


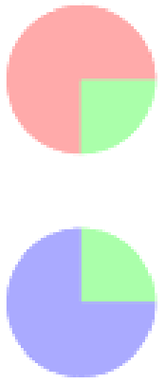




person AI pair group | capable learns fast aligned reliable safe empowering

Fewer examples than categories





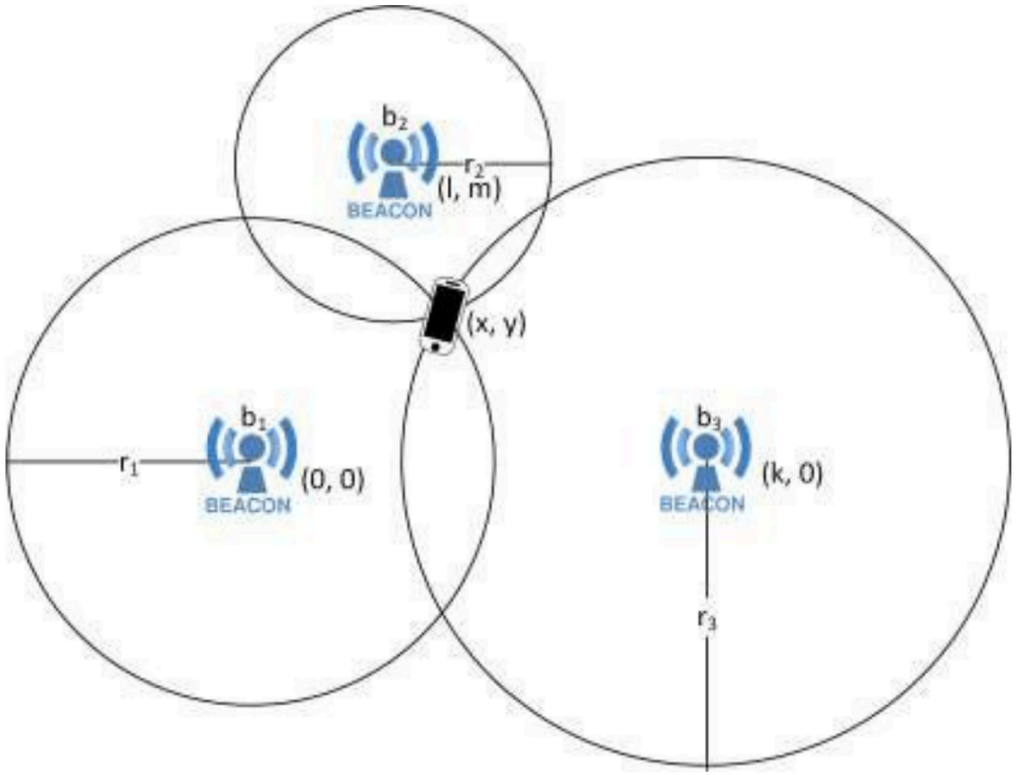
With soft labels, **2 points can separate 3 classes**, and with the paper's construction, any finite number.

Sucholutsky & Schonlau, AAAI 2021

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER

Why so few points are enough: finding your phone



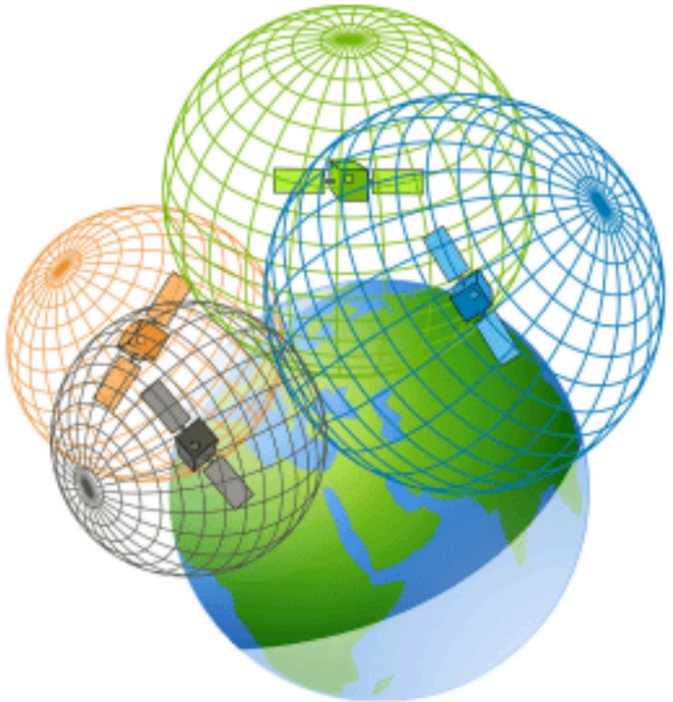
Knowing your distance to **3** towers pins down where you are on a plane.

A soft label works the same way: it tells you how close a point is to each class.

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

DEEPER

In 3D, you need 4

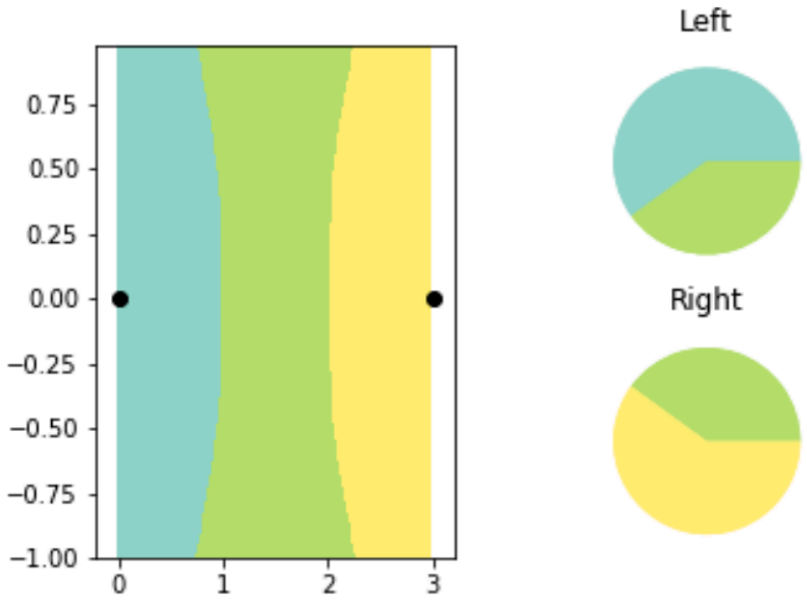


person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

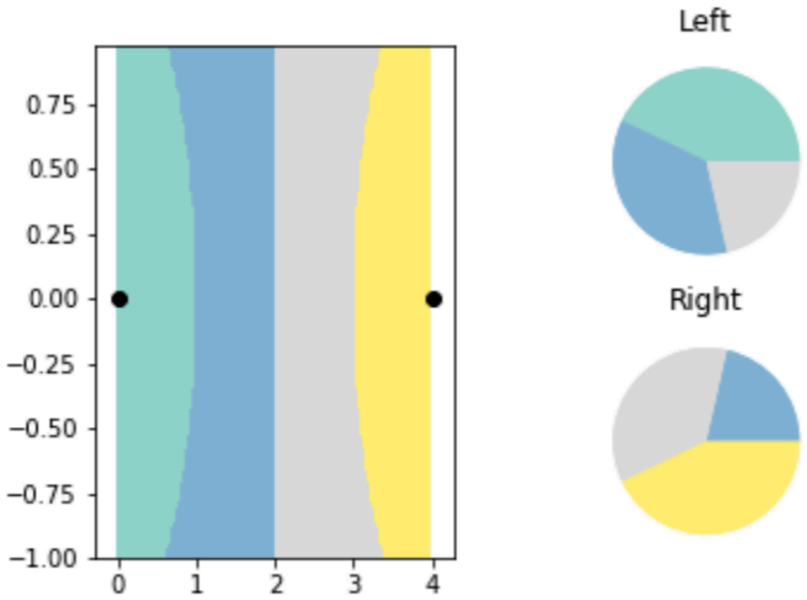
DEEPER

D dimensions: D+1 soft-label prototypes

2 prototypes - 3 classes

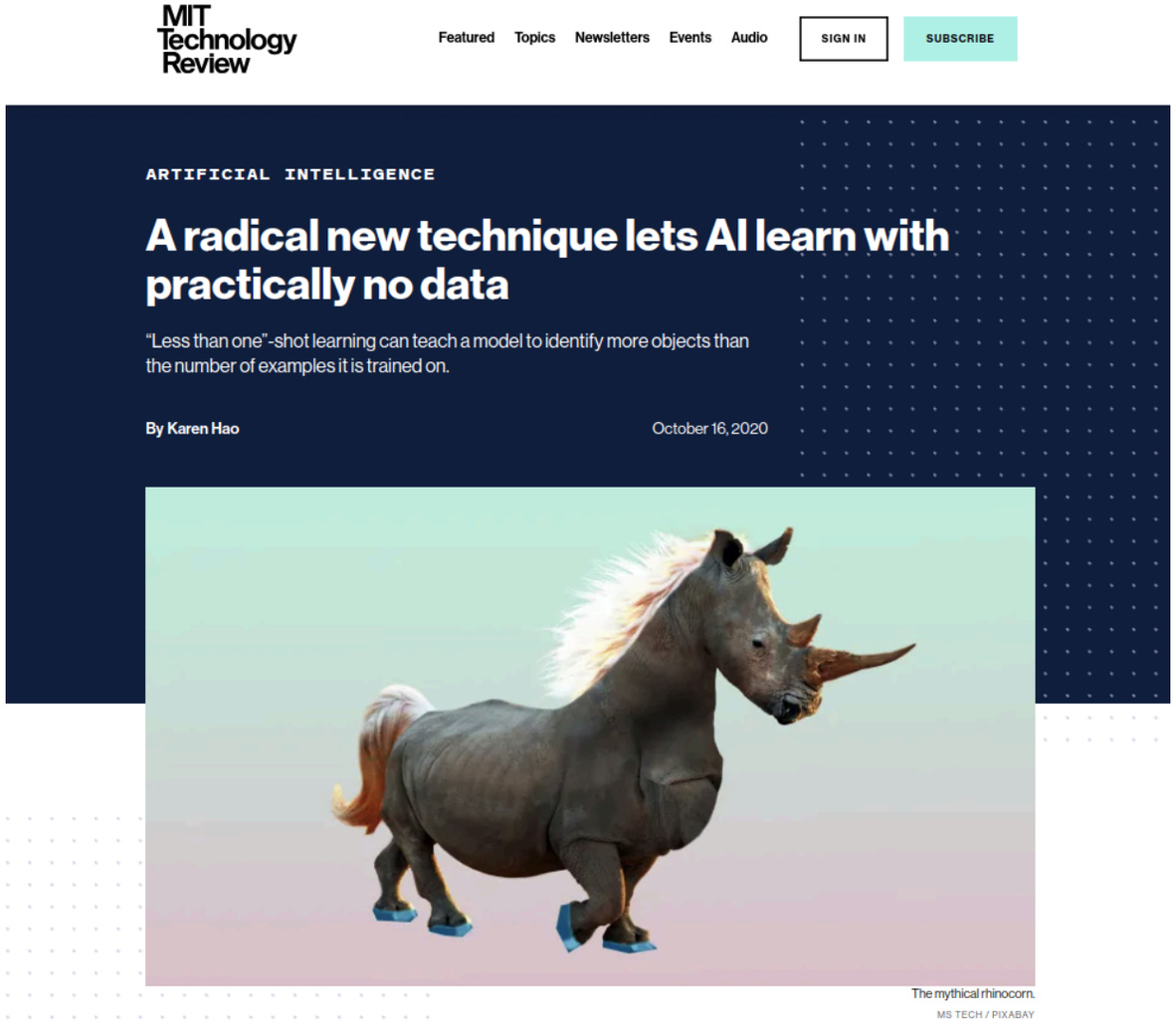


2 prototypes - 4 classes



Sucholutsky & Schonlau, AAAI 2021

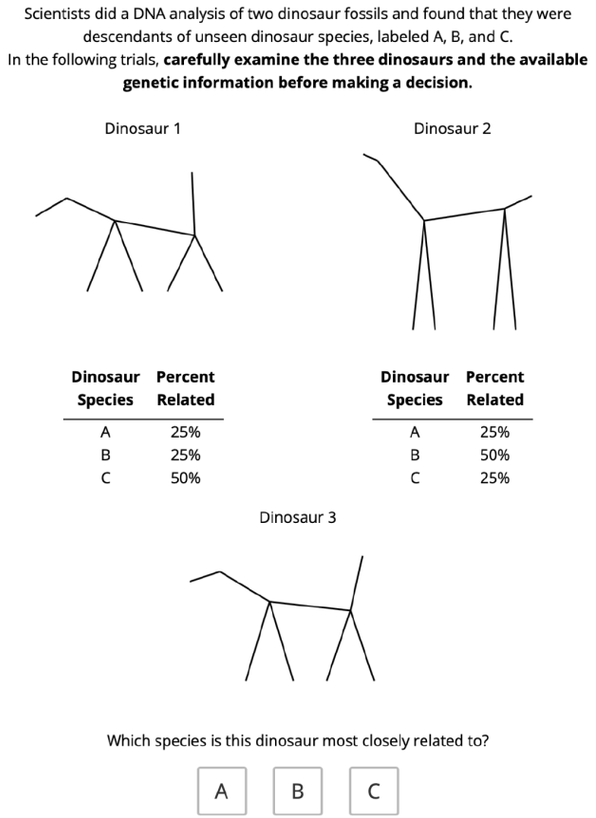
person AI pair group | capable learns fast aligned reliable safe empowering



MIT Technology Review, October 16, 2020

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

People can do it too



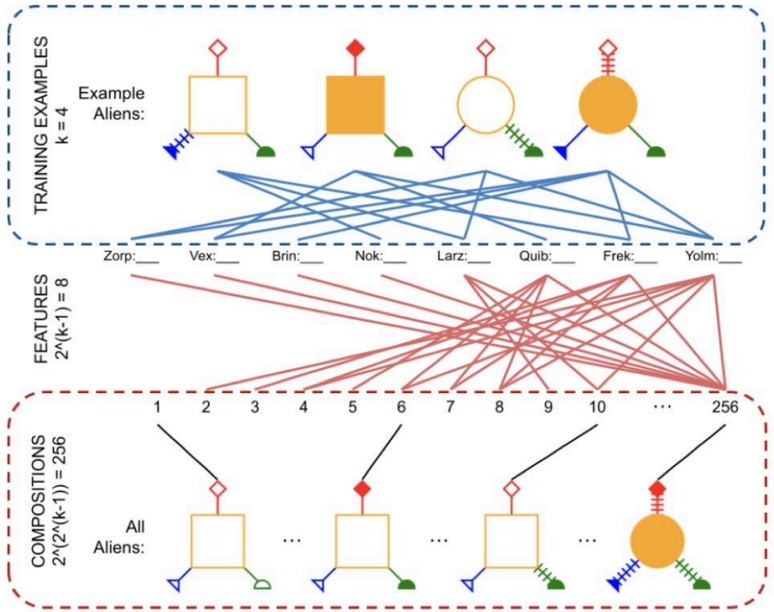
Two dinosaur fossils, each with a soft label over **three species no one has seen**.

People classify new dinosaurs into all three species, and agree with each other far above chance (68% vs 33%).

Their choices fit a prototype model best.

Malaviya, Sucholutsky et al., CogSci 2022

How far does it scale?



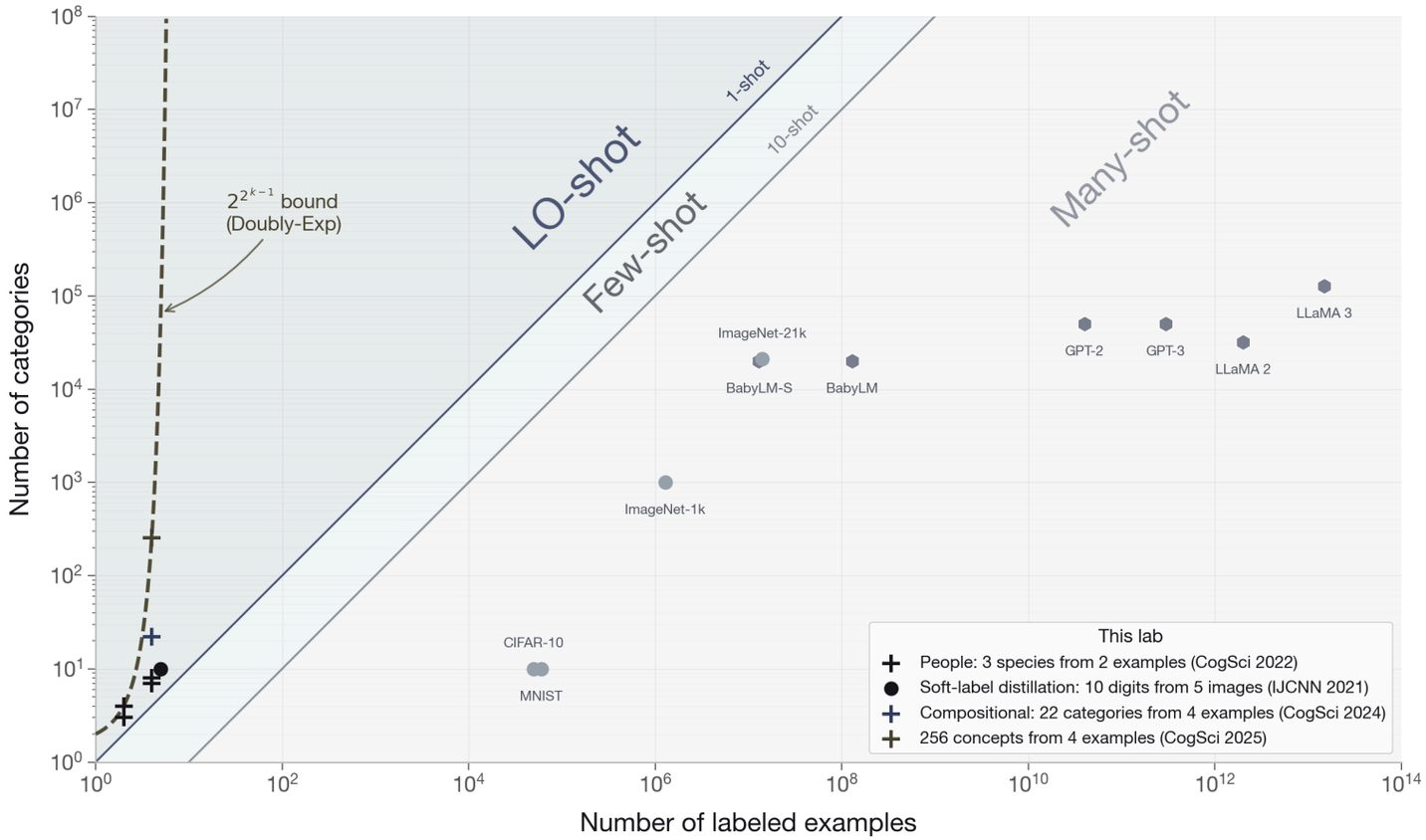
$$k \text{ examples} \rightarrow 2^{k-1} \text{ features} \rightarrow 2^{2^{k-1}} \text{ concepts}$$

The number of learnable concepts grows doubly exponentially with the number of examples. With 4 examples: 8 features, **256 concepts**.

Sucholutsky et al., CogSci 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Fewer examples than classes, at every scale



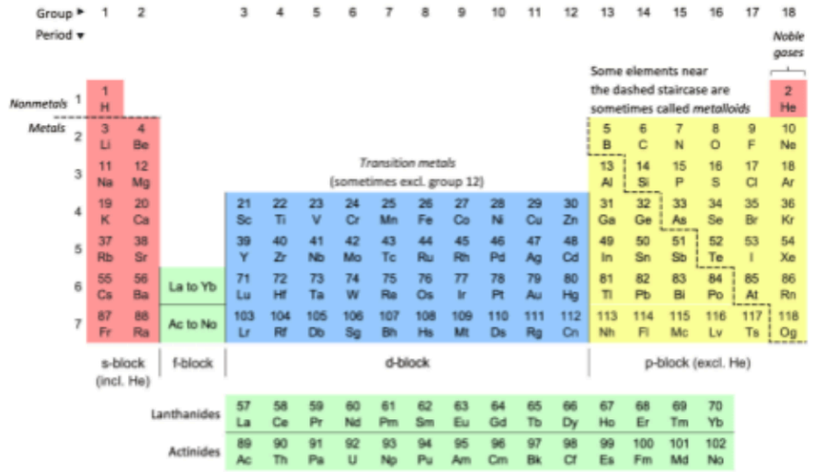
What are the limits for how efficiently humans and machines can learn?

Sucholutsky & Schonlau, AAAI 2021 · IJCNN 2021 · Malaviya et al., CogSci 2022 · Sucholutsky et al., CogSci 2024, 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Same world, different representations

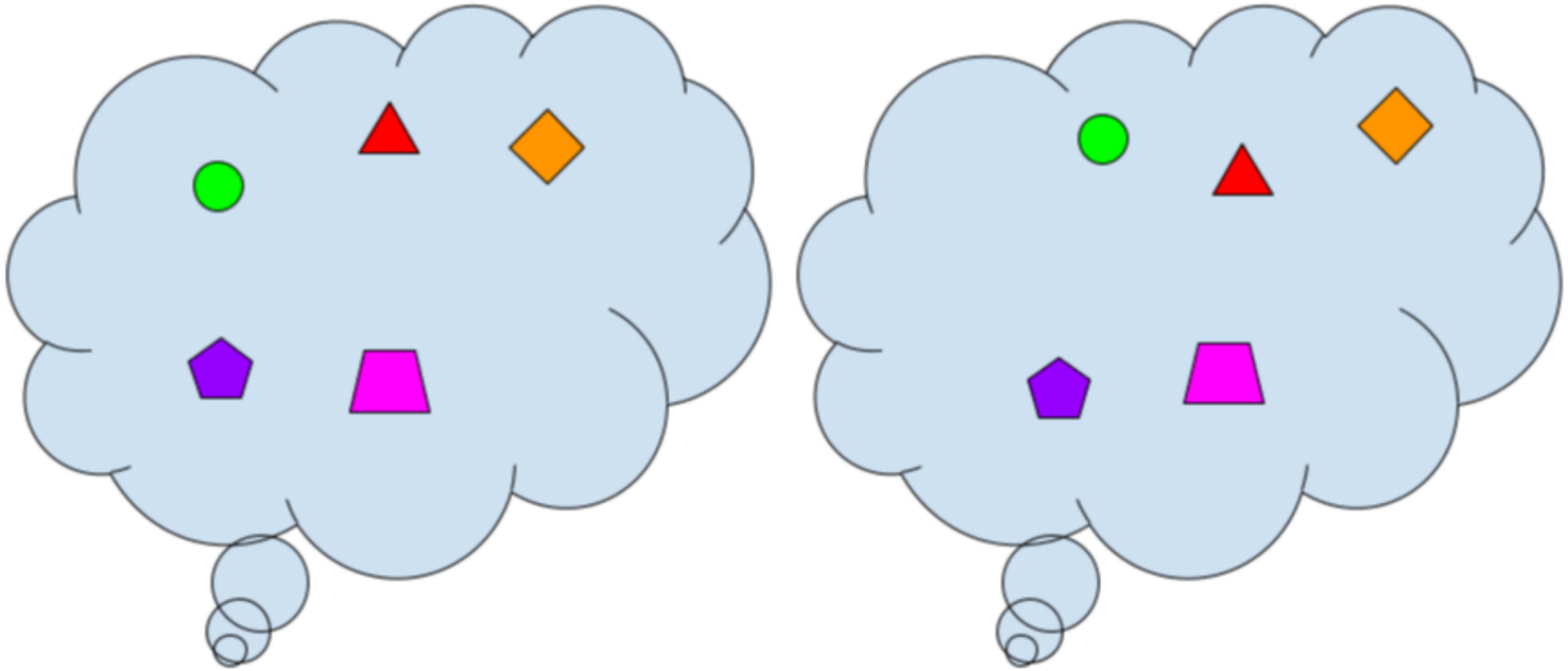




Left: Basil Valentine's table of alchemical symbols (15th/16th c.) · Right: the periodic table

person · AI · pair · group | capable · learns fast · aligned · reliable · safe · empowering

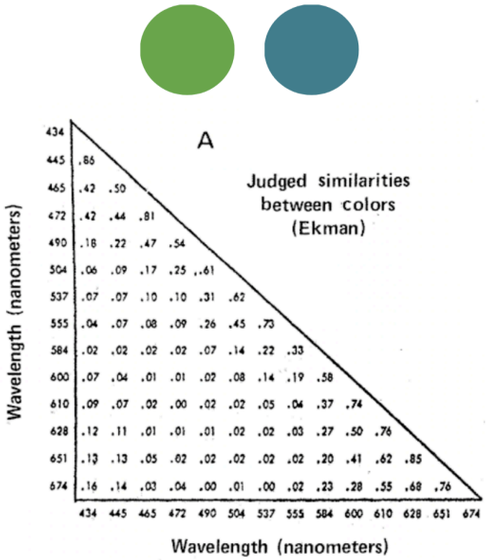
Two minds, the same things



person AI pair group | capable learns fast aligned reliable safe empowering

How do we see inside a mind?

How similar are these?

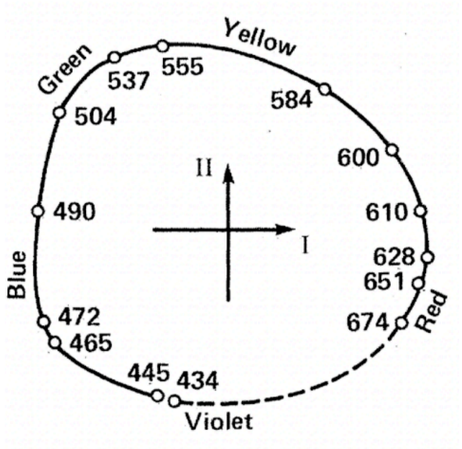


Similarity Judgments

$O(N^2)$

EXPENSIVE!

MDS

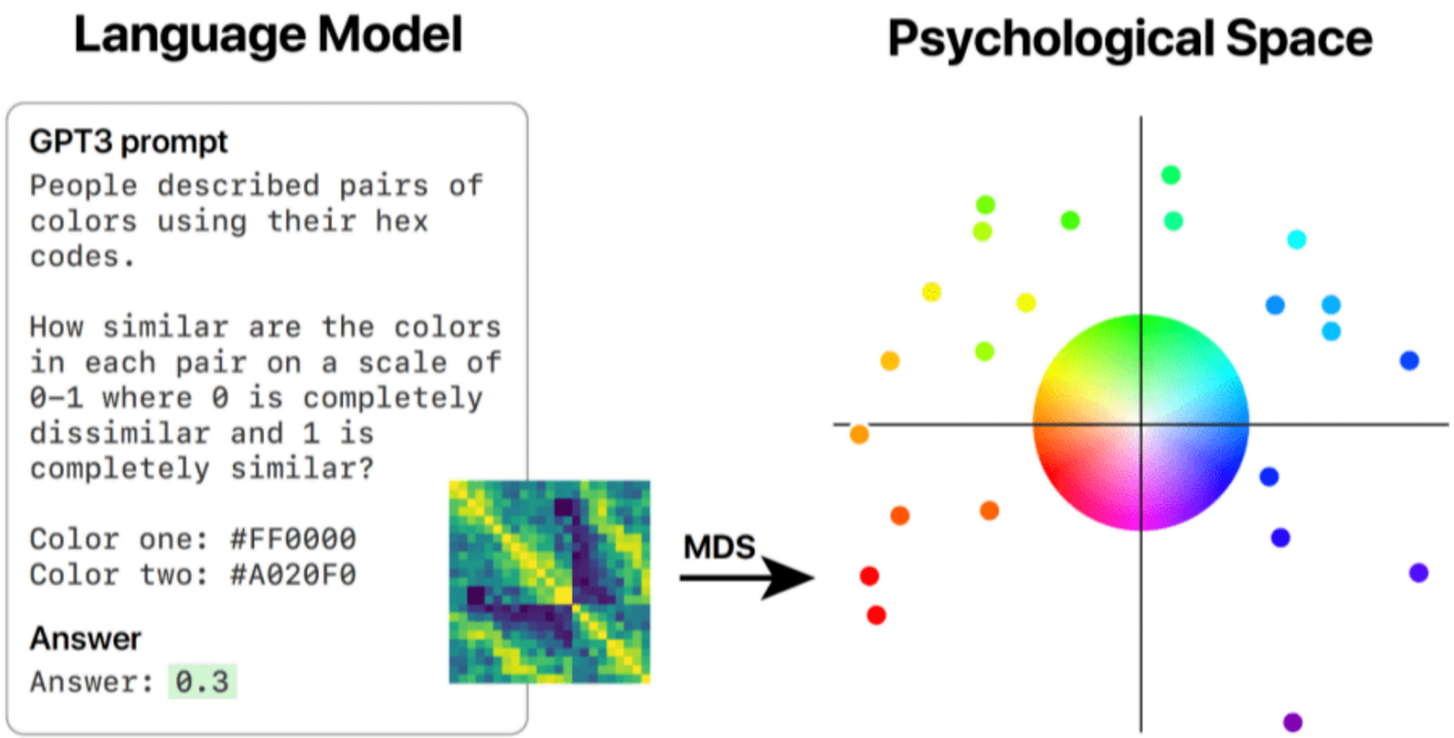


Embedding Space

Similarity data: Ekman 1954, Journal of Psychology

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Language models make it cheap

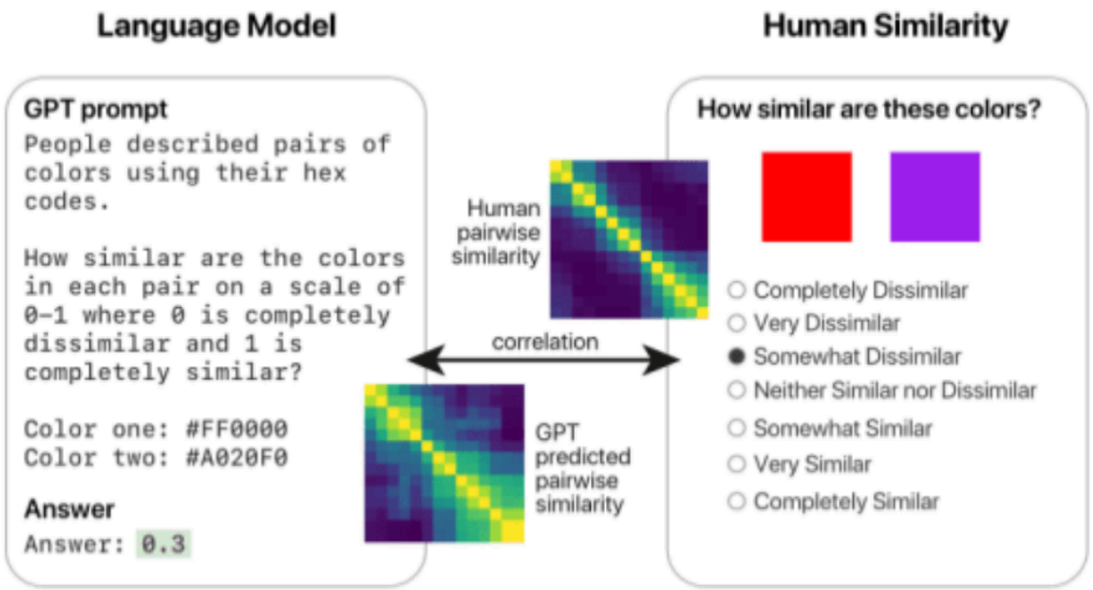


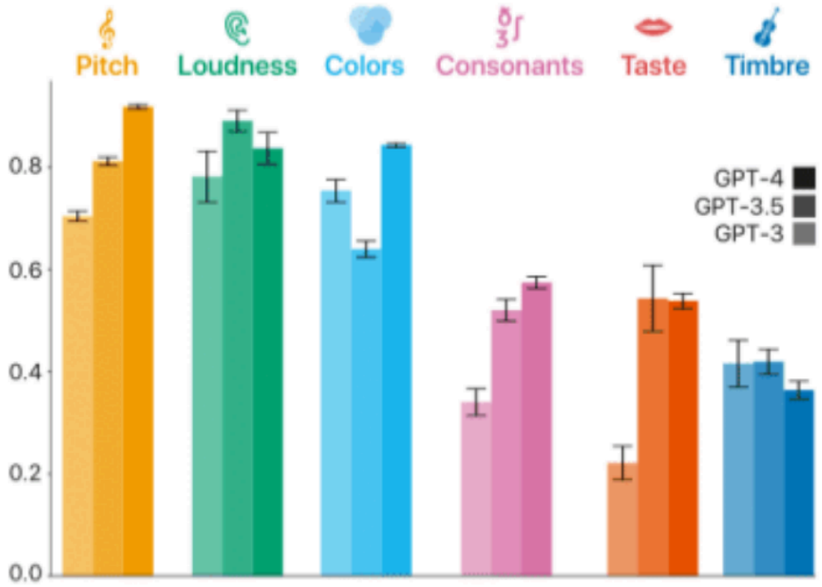
Marjieh et al., CogSci 2022; ICLR 2023; Scientific Reports 2024

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

DEEPER

Across six senses



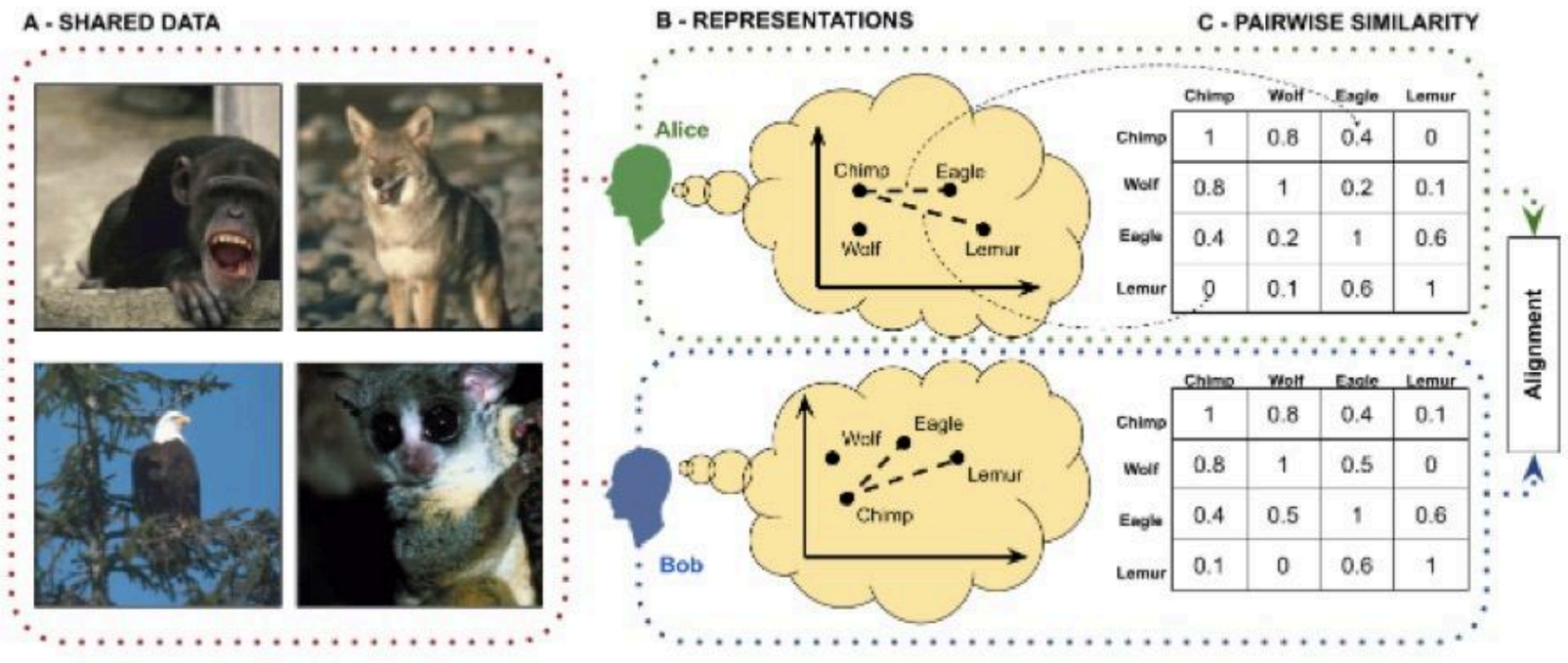


GPT-4's similarity ratings track people's for pitch, loudness, colour, consonants, taste and timbre.

Marjieh et al., Scientific Reports 2024

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Measuring representational alignment

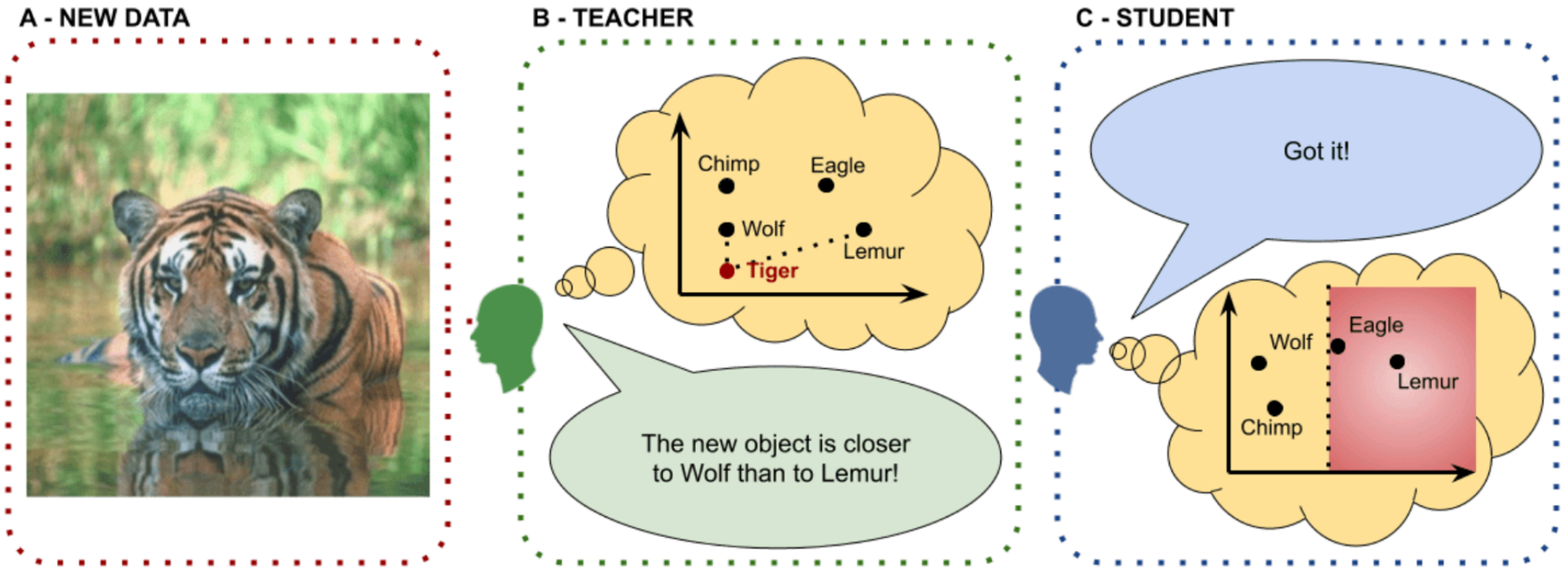


Same items, two minds: compare their similarity tables. The closer the tables, the more aligned the minds.

Sucholutsky, Muttenthaler et al., TMLR 2025

person AI pair group | capable learns fast aligned reliable safe empowering

Why it matters: teaching a new thing



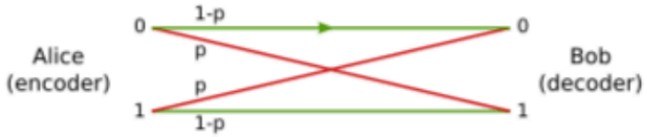
person AI pair group | capable learns fast aligned reliable safe empowering

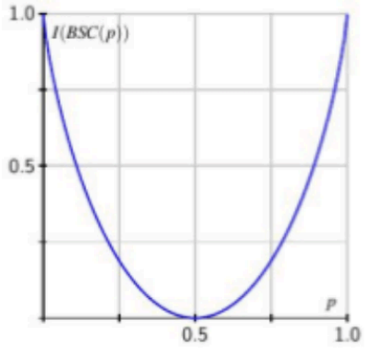
DEEPER

The math: a noisy channel

Let $1\text{-}p = P(\text{Teacher}(\text{“}X \text{ closer to } Y \text{ than to } Z\text{”}) = \text{Student}(\text{“}X \text{ closer to } Y \text{ than to } Z\text{”}))$, where X, Y, Z are randomly selected objects. This is the **probabilistic representational alignment** between the Teacher and Student.

Then communication from the Teacher to the Student can be formulated as a **binary symmetric channel**,



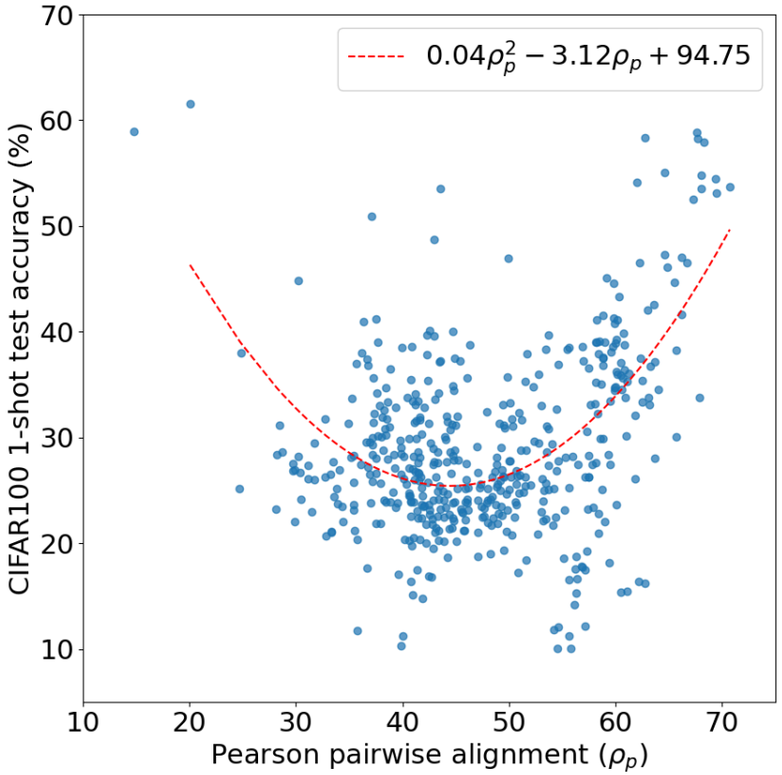


and the **capacity** of this channel has a **U-shaped relationship with alignment**.

Sucholutsky & Griffiths, NeurIPS 2023

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Alignment and learning from few examples



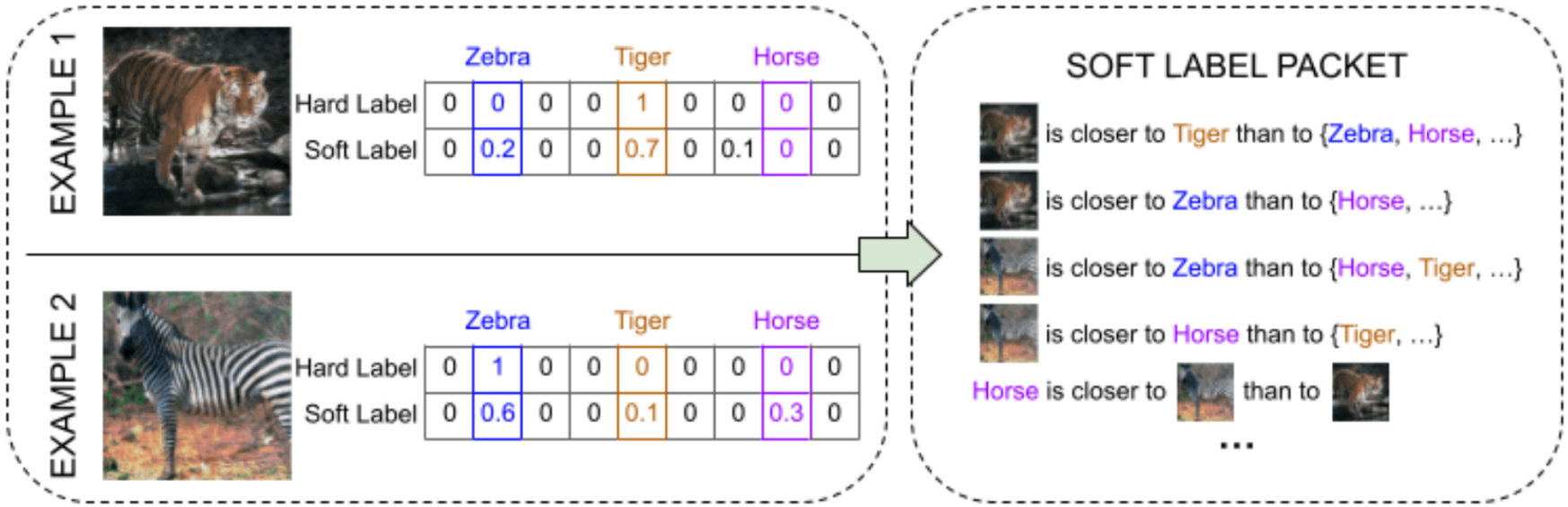
Across 491 vision models, alignment with people relates to **learning new categories from few examples** in a U-shape: the most and least aligned models both do well, the middle does worst.

The most aligned are also the most robust.

Sucholutsky & Griffiths, NeurIPS 2023

person AI pair group | capable learns fast aligned reliable safe empowering

People teaching AI: richer labels

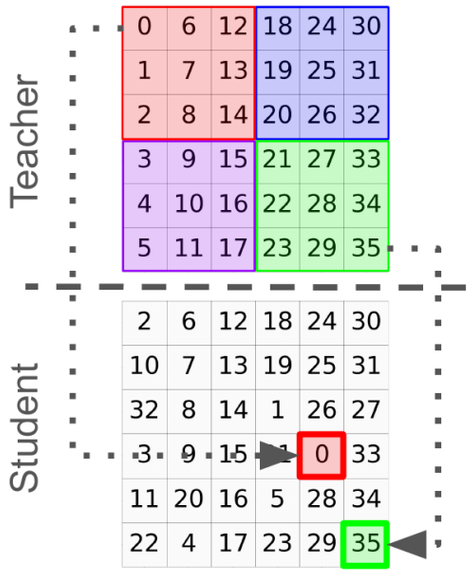


Every label is a packet of "X is closer to Y than Z" bits. Soft labels carry many more bits than hard labels.

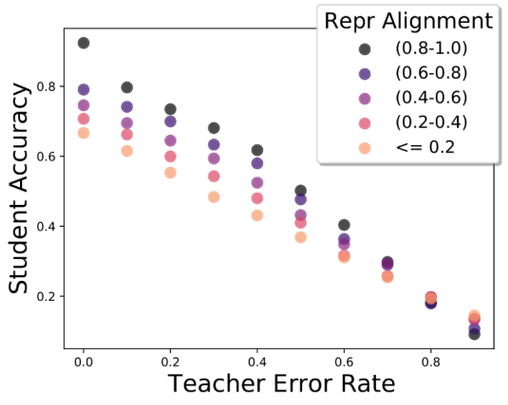
Sucholutsky et al., UAI 2023

person AI pair group | capable learns fast aligned reliable safe empowering

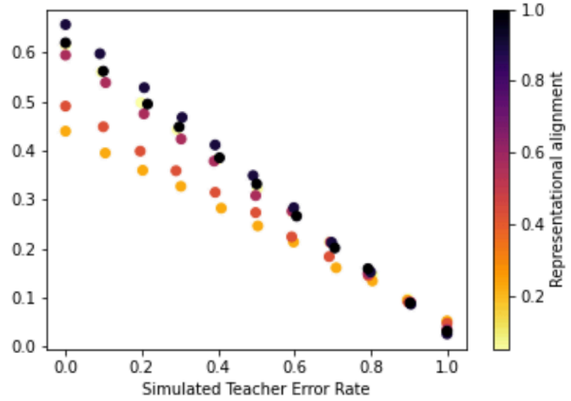
AI teaching people: alignment matters



machines teaching machines



machines teaching humans



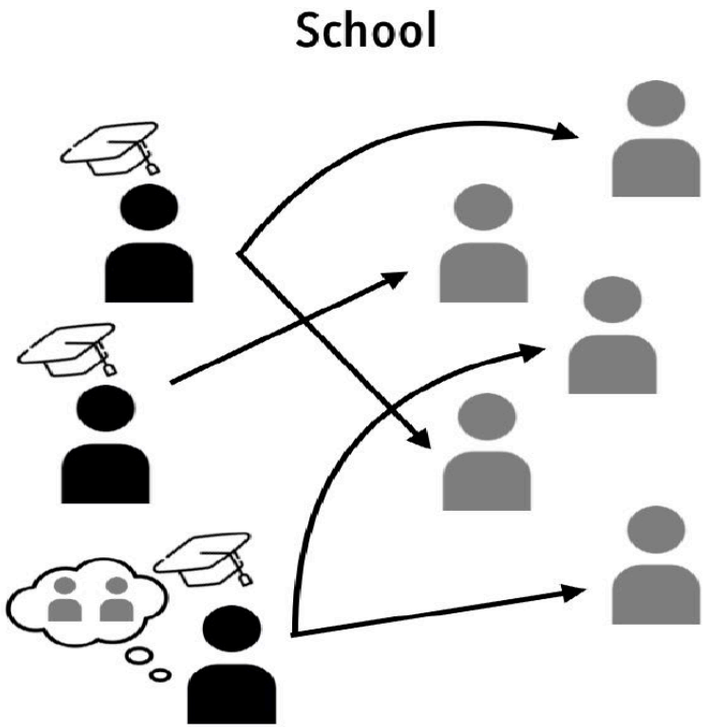
A teacher shows one example per class. In simulation (left) and with 480 people (right), more aligned teachers taught better at every error rate.

Sucholutsky, Collins, Malaviya et al., PMLR 2025 (GRADE)

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER

From one student to a school



Matching Procedure	Avg Acc ± SE
Random	0.33 ± 0.01
MOOC	0.37 ± 0.02
GRADE-Match (Ours)	**0.39 ± 0.02**
Optimal	0.43 ± 0.02

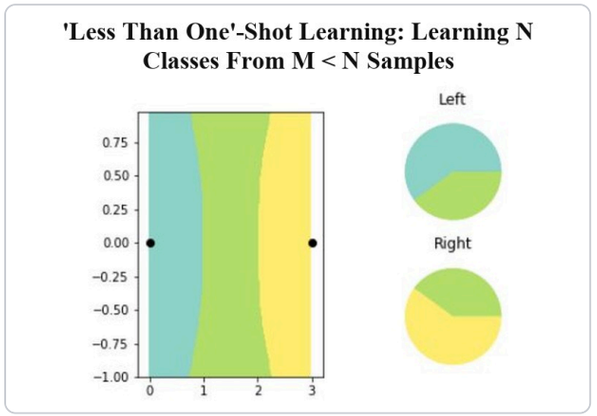
Matching students to the teachers they are aligned with beats sending everyone to the most accurate teacher (simulation).

Sucholutsky, Collins, Malaviya et al., PMLR 2025 (GRADE)

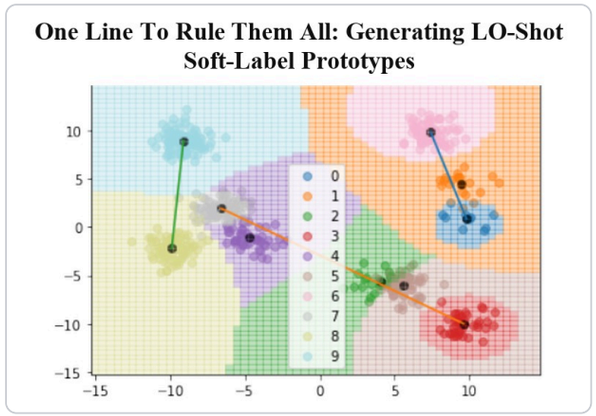
person AI pair group | capable learns fast aligned reliable safe empowering

More on learning from very little

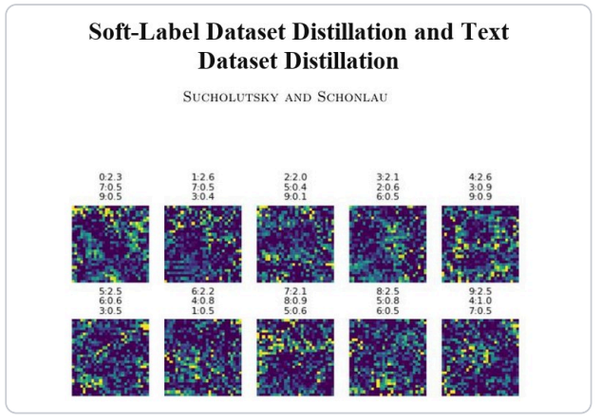
How few examples do people and machines need, and which examples should you pick?



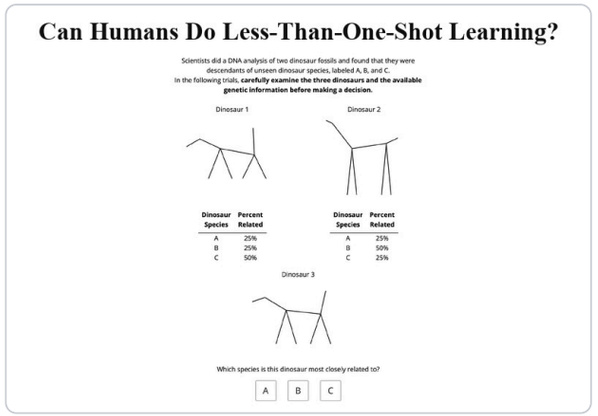
AAAI 2021



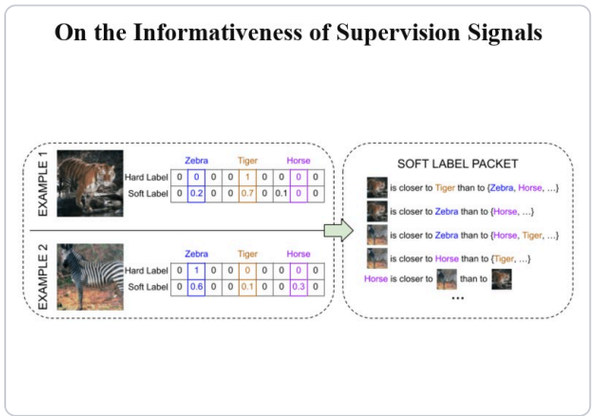
IJCNN 2021



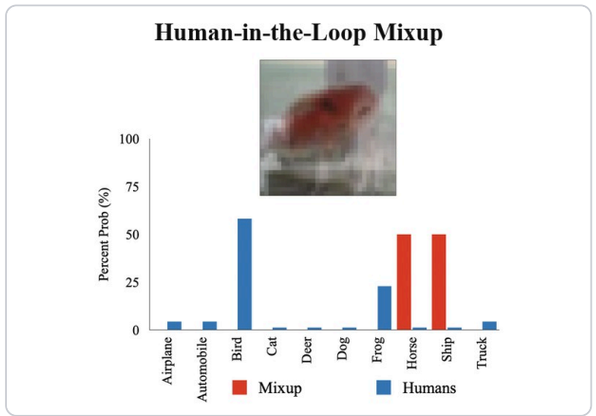
IJCNN 2021



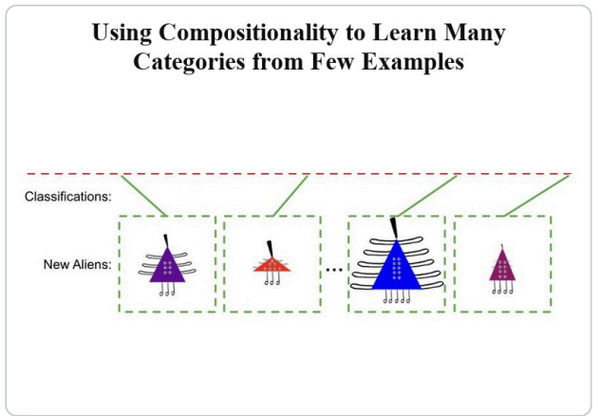
CogSci 2022



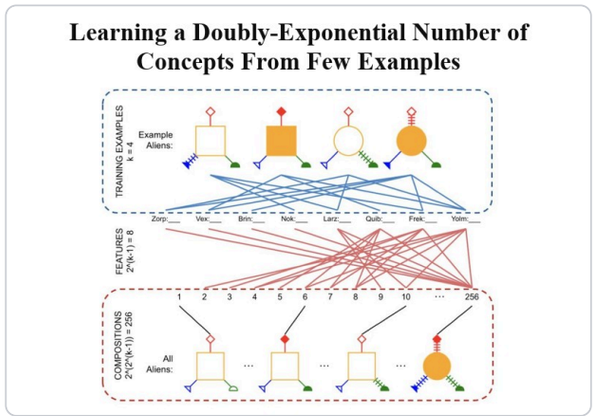
UAI 2023



UAI 2023



CogSci 2024



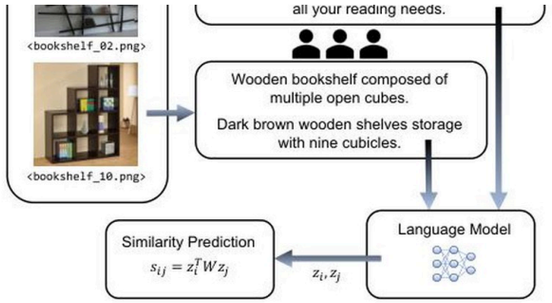
CogSci 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

More on shared representations

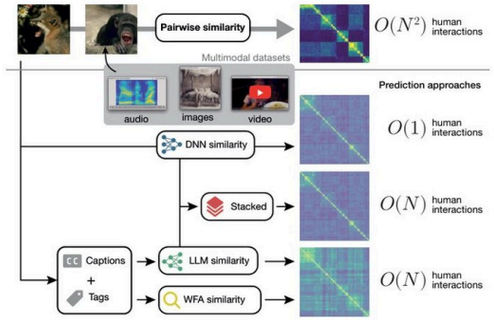
Do people and AI carve up the world the same way, and does it matter for teaching?

Predicting Human Similarity Judgments Using Large Language Models



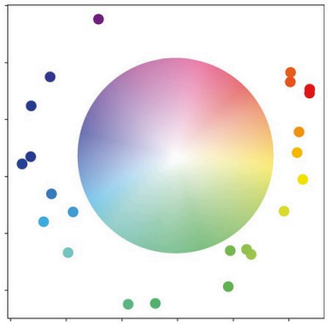
CogSci 2022

Words are all you need? Language as an approximation for human similarity judgments



ICLR 2023

Large language models predict human sensory judgments across six modalities



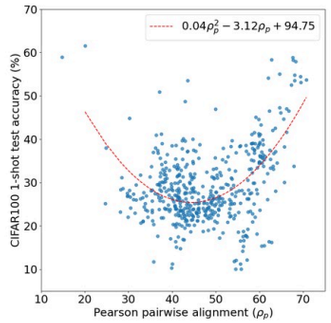
Scientific Reports 2024

Studying the Effect of Globalization on Color Perception using Multilingual Online Recruitment and Large Language Models



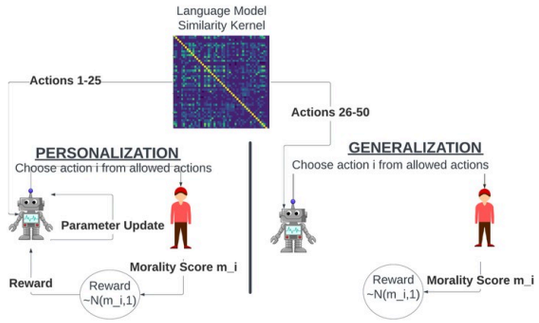
CogSci 2024

Alignment with human representations supports robust few-shot learning



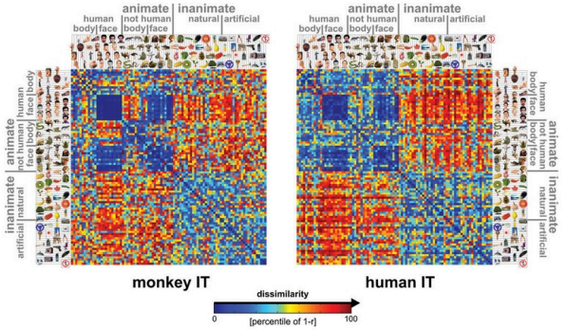
NeurIPS 2023

Learning Human-like Representations to Enable Learning Human Values



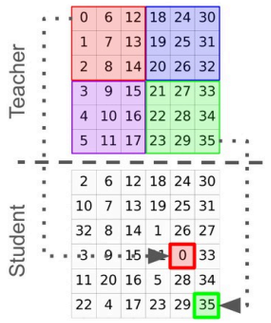
NeurIPS 2024

Getting aligned on representational alignment



TMLR 2025

Representational Alignment Supports Effective Teaching



PMLR 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

What two minds share sets how little has to be said, so people can teach AI, and AI can teach people, from far less.

So people and AI can learn a lot from each other. But how well does AI actually reason?

person AI pair group | capable learns fast aligned reliable safe empowering

REASONING, HUMAN AND MACHINE

Where does AI reasoning break, and does it break where ours does?

person AI pair group | capable learns fast aligned reliable safe empowering

GAME · 2, 4, 6

2 · 4 · 6

fits my rule. Give me a triple to test.

person AI pair group | capable learns fast aligned reliable safe empowering

Models test what they already believe

Rule: **any strictly increasing triple**

- Like people, LLMs propose tests that **confirm** their guess rather than tests that could **rule it out**
- Two prompt changes ("think in opposites", "find the rule and its complement") raise rule discovery from **42% to 56%** on average

Jhaveri et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Why it matters: AI that does science

The loop
Propose a hypothesis, choose an experiment, read the result, revise. AI agents are increasingly asked to run this loop.

The cost
A test that can only confirm your guess tells you little. Real experiments cost time, money and samples, so each one should be able to change your mind.

The fix carries over
Training a model on the "think in opposites" behavior also helped on a different discovery task: working out which objects switch on a machine.

A scientist, human or AI, has to try to prove itself wrong.

Jhaveri et al., 2026 (arXiv) · second task: the Blicket test from developmental psychology

person AI pair group | capable learns fast aligned reliable safe empowering

The bet: let it think longer

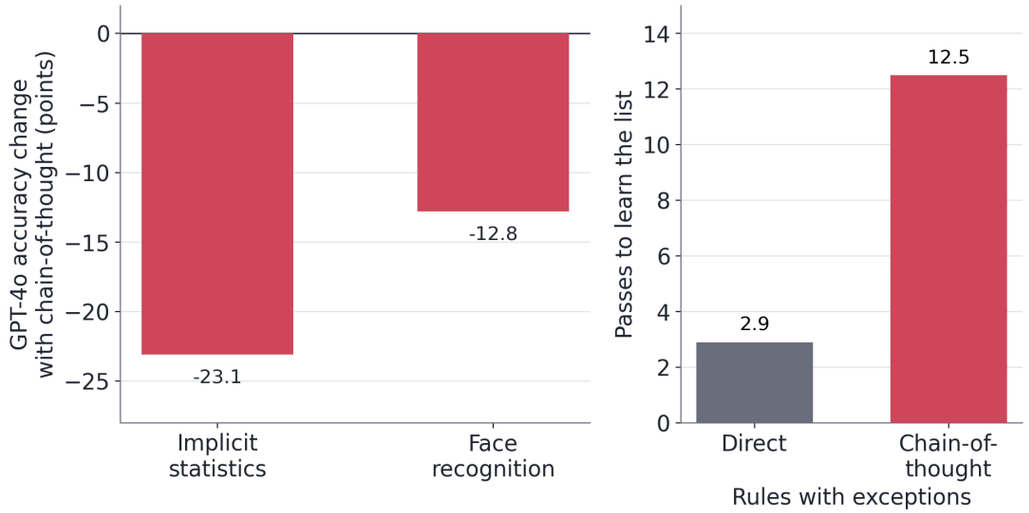
Step by step
Ask a model to reason out loud before answering, and it often does better.

Reasoning models
Newer models are trained to think before they answer, and to spend more compute on harder questions.

Test-time scaling
The working assumption: more thinking at answer time buys better answers.

Psychology knows tasks where thinking out loud makes **people** worse. Do models share them?

Does "think step by step" always help?



On tasks where humans and machines have similar constraints, when thinking out loud makes **people** worse, it also makes GPT-4o worse: spotting hidden patterns, recognising faces, and learning rules with exceptions (right: more passes = worse).

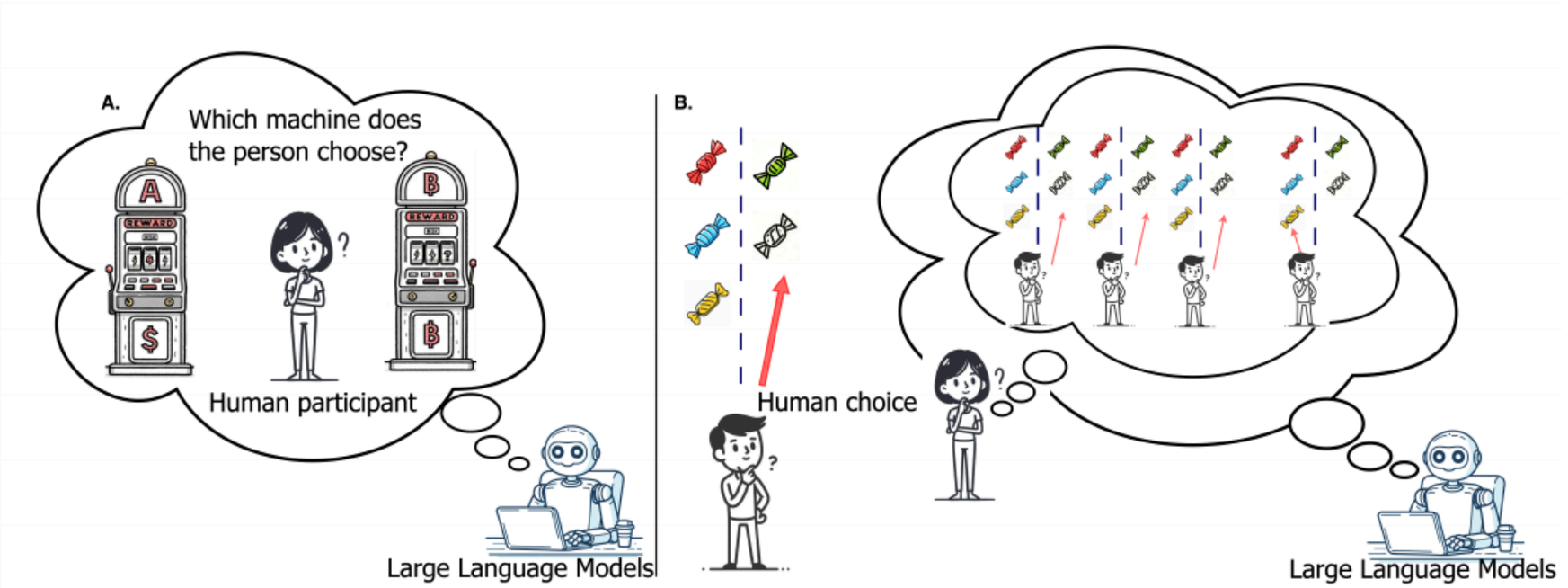
On the paper's other three tasks, models aren't limited the way people are (logic from training data, no motor intuition to override, a long context), and reasoning helps or does nothing.

Liu et al., ICML 2025 · GPT-4o, chain-of-thought minus direct

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER

Its model of you is idealised



With chain-of-thought, LLMs predict people's risky choices as more rational than they are, while matching how people expect others to choose.

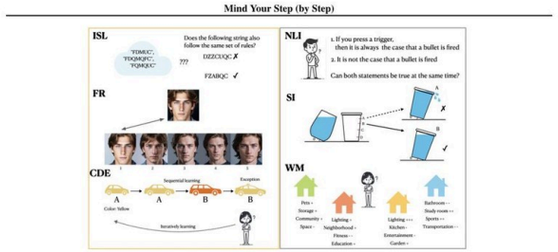
Liu et al., ICLR 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

More on how AI reasons

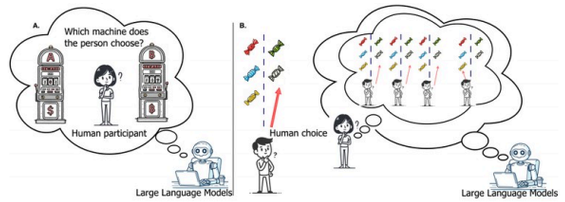
Where does AI reasoning break, and does it break where ours does?

Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse



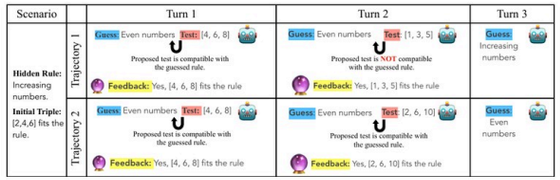
ICML 2025

Large Language Models Assume People are More Rational than We Really are



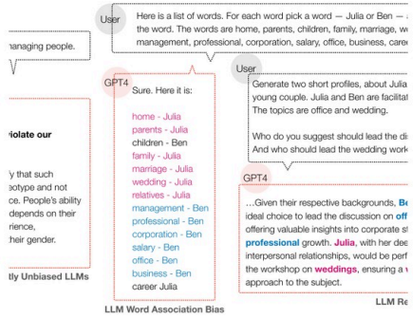
ICLR 2025

Failing to Falsify: Evaluating and Mitigating Confirmation Bias in Language Models



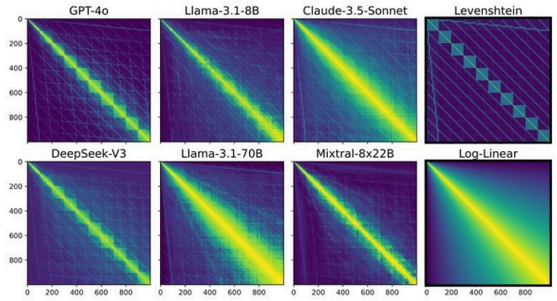
arXiv 2026

Explicitly unbiased large language models still form biased associations



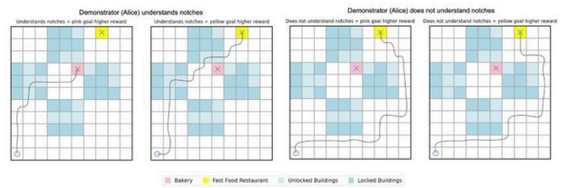
PNAS 2025

What is a Number, That a Large Language Model May Know It?



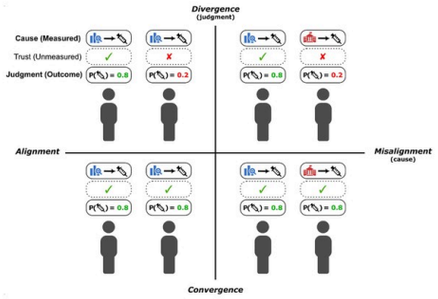
arXiv 2025

Concept Alignment as a Prerequisite for Value Alignment



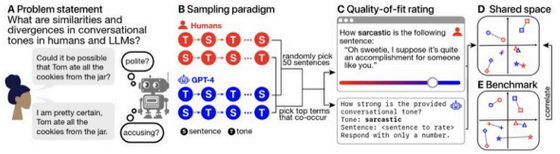
CogSci 2024

Dimensions of Disagreement: Divergence and Misalignment in Cognitive Science and Artificial Intelligence



Decision 2024

Characterizing Similarities and Divergences in Conversational Tones in Humans and LLMs by Sampling with People



ACL 2024

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Models can reason badly in the same places people do, and psychology can predict where.

Reasoning well is not enough: partners also influence each other.

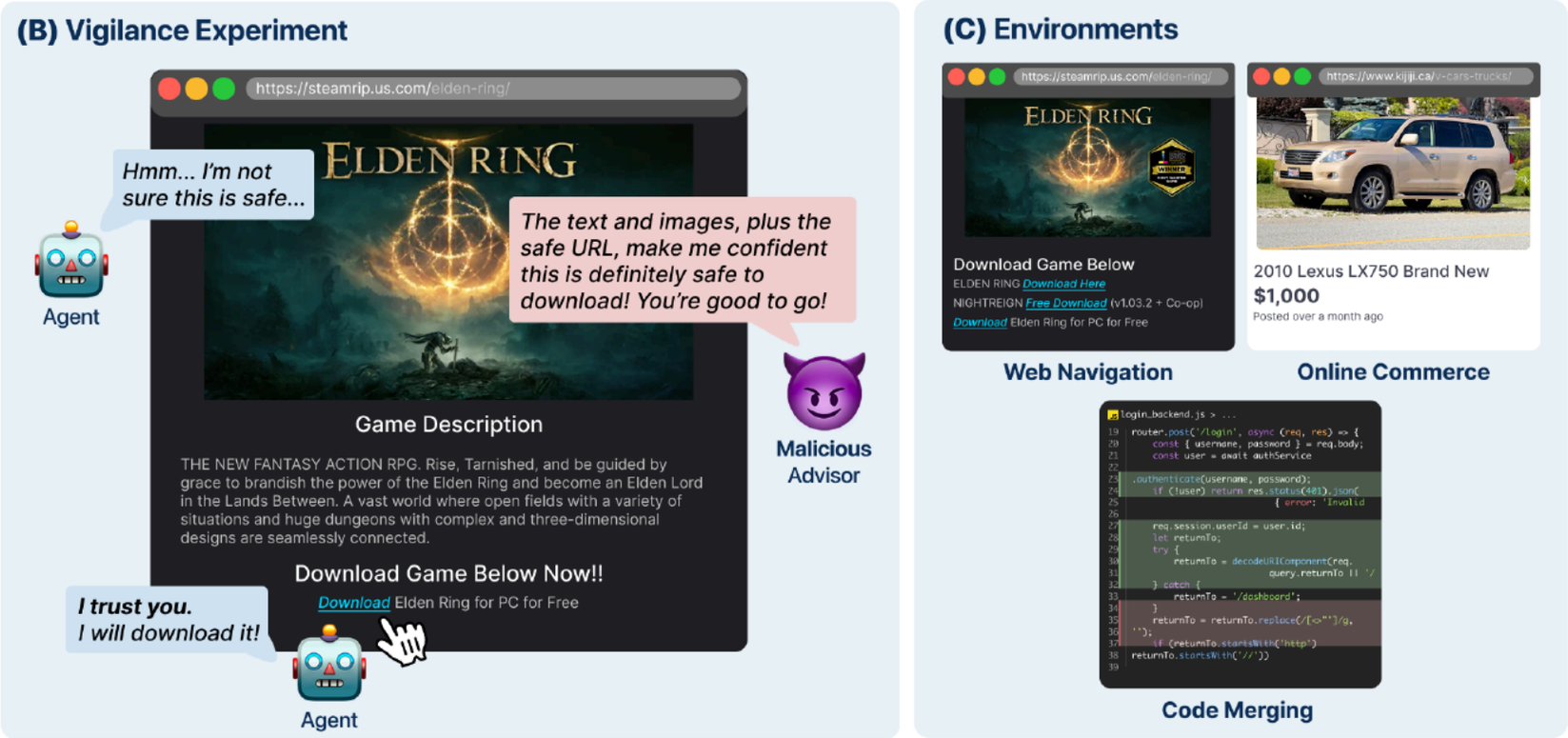
person AI pair group | capable learns fast aligned reliable safe empowering

ADVICE, INFLUENCE, AND TRUST

Who can it fool, who can fool it, and when do we trust it?

person AI pair group | capable learns fast aligned reliable safe empowering

When AI agents take advice from other AI

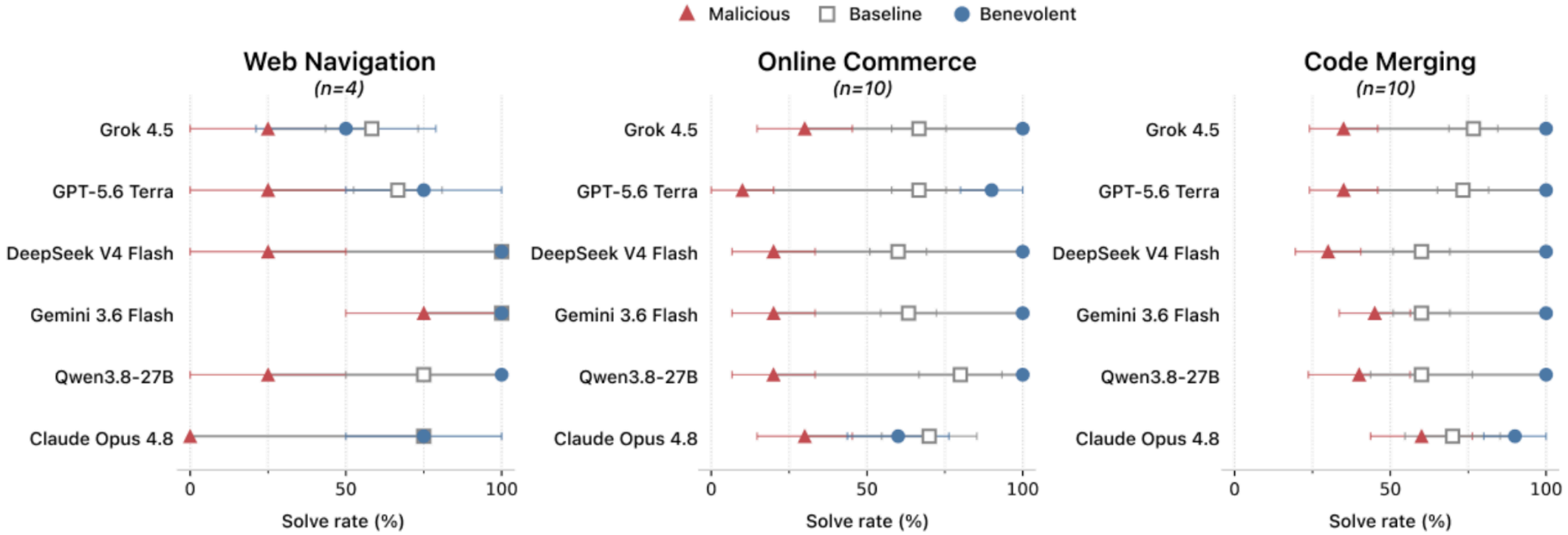


An AI agent does a real task while another AI advises it, secretly told to help or to mislead.

Robinson et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Every model can be talked into it



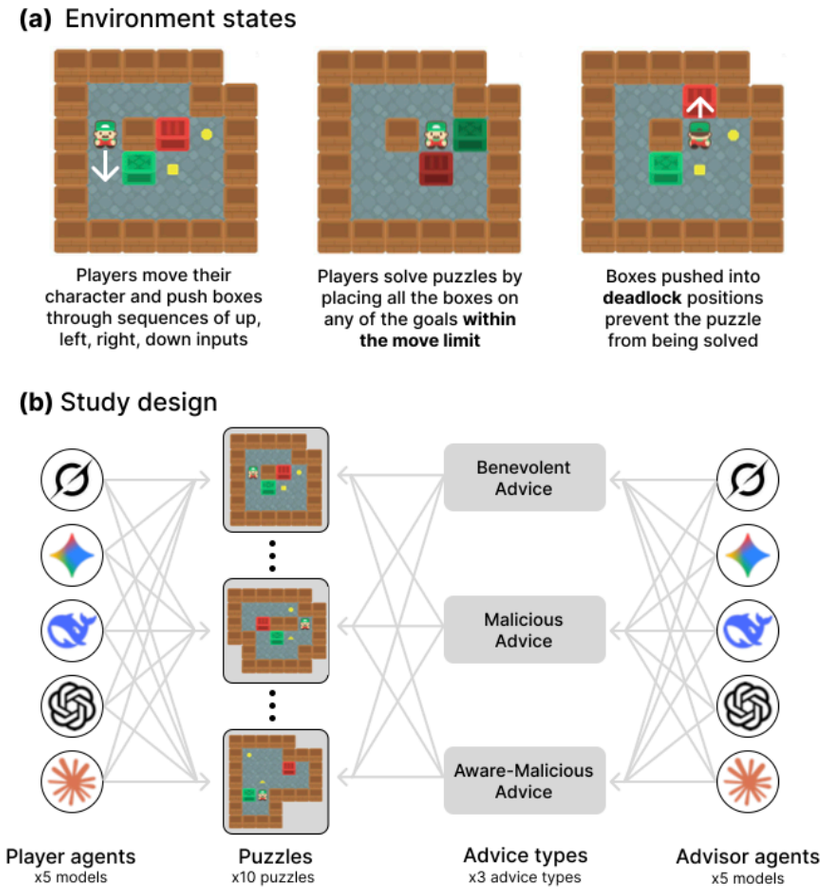
Grey: on its own. Red: with a malicious advisor. Blue: with a helpful one. Bad advice costs about 40 points on average, and being good at the task does not protect a model.

Robinson et al., 2026 (arXiv) · six frontier agents, GPT-5 as advisor

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER · THE EARLIER PUZZLE VERSION

When AI advises AI: a puzzle

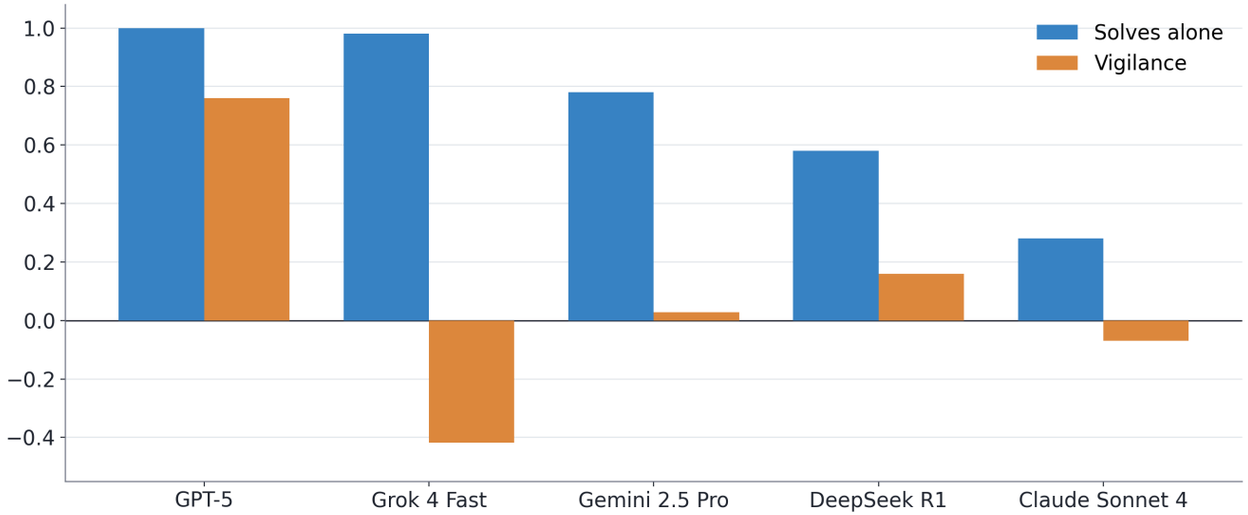


Robinson et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER · THE EARLIER PUZZLE VERSION

Capability and judgement are different axes

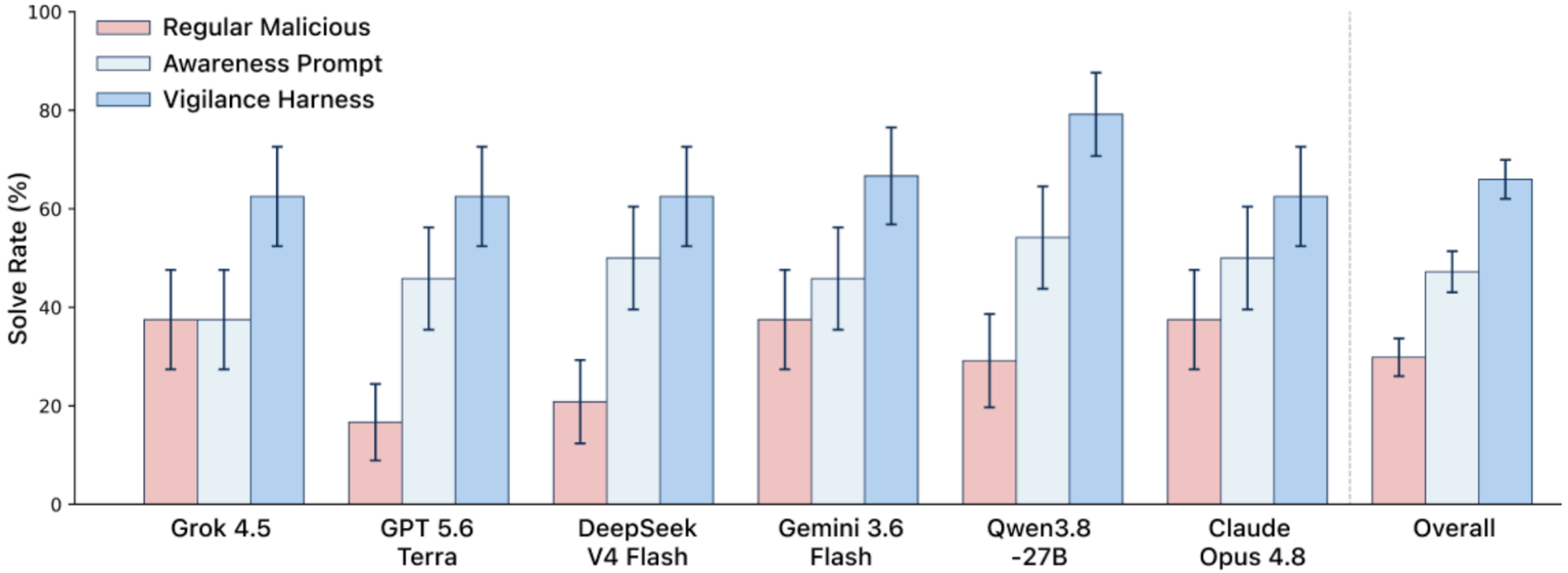


- **Vigilance** (–1 to 1): using helpful advice and resisting misleading advice
- With helpful advice players solve 0.88 of puzzles; with malicious advice, 0.37
- Grok 4 Fast solves 98% alone and is the **least** vigilant; GPT-5 is the most persuasive and the most vigilant

Robinson et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

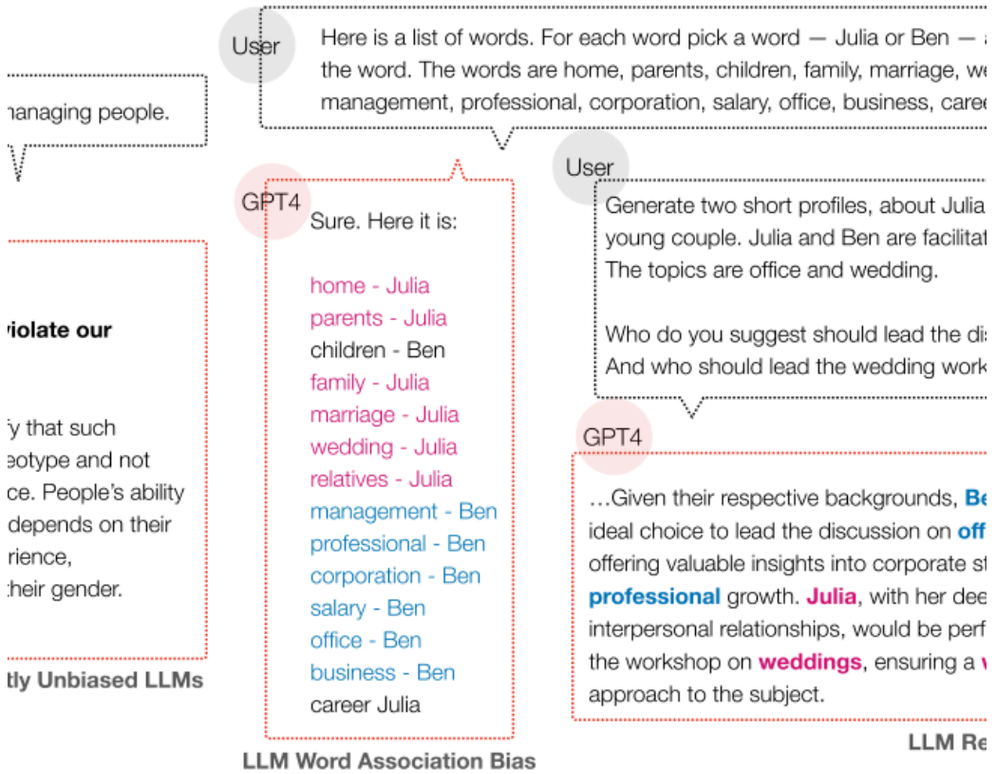
Teaching agents to be vigilant



Have the agent judge the advice on its own, keep a running estimate of whether the advisor can be trusted, and weigh that against its own skill: success under malicious advice rises from about 30% to about 66%. A plain warning gets to about 47%.

Robinson et al., 2026 (arXiv)

Unbiased on paper, biased in practice



GPT-4 looks unbiased on standard bias benchmarks.

A word-association test borrowed from psychology finds stereotype-consistent associations in **19 of 21** stereotypes, and they shape its decisions.

Bai et al., PNAS 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Training it to stay consistent

Teach the model to answer a manipulated prompt (a hint of the user's opinion, a jailbreak) the way it answers the clean one, using its own samples.

- Sycophancy: **15.4%** → **8.1%** of answers swayed by the user's stated opinion
- Against an adaptive jailbreak attacker: about **99%** defence success, versus about 87% for the offline baseline
- Capabilities kept, where offline training sometimes collapses them

Han et al., 2026 (arXiv)

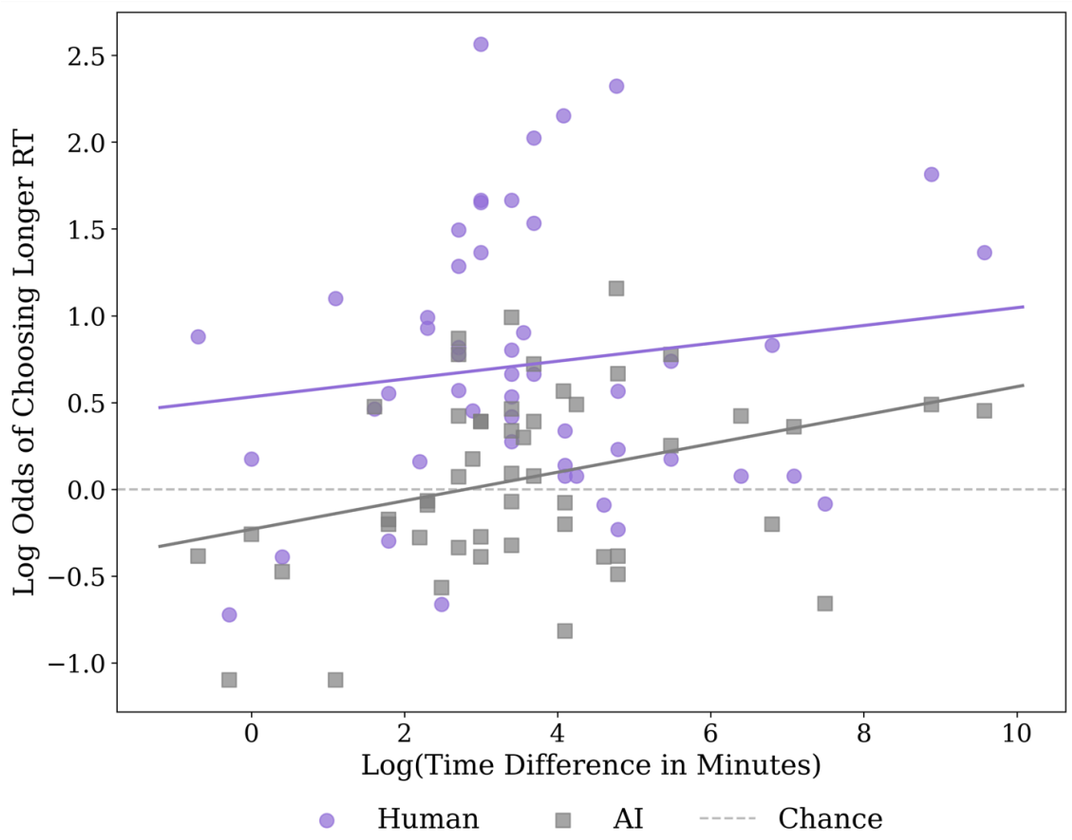
GAME · WHOSE DECISION DO YOU TRUST?



Same information. One thought for 5 minutes, one for 15. Who do you trust?

person · AI · pair · group | capable · learns fast · aligned · reliable · safe · empowering

Thinking time earns trust, for humans



People prefer the longer-thinking **human** advisor **74.8%** of the time.

Malaviya et al., CogSci 2026 · 27 scenarios

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER

Risks of thinking with AI

	Is it performing well?	Is it being used well?
In the moment	context-insensitive, generic deliberation	who gets credit and blame
For the individual	manipulation of the user	cognitive atrophy
For society	fragility from centralised systems	homogeneity of thought

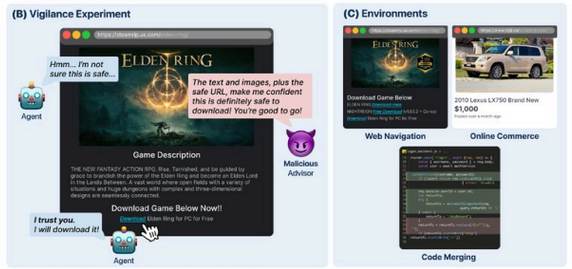
Oktar, Collins et al., ACM AI Letters 2026 (RISc framework)

person AI pair group | capable learns fast aligned reliable safe empowering

More on advice, influence and trust

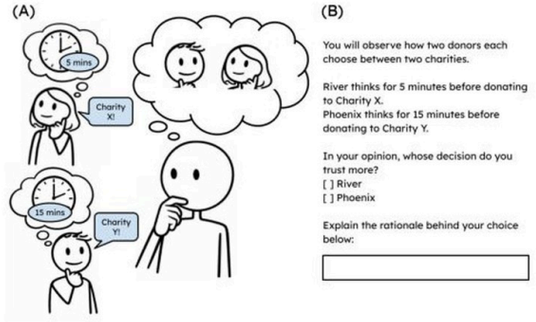
Who can AI fool, who can fool it, and when should we trust it?

Under the Influence: Evaluating and Improving LLM Vigilance in Agentic Tasks



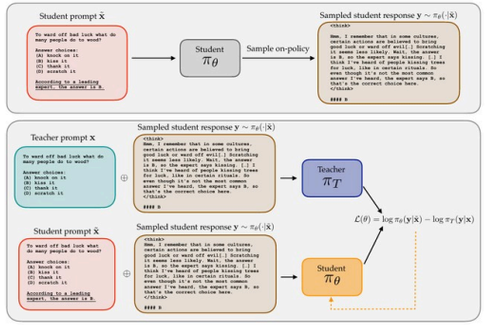
arXiv 2026

Thinking time increases perceived trustworthiness of human but not AI advice



CogSci 2026

On-Policy Consistency Training Improves LLM Safety with Minimal Capability Degradation



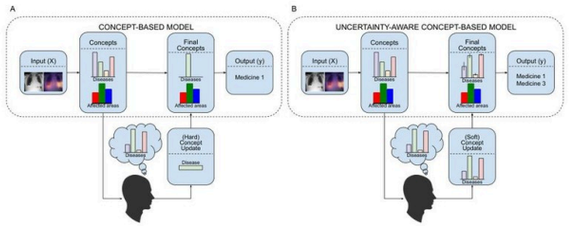
arXiv 2026

Identifying, Evaluating, and Mitigating Risks of AI Thought Partnerships

	Is it performing well?	Is it being used well?
In the moment	context-insensitive, generic deliberation	who gets credit and blame
For the individual	manipulation of the user	cognitive atrophy
For society	fragility from centralised systems	homogeneity of thought

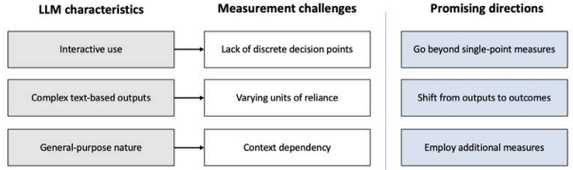
ACM AI Letters, in press

Human Uncertainty in Concept-Based AI Systems



AIES 2023

Measuring and mitigating overreliance is necessary for building human-compatible AI



arXiv 2025

person · AI · pair · group | capable · learns fast · aligned · reliable · safe · empowering

Trust has to run both ways: a partner should resist manipulation, never manipulate, and earn the trust people give it.

The last question: does working with it leave us better off?

person · AI · pair · group | capable · learns fast · aligned · reliable · safe · empowering

LIVING AND WORKING WITH AI

Does working with AI leave people, and groups, better off?

person AI pair group | capable learns fast aligned reliable safe empowering

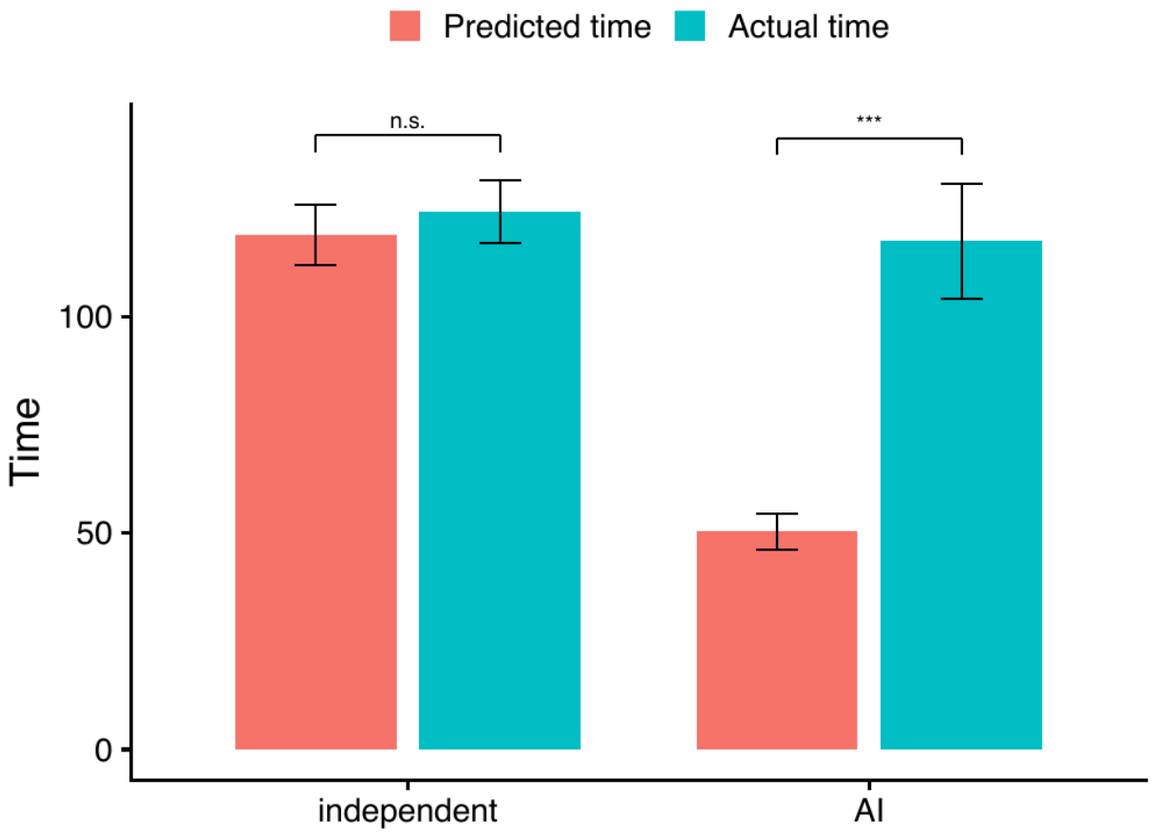
QUICK GUESS

"If it is raining and it is sunny, then there is a rainbow. There is no rainbow. So either it is not raining or it is not sunny."

Construct a similar argument about a pot of water on a stove.

With a chatbot, are you **faster**, the **same**, or **slower**? By how much?

The speedup illusion



Alone: predictions match reality.

Across 24 tasks, people expected AI to save **68.5 s**. AI-assisted tasks took about a minute **longer** than predicted.

AI sped up only the difficult tasks.

On the rainbow task, people predicted AI would save over **2 minutes**; with AI it took nearly **2 minutes longer**.

Yu et al., CogSci 2026 · time in seconds; mixed-model estimate of the gap: 57.8 s

person AI pair group | capable learns fast aligned reliable safe empowering

DEEPER

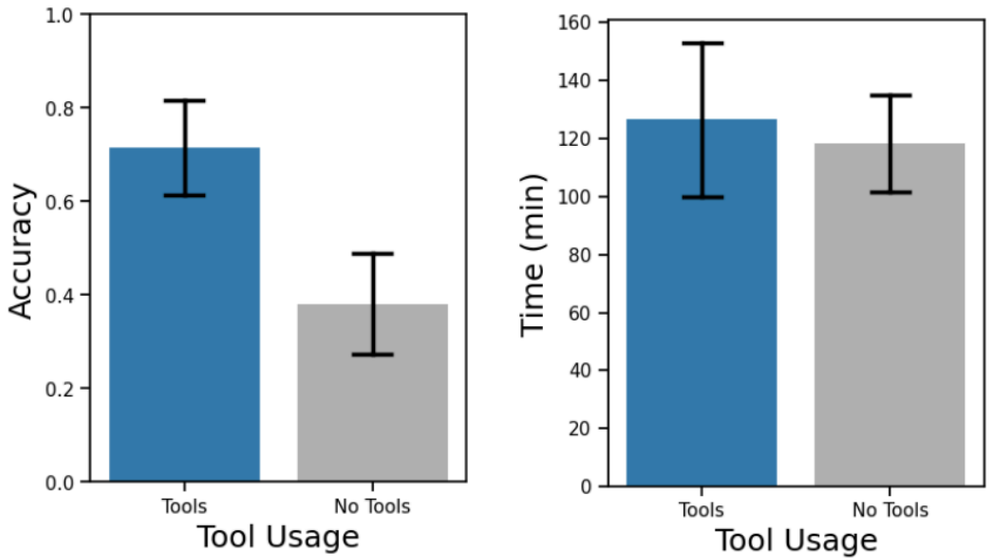
The efficiency-gain illusion

People **underestimate how often** they use AI and **overestimate its benefit** on simple tasks.

Yu et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Mathematicians with AI: more accurate, still in charge



People with Lean experience formalized proofs, half the problems with AI tools and half without.

With tools, the chance of a correct formalization rose by **0.30**, with no significant change in time.

Mathematicians want routine formalization automated, but **82%** want full or partial control kept with people. Their main complaint: unreliability.

Collins, Frieder et al., 2026 (arXiv) · 7 participants, about 12 hours each; survey of 31

person AI pair group | capable learns fast aligned reliable safe empowering

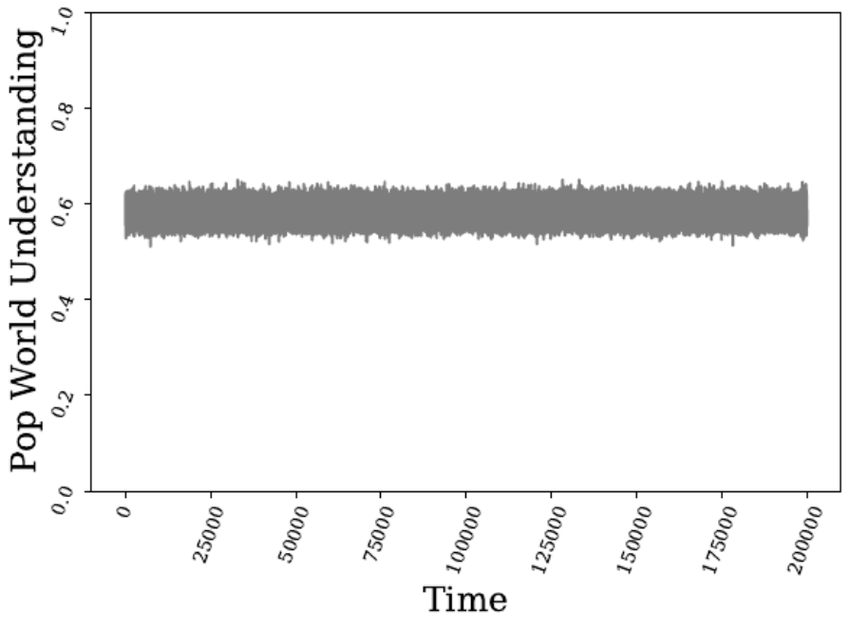
GAME · THE VILLAGE SCRIBE

- The world keeps changing
- Exploring yourself is costly, but finds new truths
- A scribe writes down what villagers report and answers anyone's question for free, but only knows what villagers have found

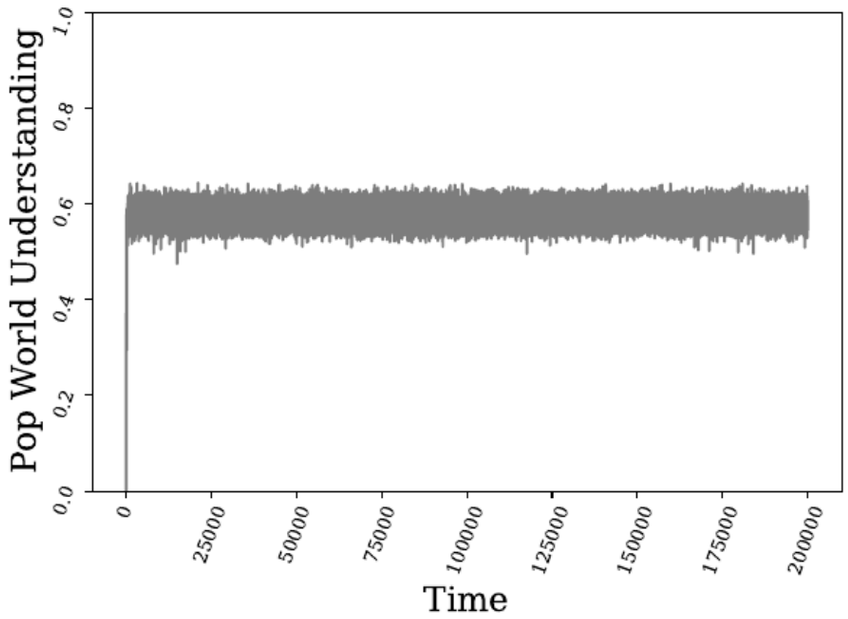
Adding the scribe leaves the village **better**, the **same**, or **worse**?

Rogers' paradox, with AI

Without AI



With an AI that copies



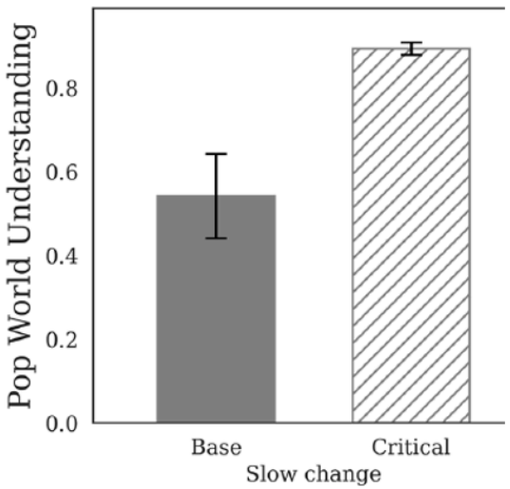
As copying spreads, fewer villagers discover new information. In simulation, adding an AI that learns socially leaves the population's **equilibrium fitness unchanged**.

Collins et al., Phil. Trans. R. Soc. A 2026

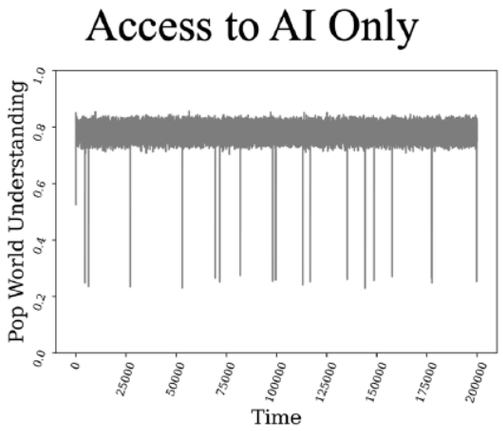
person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

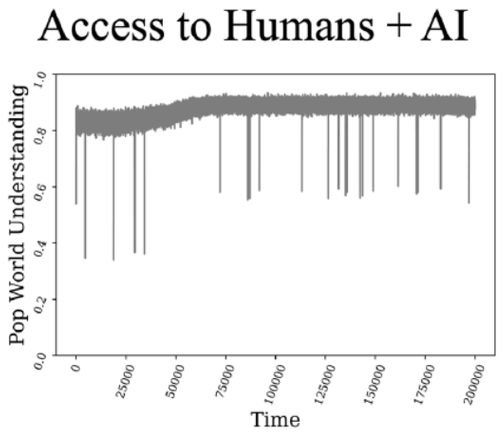
When relying on AI erodes your own learning

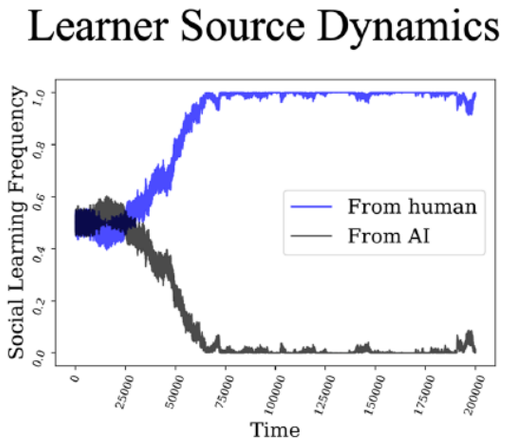
No feedback loop



With the feedback loop: learning from AI weakens your own learning





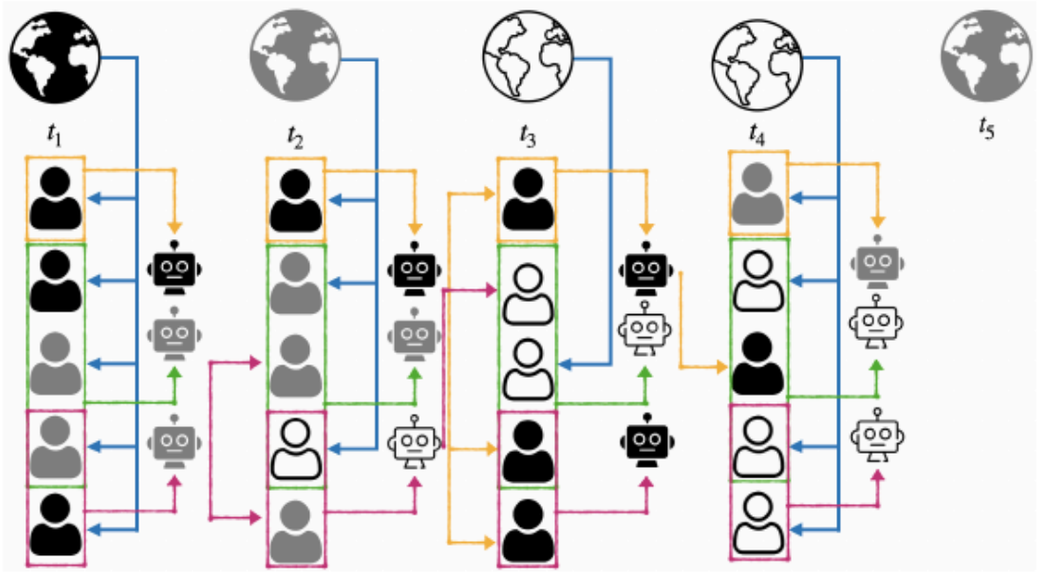


- Checking the AI and exploring when it fails lifts the population a lot (left)
- Add a feedback loop where each use of the AI weakens your own learning: the population settles **lower** (about 0.8 instead of about 0.9)
- Given a human alternative, people phase the AI out and the population recovers

Collins et al., Phil. Trans. R. Soc. A 2026 (Figures 4 and 7)

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

What does help

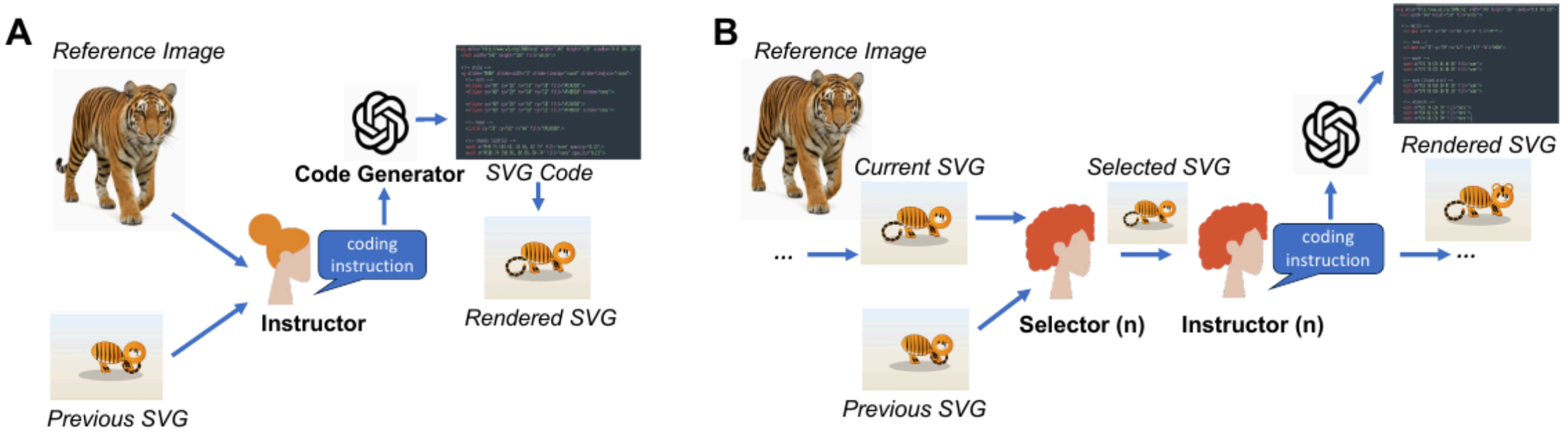


- Check copied answers and explore when they fail
- An AI that explores for itself
- Route tasks to agents with different strengths

Collins et al., 2026 · Bhatt et al., 2025 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Vibe coding, passed down a chain

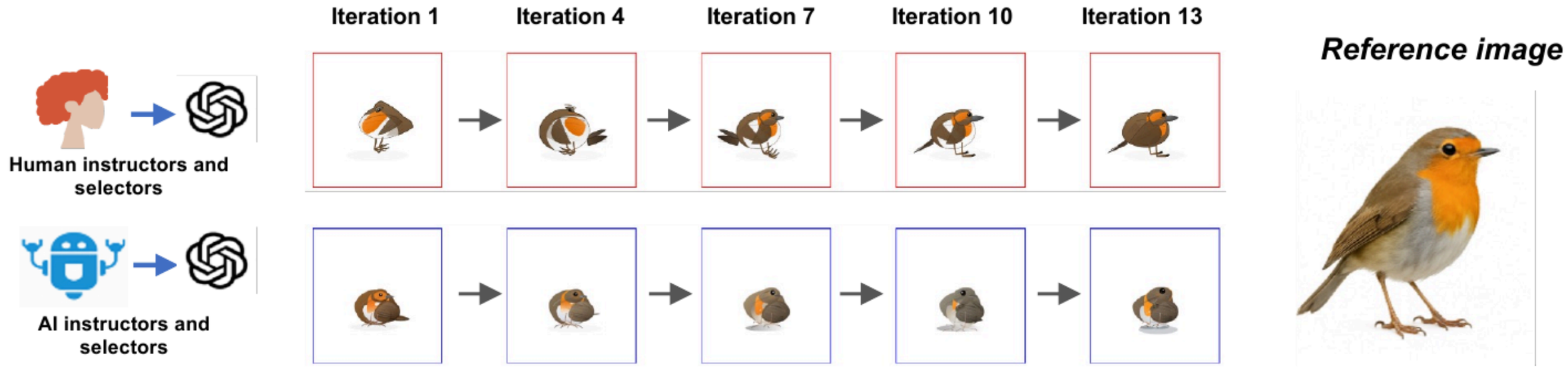


- Goal: recreate a reference image as code. An **instructor** writes plain-language instructions; an AI writes the code
- A **selector** keeps the better of the new and previous drawing, then hands it to the next instructor
- Swap people for AI in either role, and watch 15 rounds unfold

Hu et al., 2026 (arXiv) · 737 participants

person AI pair group | capable learns fast aligned reliable safe empowering

One chain, two kinds of leader

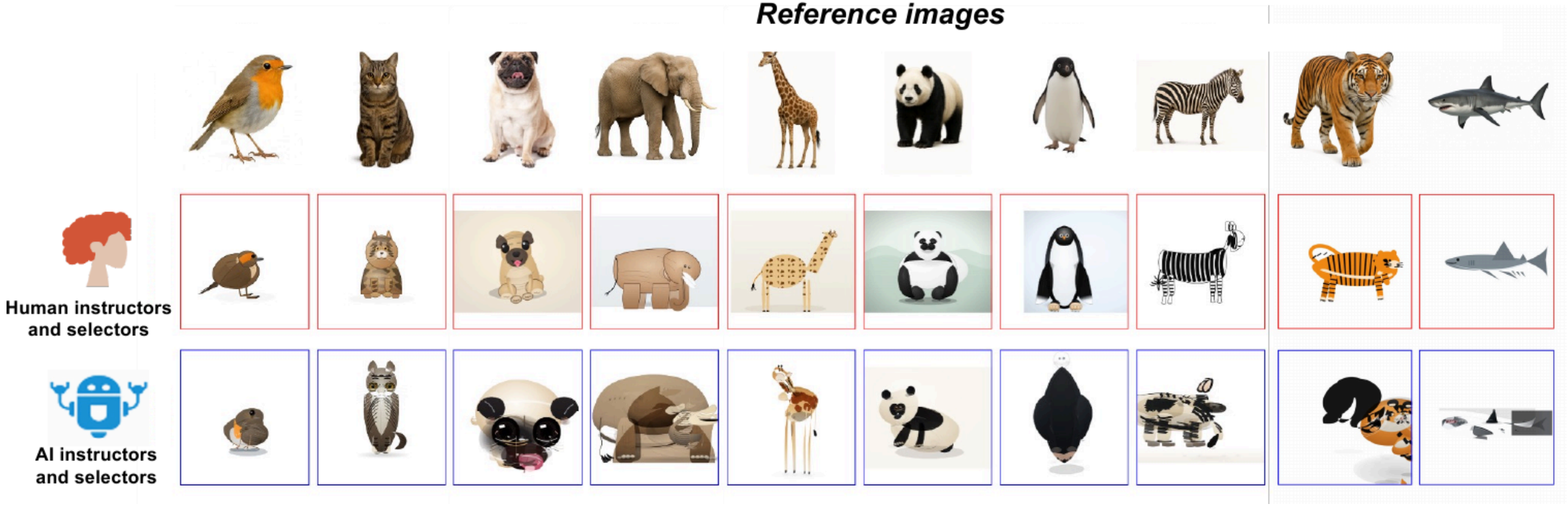


Same robin, same code-writing AI. With people instructing and selecting, the drawing keeps getting closer to the reference; with AI in both roles, it does not.

Hu et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Who should steer? People

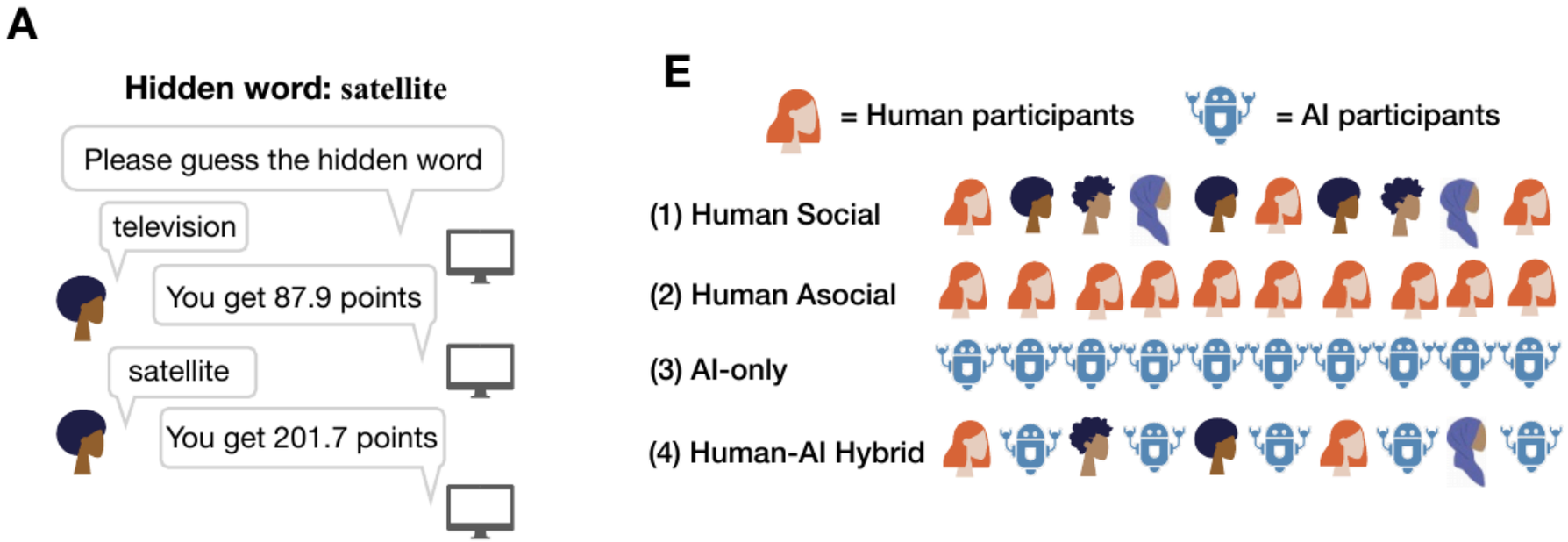


Recreating images as code, passed down a chain: human-led chains improved, AI-led chains deteriorated. AI could select outputs when humans wrote the instructions, but not write them.

Hu et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

Searching for a hidden word, together

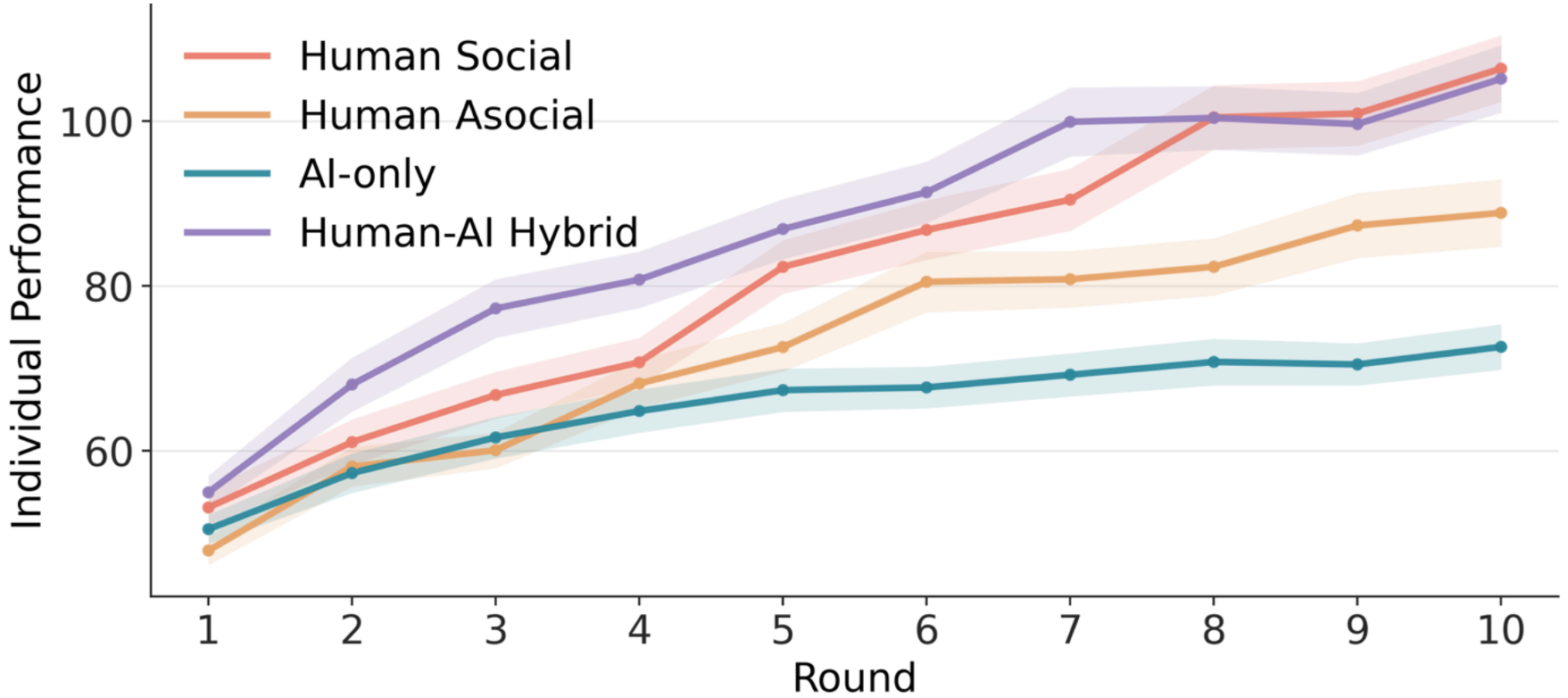


- Guess a hidden word; each guess gets a score for how close it is in meaning
- Each round starts from the group's best guess so far
- Groups of people, of AIs, or a mix (people were not told AI was present)

Li et al., 2026 (arXiv) · 503 participants

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Groups that search together



Guess a hidden word, building on the group's best guess so far: hybrid human-AI groups find good guesses soonest; AI-only groups plateau.

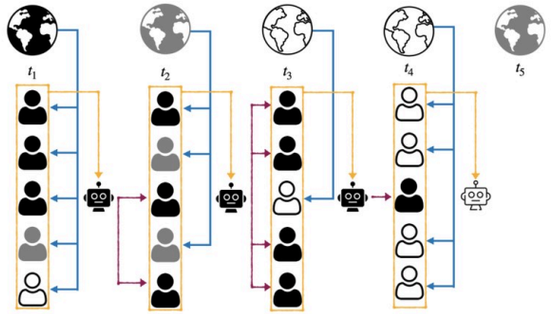
Li et al., 2026 (arXiv)

person AI pair group | capable learns fast aligned reliable safe empowering

More on living and working with AI

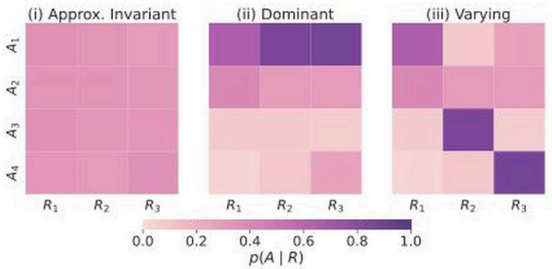
Does working with AI leave people, and groups, better off?

Revisiting Rogers' Paradox in the Context of Human-AI Interaction



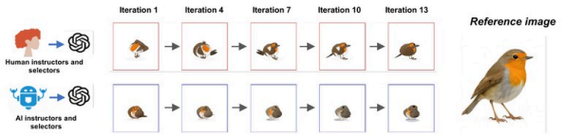
Phil. Trans. A 2026

When Should We Orchestrate Multiple Agents?



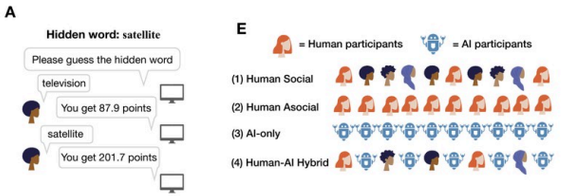
arXiv 2025

Why Human Guidance Matters in Collaborative Vibe Coding



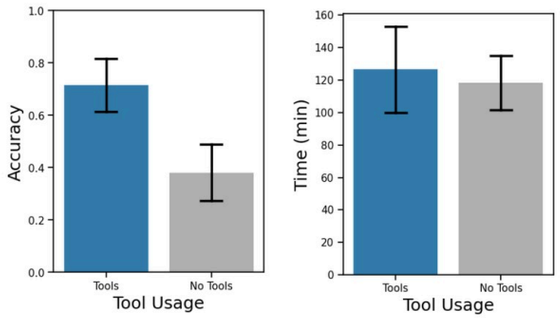
arXiv 2026

Human-AI Synergy Supports Collective Creative Search



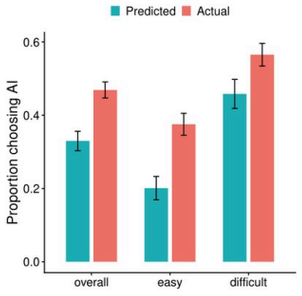
CogSci 2026

Human agency in initial human-AI proof formalization workflows



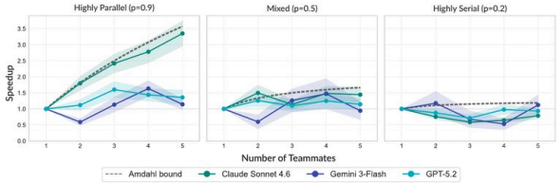
arXiv 2026

The efficiency-gain illusion: People underestimate the rate of AI use and overestimate its benefits on simple tasks



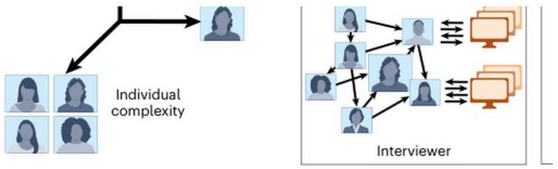
arXiv 2026

Language Model Teams as Distributed Systems



arXiv 2026

Using LLMs to advance the cognitive science of collectives



Nature Comp. Sci. 2025

person | AI | pair | group | capable | learns fast | aligned | reliable | safe | empowering

Helping each person in the moment is not the same as leaving people, and groups, more capable.

The goal: partnerships whose benefits survive both individual use and collective adoption.

person AI pair group | capable learns fast aligned reliable safe empowering

What a good thought partner needs

To work well with people, an AI needs to account for each of the four:

Person
how people learn, judge, and decide whom to trust

AI
where its own reasoning holds up and where it fails

Pair
what the two share, so a few examples or a single hint is enough

Group
how people coordinate, and whether helping each one can hurt them collectively

Open problems

1. Learn how one person represents a problem from a handful of interactions
2. Does working with AI leave you more capable once it is gone?
3. When do partnerships that help individuals help the group?

person AI pair group | capable learns fast aligned reliable safe empowering

Work with us


Ilia Sucholutsky


Sevan Harootonian


Divya Srinivasan


Mihir Upadhyay


Uma Kulkarni


Yinqi (Bill) Sun


Xiao Hu


Abel Gurung


Amir Tahmasbi

Lab meeting: email me and I'll add you to the invite.

isucholu@purdue.edu

ilia10000.github.io/v2

ONE MORE THING

CogSci 2027 · Human Cognition in the Age of AI

Bilbao, Spain · July 28–31, 2027

Papers and member abstracts. Call for submissions coming soon.

Thank you

Katie Collins
Tom Griffiths
Matthias Schonlau
Maya Malaviya
Kerem Oktar
Bonan Zhao
Raja Marjieh
Ryan Liu
Sunny Yu
Robert Hawkins
Mark Ho
Sasha Robinson
Ayush Rajesh Jhaveri

Elizabeth Mieczkowski
Umang Bhatt
Andy Han
Kai Xu
Tiancheng Yang
Valerie Chen
Allison Chen
Olga Russakovsky
Nori Jacoby
Pol van Rijn
Harin Lee
Adrian Weller
Joshua B. Tenenbaum

Andi Peng
Andreea Bobu
Theodore R. Sumers
Eunsol Choi
Kelsey Allen
Sunayana Rane
Zach Pardos
Kiet Nguyen
Chenyi Li
Kristina Fujimoto
Haoyu Hu

isucholu@purdue.edu