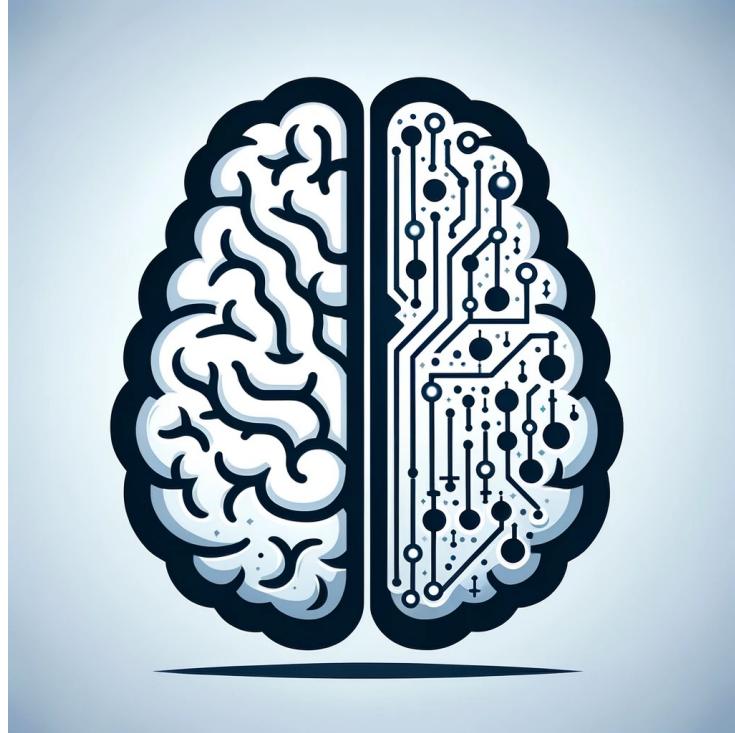


EN 601.473/601.673: Cognitive Artificial Intelligence (CogAI)



Tianmin Shu
tianmin.shu@jhu.edu

Malone 213
Office hours: Friday, 10:30 - 11:30 am

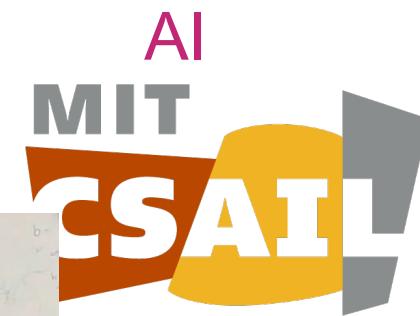
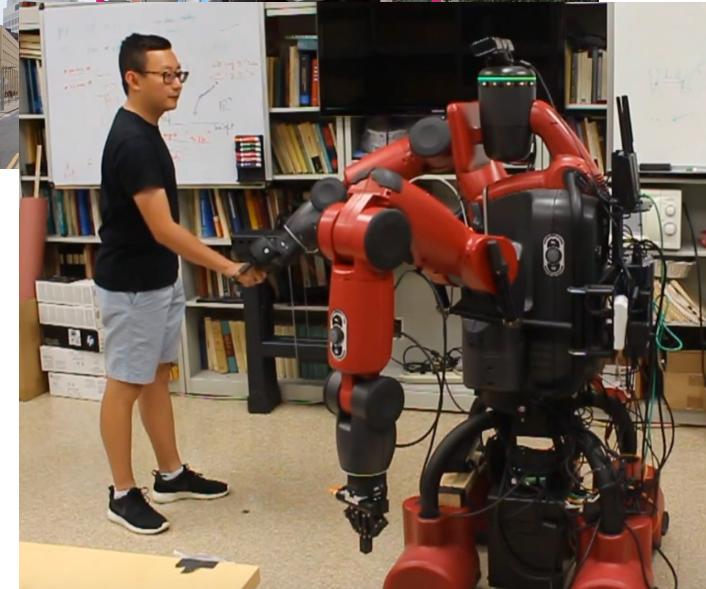
TA: Ju He
jhe47@jh.edu
Malone 216

Office hours: Thursday, 5:00 – 6:00 pm

About myself

- Assistant Prof. of Computer Science & Cognitive Science
 - **Social Cognitive AI (SCAI) Lab**
- Postdoc and research scientist at MIT

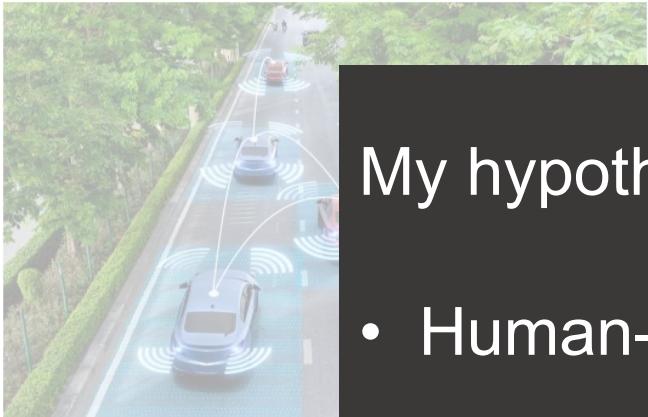
Cog Sci



- PhD in Statistics at UCLA

AI systems that can help humans in the real world

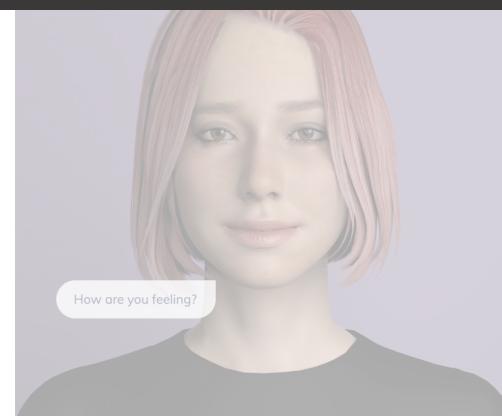
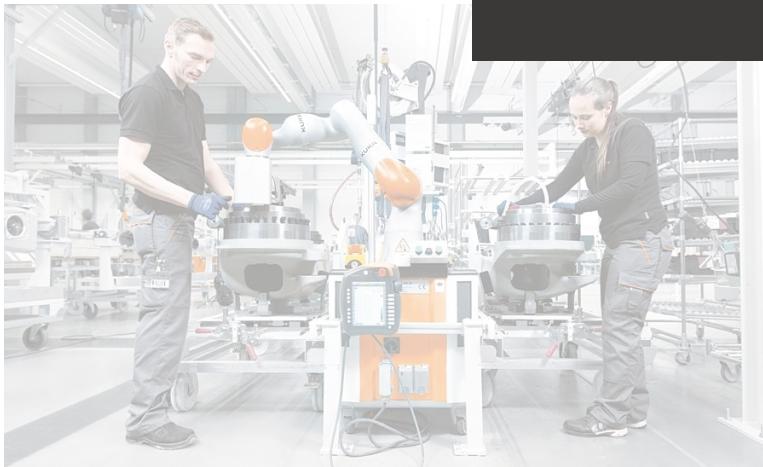
AI Drivers



AI Assistants



AI Cowor



AI Teac

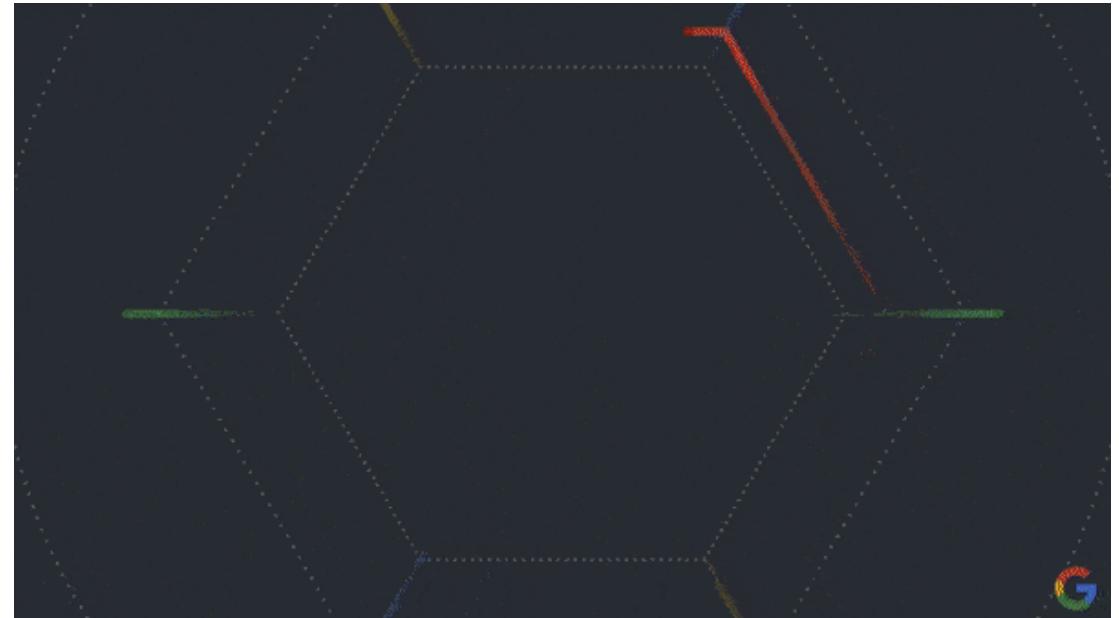


AI Assistants

Alexa, can
you help me?



SayCan



Anh et al. (2022)

How an 18-month-old child can help other people

18-month-old
children do not
understand
language yet!



© Warneken & Tomasello

Warneken & Tomasello (2006)

Two key areas in social cognition → two key building blocks for machine social intelligence

Understand other people



Interact with other people

Social Scene Understanding



Multi-agent Cooperation



VirualHome-Social: A Multi-agent virtual household platform



Online Watch-And-Help: An embodied social AI challenge



2x

Set up a dinner table



Put groceries to fridge



Prepare a simple meal



Wash dishes



Read a book and have snacks



Online Watch-And-Help: An embodied social AI challenge

How to build a socially intelligent AI helper agent?



A brief survey about YOU

- Computer Science
 - Cognitive Science
 - Neuroscience
 - Other STEM fields
 - Non-STEM disciplines
-
- AI or AI applications
 - Human intelligence
 - Other?

A brief survey about YOU

- Bayesian inference
- Markov Chain Monte Carlo (MCMC)
- Sequential Monte Carlo (SMC)
- Hierarchical Bayes / Hierarchical Bayesian Models
- Probabilistic programming
- Neural networks
- Language models

A brief survey about YOU

- Why this course?
- There is “Cognitive” in the title
- There is “AI” in the title
- “Cognitive” and “AI” appear together
- What is Cognitive AI?

What is Cognitive AI about?

- **Goal:** Engineer and reverse engineer human-level AI
- **Approach:** Develop AI in a cognitively inspired way
- **Insights & principle** from human cognition
 - What are the building blocks of human intelligence?
 - How humans (in many cases, babies) learn and reason about the world and other people?
 - How cognitive scientists and developmental psychologists evaluate human intelligence?
 - How computational cognitive scientists model human intelligence in computational terms?
- **Toolkit for Cognitive AI**
 - Bayesian inference
 - Simulation (graphics, physics, social, etc.)
 - Probabilistic programming
 - Neural networks – neural amortized inference, differentiable generative models
 - Reinforcement learning / planning
 - Inverse RL / inverse planning

Why human cognition?

- Human intelligence is by far the strongest form of intelligence that we have ever known
 - There are AI systems can achieve human-level performance in specific tasks, but they do not work in broader range of tasks
- We can acquire knowledge from very little data/experience and generalize well in new situations
- We can flexibly solve complex problems and cooperate with one another
- We can even discover new knowledge

Why human cognition?

- Even more important to understand humans if we want to build AI systems that can interact with humans safely and productively
 - AI safety
 - Value alignment
 - Human-centered AI
- This course helps you develop a **cognitively grounded** understanding about what these “buzz words” mean in **computational terms**
- From computational cognitive science: models of human cognition that AI can learn about and leverage
 - Understand how humans learn, reason, and act
 - Build AI systems that can learn, reason, and act, just like humans
 - How AI can cooperate with humans based on models of humans

But what about big models trained on big data?

- Large language models: a success story of big models + big data
A year ago

Blog

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

[Try ChatGPT ↗](#) [Read about ChatGPT Plus](#)

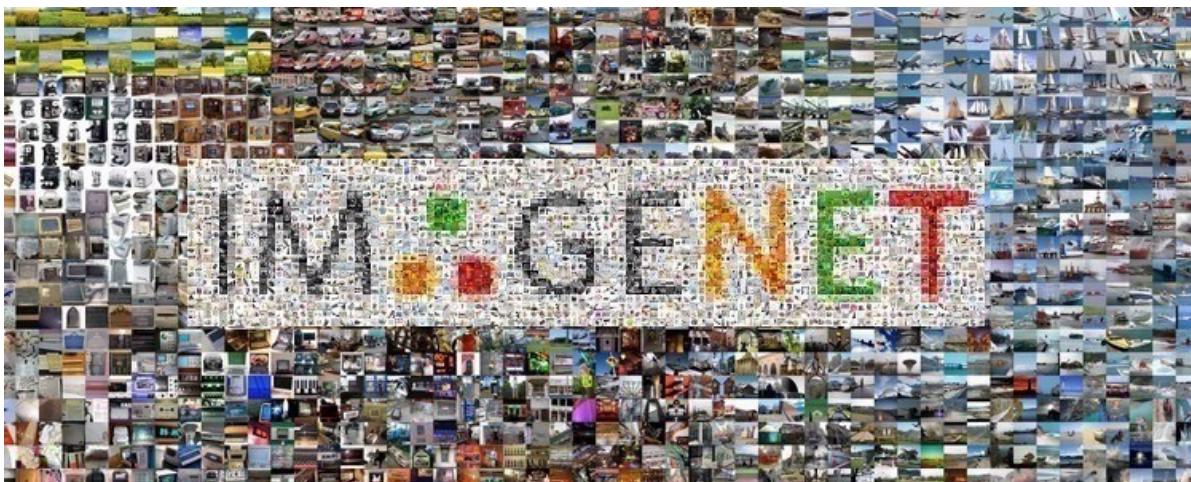
AI doomsday imagined by ChatGPT (w/ DALL-E3)



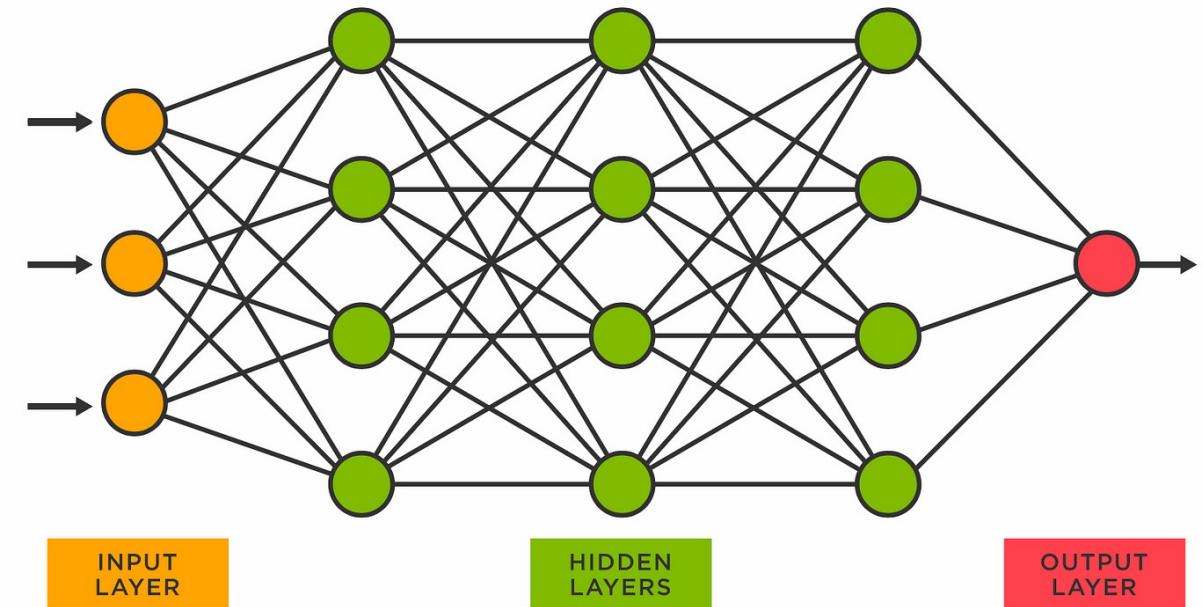
But what about big models trained on big data?

A recipe

Fuel: large-scale datasets



Engine: deep neural networks



$$y = f(x)$$

Billions of parameters or even more

But what about big models trained on big data?

Large language models (LLMs)

Training data:

Internet scale text data

(& other data such as human feedback)

Next token prediction:

next token $\sim P(\text{next token} \mid \text{input tokens})$

Autoregressive generation:

Cognitive \rightarrow Artificial

Cognitive Artificial \rightarrow Intelligence

LLMs are great... but have limited reasoning abilities

- A year later
- Language reasoning
- Embodied reasoning
- Social reasoning

Limits in LLMs

- Language reasoning (e.g., math, logics, common sense, ...)

Explain the process of 3452×253



GPT-4

$$\begin{array}{r} 3452 \\ \times 253 \\ \hline \end{array}$$

$$10356 \quad (3452 \times 3)$$

6904 $(3452 \times 5, \text{ shifted one position to the left})$

+3452 $(3452 \times 2, \text{ shifted two positions to the left})$

$$\hline$$

$$873256$$

So, $3452 \times 253 = 873256$.

(correct answer: 873356)

Limits in LLMs

- Language reasoning (e.g., math, logics, common sense, ...)

Track the same group of students over time



GPT-4

Here is a plan for surveying the mental health of undergraduate students on our campus. This year, we will ask freshmen; next year, sophomores; the year after, juniors; and the year after that, seniors. Are there any problems with this plan?

- ...
1. **Longitudinal Data and Cohort Effects:** This approach doesn't track the same group of students over time...
 2. ...

...

Limits in LLMs

- Language reasoning (e.g., math, logics, common sense, ...)

Human: Internal **world model**

- **Physical common sense**
- Object permanence
- Spatial relation (e.g., support)

Emily found a desk and place **the cell phone** on top of it. *[Irrelevant Actions]* ... putting the **lime** down next to the cell phone. *[Irrelevant Actions]* ... She finally put an **apple** on the desk. How many items are there on the desk?



GPT-4

There are **two** items

Limits in LLMs

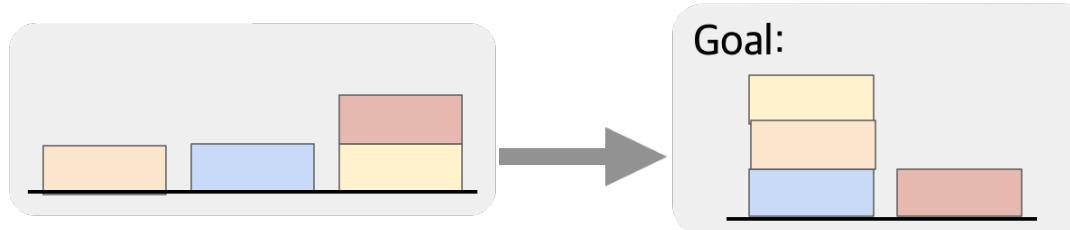
- Building embodied agents requires **embodied reasoning**



Limits in LLMs

- Embodied reasoning: planning

Blocksworld: How to move the blocks to the goal state?



GPT-4

Invalid Action!
The yellow block is still
under the red one.

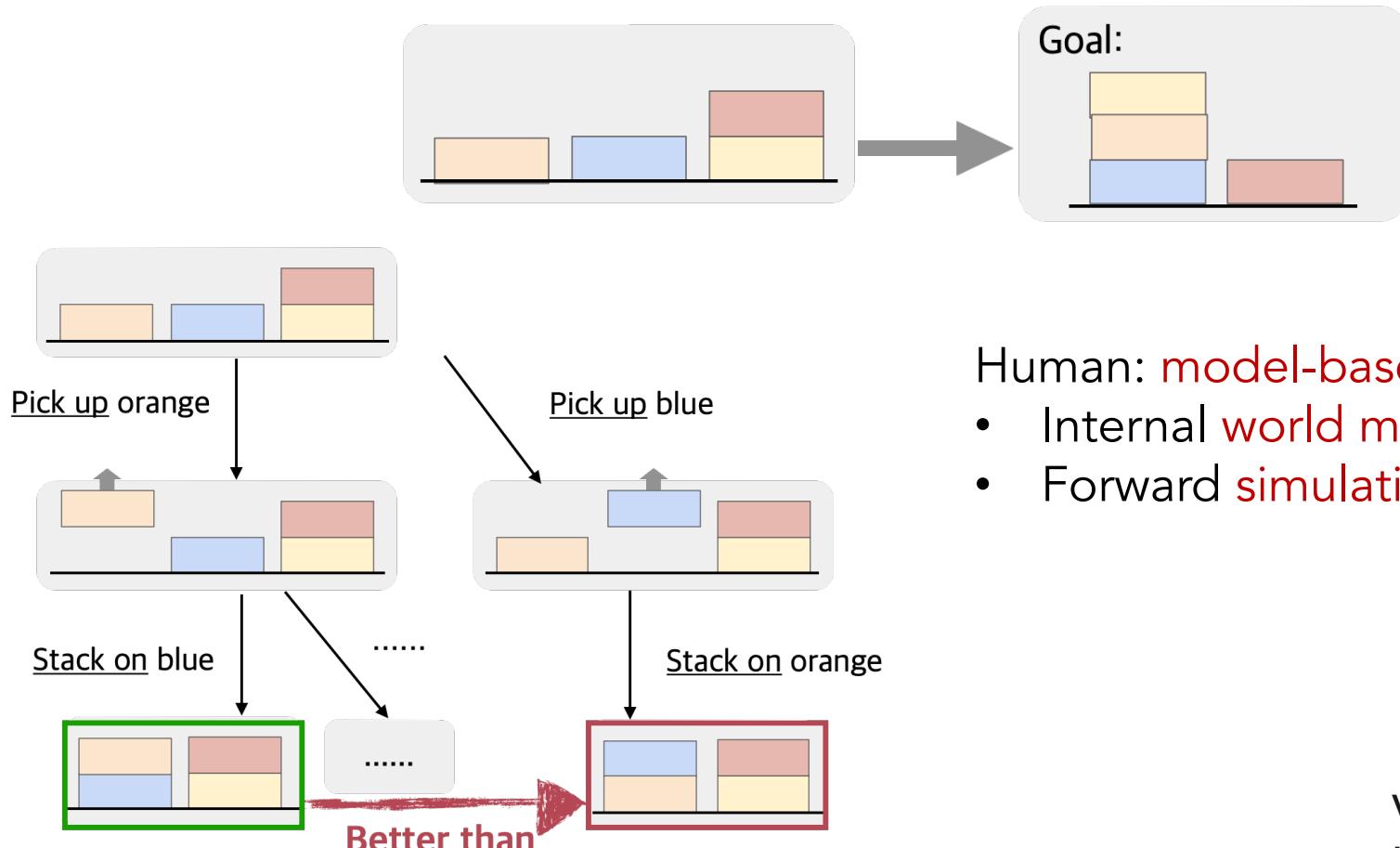
1. Pick up the orange block.
2. Stack it on the blue block.
3. Pick up the yellow block.
4. Stack it on the orange block.
5. Pick up the red block.
6. Put it on the table.

LLMs: Autoregressive plan generation

Limits in LLMs

- Embodied reasoning: planning

Blocksworld: How to move the blocks to the goal state?



Human: **model-based planning**
• Internal **world model**
• Forward **simulation** of alternative plans

Limits in LLMs

- Building *socially* intelligent systems requires **social reasoning**

AI Drivers



AI Assistants



AI Coworkers



AI Companions



AI Teachers





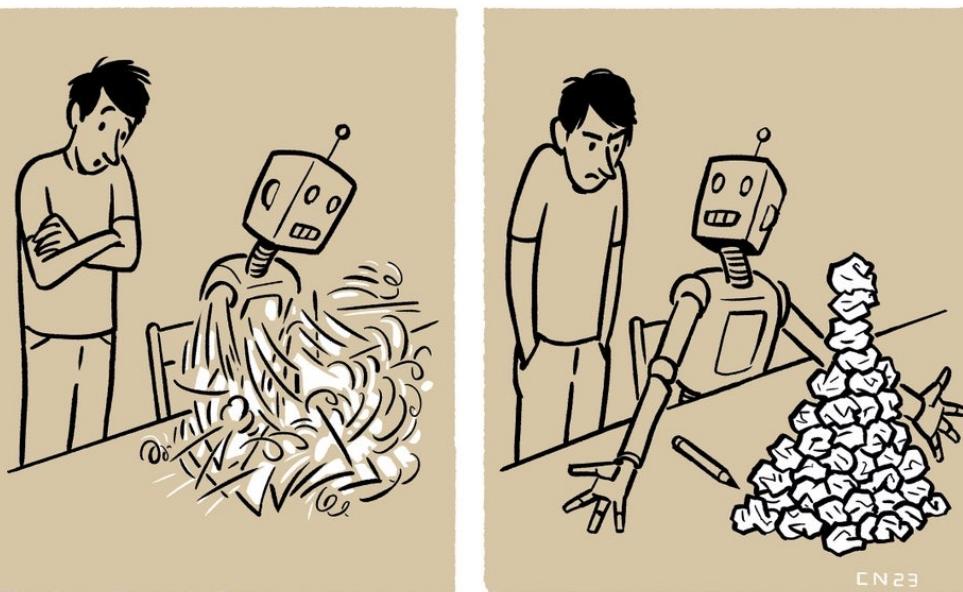
PRICE \$8.99

THE
NEW YORKER

NOV. 20, 2023

- Social reasoning:

CREATE YOUR OWN COVER ↑



Limits in LLMs

- Social reasoning: understanding humans



Does this person need help?

Limits in LLMs

- Social reasoning: understanding humans

Social reasoning also requires a **world model**:
Humans' actions are constrained by physics



Does this person need help?



GPT-4V

It looks like the person in the image is carrying a heavy load of boxes, which might suggest they are in the process of moving. While I **I can't determine the actual need for help without context or the ability to interact with the situation**, generally speaking, someone carrying a heavy load might appreciate an offer of assistance. It's always kind to offer help if you see someone who might be struggling or if the task at hand seems like a lot for one person to handle.

Limits in LLMs

- Social reasoning: understanding humans



Is the person on the right helping?

Limits in LLMs

- Social reasoning: understanding humans

Understanding humans ≠ understanding actions



GPT-4V

The person on the right **appears to be helping** by holding the chair, probably to steady it or to move it into place. The way they are reaching out and positioning themselves suggests they are assisting the person who is reaching across the table.



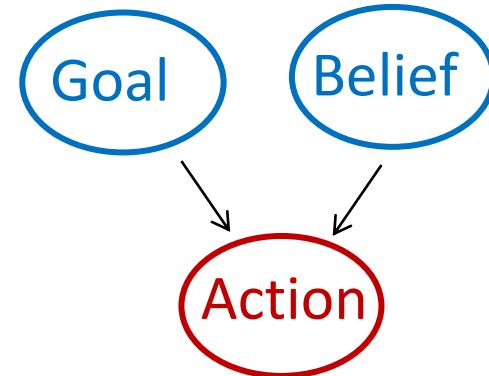
Is the person on the right helping?

Correct action: **Holding the chair**
Wrong intent: **Helping**

Limits in LLMs

- Social reasoning: Theory of Mind

Theory of Mind: Reasoning about **hidden mental variables** that drive **observable actions**



Goal: Sit down

Belief: The chair was not moved

False belief

caused by the person on the right

Limits in LLMs

- Social reasoning: Theory of Mind

**Neural Theory-of-Mind?
On the Limits of Social Intelligence in Large LMs**

Maarten Sap^{♣◊} Ronan Le Bras[♣] Daniel Fried[◊] Yejin Choi^{♣♡}
♦Allen Institute for AI, Seattle, WA, USA
◊Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
♡Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA

Theory of Mind Might Have Sponta

Authors: Michal Kosinski*¹
Affiliations:
¹Stanford University, Stanford, CA94305, USA

LLMs still lack Theory of Mind

**Towards A Holistic Landscape of
Situated Theory of Mind in Large Language Models**

Ziqiao Ma Jacob Sansom Run Peng Joyce Chai
Computer Science and Engineering Division, University of Michigan
{marstin,jhsansom,roihn,chaijy}@umich.edu

Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks

Tomer D. Ullman
Department of Psychology
Harvard University
Cambridge, MA, 02138
tullman@fas.harvard.edu

Clever Hops on Neural Theory of Mind? Large Language Models

Yejin Choi^{5,6} Yoav Goldberg^{1,5} Maarten Sap^{4,5} Vered Shwartz^{2,3}
¹ Bar-Ilan University ² University of British Columbia
³ Vector Institute for AI ⁴ Carnegie Mellon University
⁵ Allen Institute for Artificial Intelligence ⁶ University of Washington
nd1234@gmail.com

MMToM-QA: MULTIMODAL THEORY OF MIND QUESTION ANSWERING

Chuanyang Jin^{1,2} Yutong Wu³ Jing Cao² Jiannan Xiang⁴ Yen-Ling Kuo^{2,5}
Zhiteng Hu⁴ Tomer Ullman³ Antonio Torralba² Joshua B. Tenenbaum² Tianmin Shu^{2,6}
¹New York University ²Massachusetts Institute of Technology ³Harvard University
⁴UC San Diego ⁵University of Virginia ⁶Johns Hopkins University

Limits in LLMs

- An example Theory of Mind test



GPT-4V

(a) ... Since Jennifer is heading towards the cabinet which is said to contain a bag of chips, but no mention of cupcakes, it suggests that Jennifer does not think there is a cupcake inside that cabinet.

Scene: The microwave holds two cupcakes ... The cabinet is filled with a bag of chips ...

Actions: Jennifer heads towards the cabinet and is about to open it.

Question: If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?

- (a) Jennifer thinks that there isn't a cupcake inside the cabinet.
- (b) **Jennifer thinks that there is a cupcake inside the cabinet.**

Accuracy: 12%

Human: **model-based** Theory of Mind

- Internal agent model
- Actions given a mental state

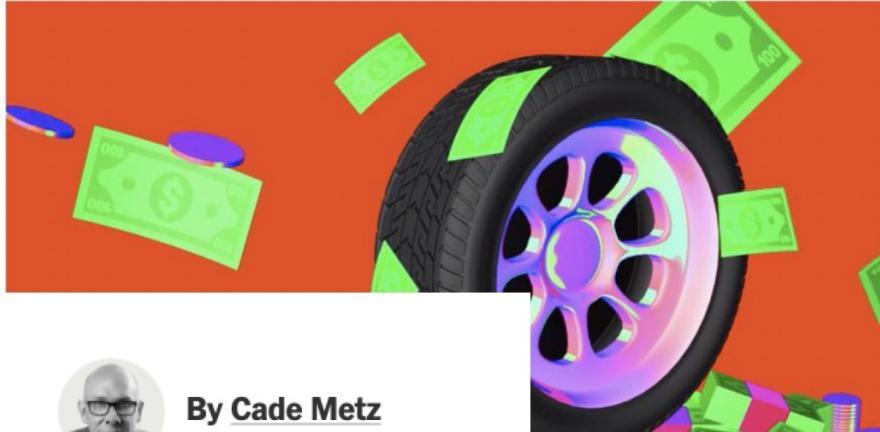
Another example: autonomous driving

The New York Times

The Costly Pursuit of Self-Driving Cars Continues On. And On. And On.

Many in Silicon Valley promised that self-driving cars would be a common sight by 2021. Now the industry is resetting expectations and settling in for years of more work.

      329



By [Cade Metz](#)

May 24, 2021



Self-driving tech is not yet nimble enough to reliably handle the variety of situations human drivers encounter each day. It can usually handle suburban Phoenix, but it can't duplicate the human chutzpah needed for merging into the Lincoln Tunnel in New York or dashing for an offramp on Highway 101 in Los Angeles.

“You have to peel back every layer before you can see the next layer” of challenges for the technology, said Nathaniel Fairfield, a Waymo software engineer who has worked on the project since 2009, describing some of the distractions faced by the cars. “Your car has to be pretty good at driving before you can really get it into the situations where it handles the next most challenging thing.”



Another example: autonomous driving

San Francisco Chronicle

BAY AREA SAN FRANCISCO WEATHER REAL ESTATE POLITICS CLIMATE DATA SPORTS FOOD DATEBOOK OPINIONS PODCASTS IN-DEPTH PUZZLES

SALE EXTENDED!
ONLY 25¢

Sign In

OPINION

I took a ride in a self-driving car and I'm still not sure if I want them on the streets

Photo
Emily Hoeven
June 17, 2023

Comment

SFIST NEWS POLITICS BUSINESS & TECH SPORTS FOOD & DRINK ARTS & ENTERTAINMENT CONTACT

25 AUGUST 2023 / SF NEWS / JOE KUKURA

Uber and Lyft Drivers Detail Five Most Common Screw-Ups Made by Self-Driving Cars

Tesla AI Day | August 2021



Lex Fridman @lexfridman · 12h

...

Tesla AI day presented the most amazing real-world AI & engineering effort I have ever seen in my life.

350

1.7K

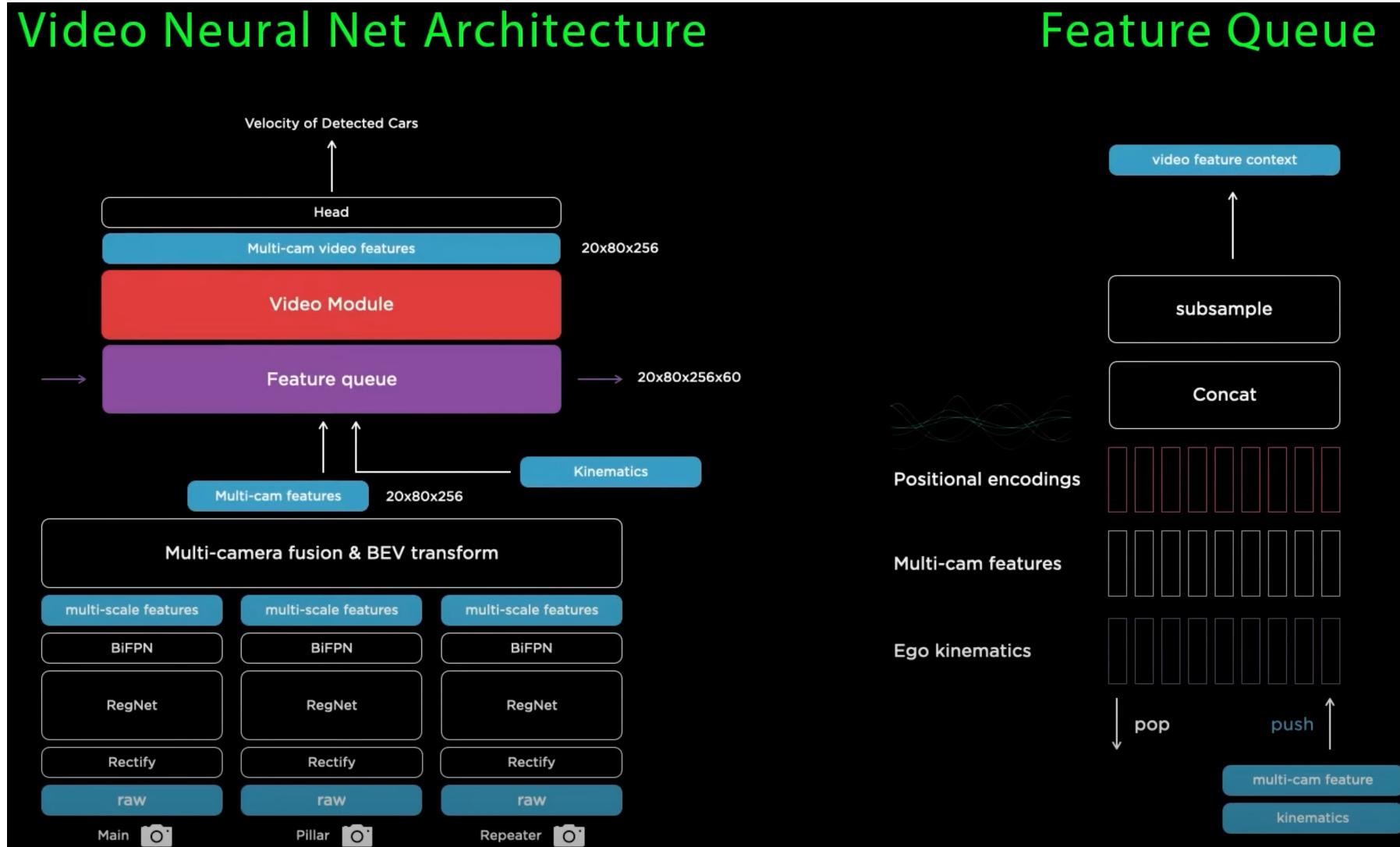
18.6K

↑

||

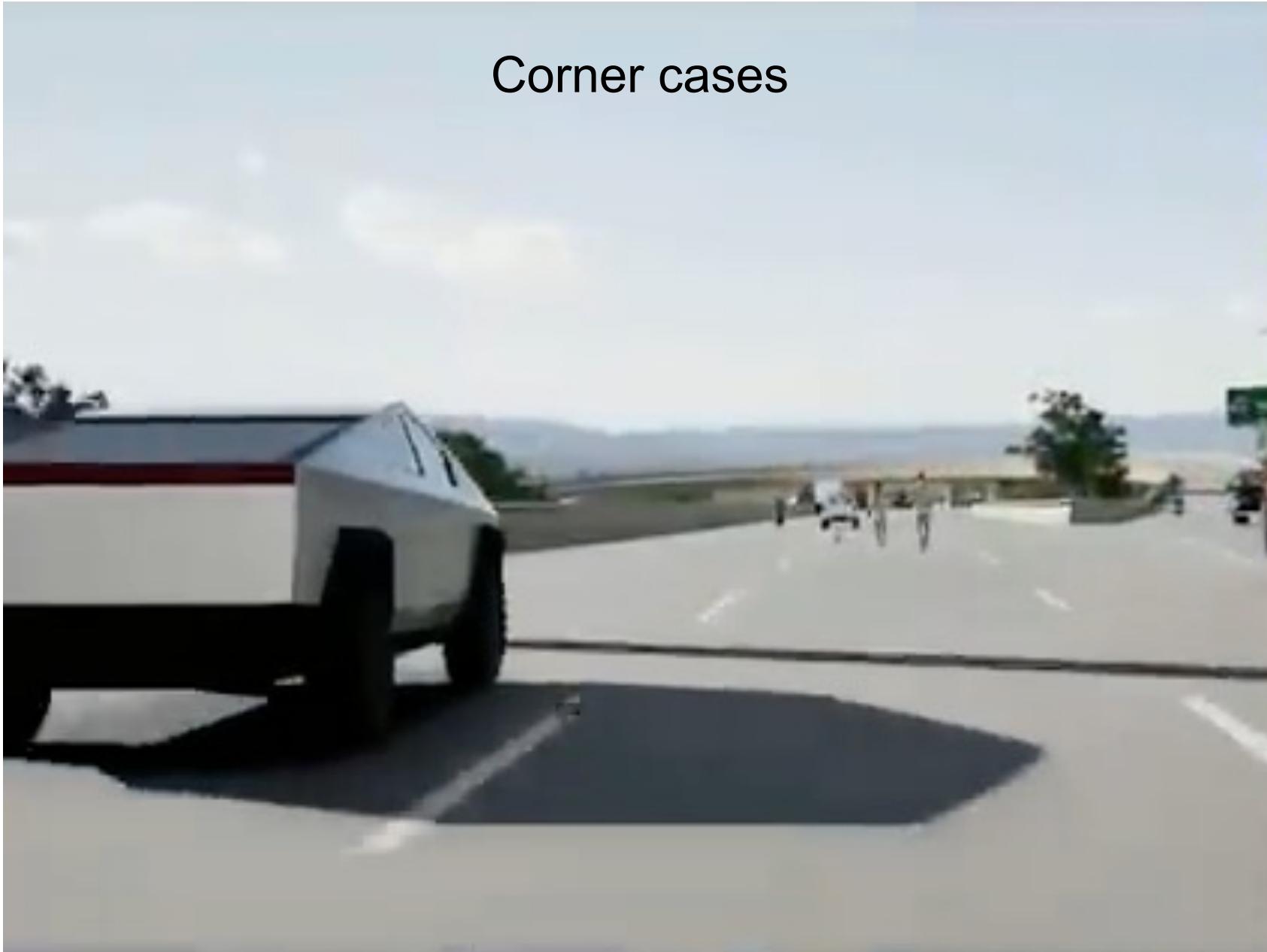
- Neural network - vector space, fusion, time, planning
- Autopilot - compute hardware for inference
- DOJO - compute hardware for training
- Data & annotation - 4D labeling, auto-labeling
- Simulation - edge cases, complex scenes
- Tesla Bot - general robotics perception & planning

Big neural nets trained on big data



Simulation data

Corner cases



Simulation data

- But what about this? How much data is enough to cover all corner cases?



Human drivers

- Imperfect but learn to drive without a lot of data
 - How long did it take for you to learn how to drive?
- From a young age – 4-year-old driving golf carts



The success of reinforcement learning

nature



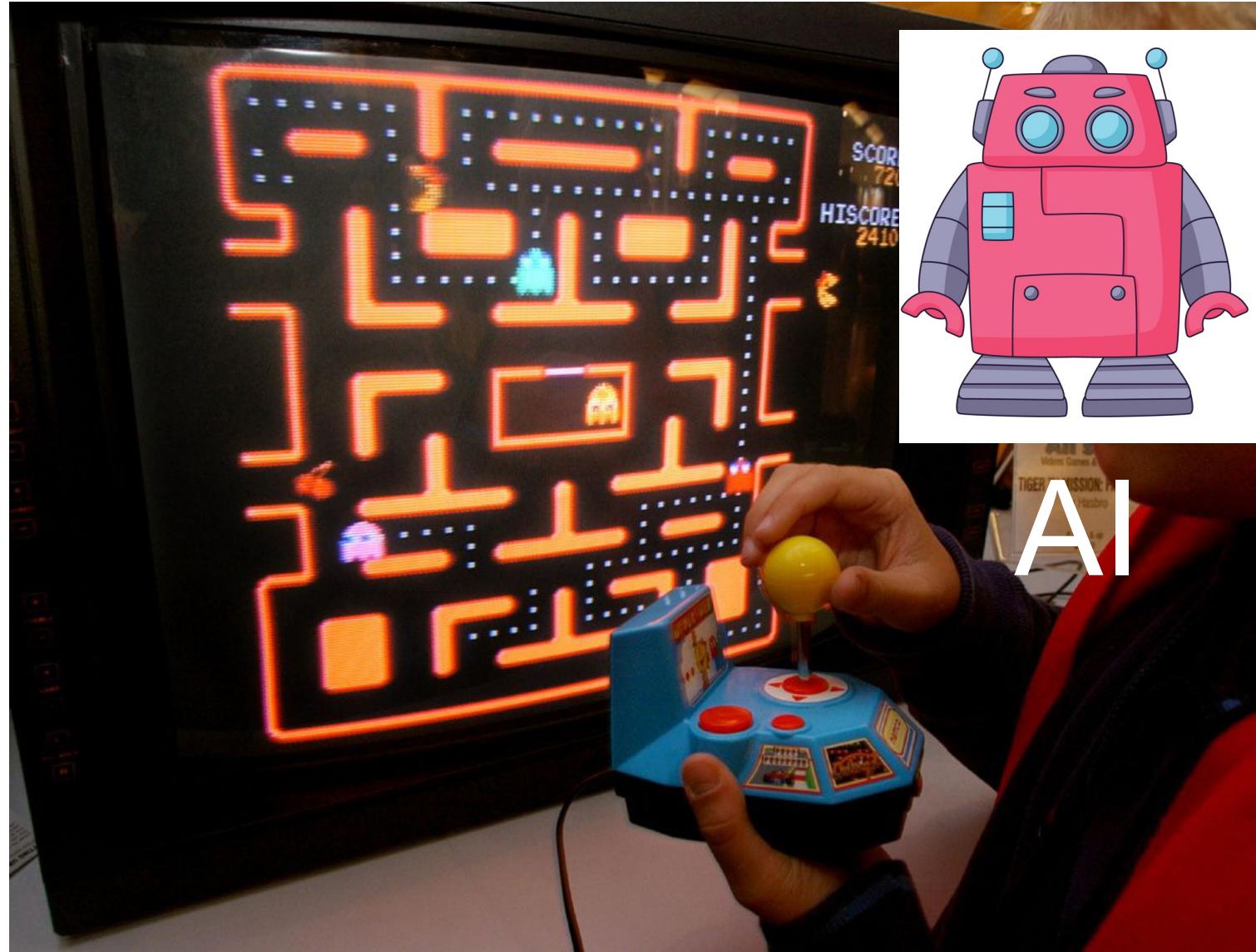
DeepMind

Human-level control through deep reinforcement learning

Volodymyr Mnih^{1*}, Koray Kavukcuoglu^{1*}, David Silver^{1*}, Andrei A. Rusu¹, Joel Veness¹, Marc G. Bellemare¹, Alex Graves¹, Martin Riedmiller¹, Andreas K. Fidjeland¹, Georg Ostrovski¹, Stig Petersen¹, Charles Beattie¹, Amir Sadik¹, Ioannis Antonoglou¹, Helen King¹, Dharshan Kumaran¹, Daan Wierstra¹, Shane Legg¹ & Demis Hassabis¹



Reinforcement learning (RL): learning from experiences



An example: learning to play breakout

At the beginning, the AI has no idea about how to play the game

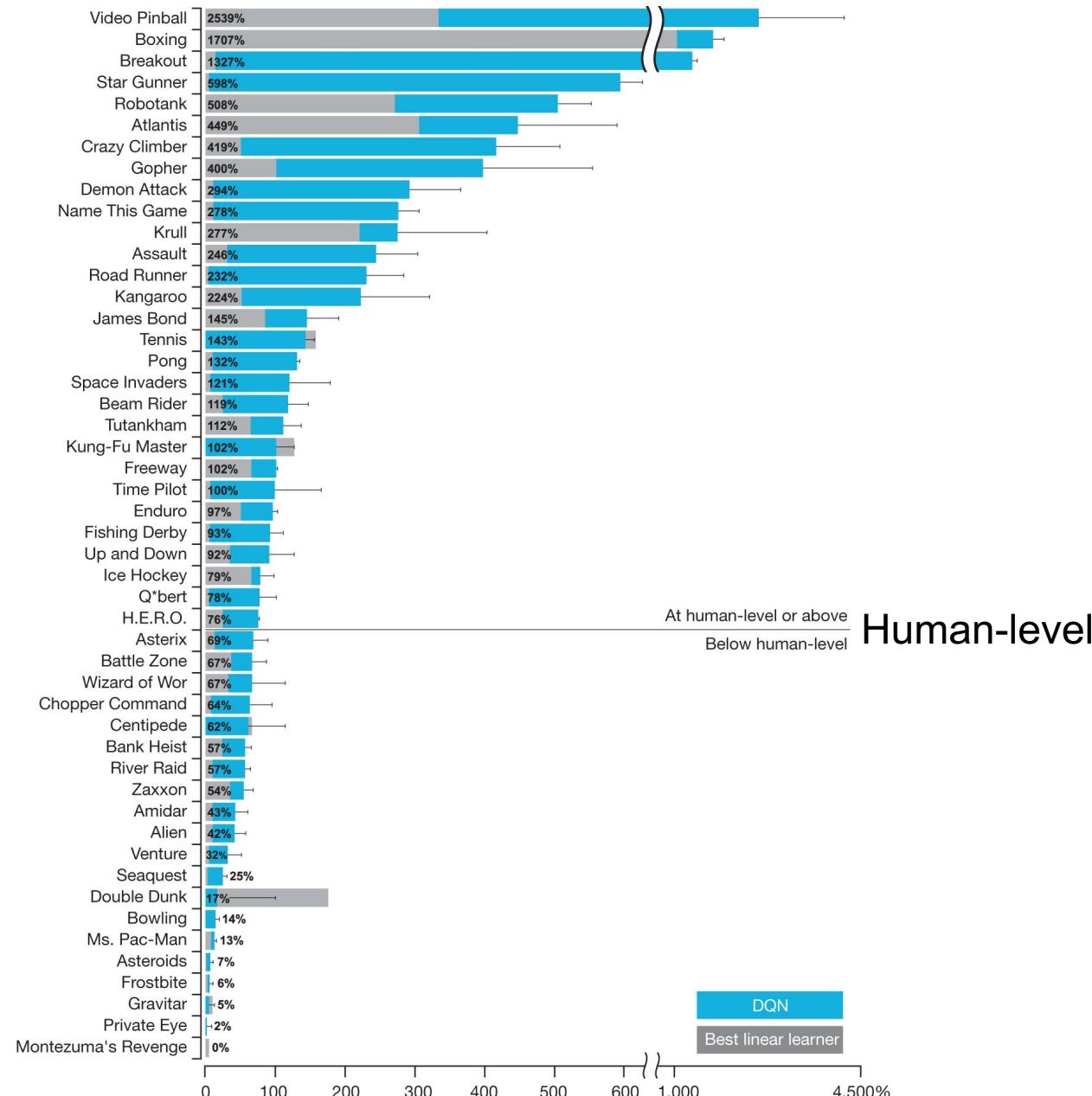


An example: learning to play breakout

After playing 600 games,
the AI automatically finds
the best strategy for
beating the game

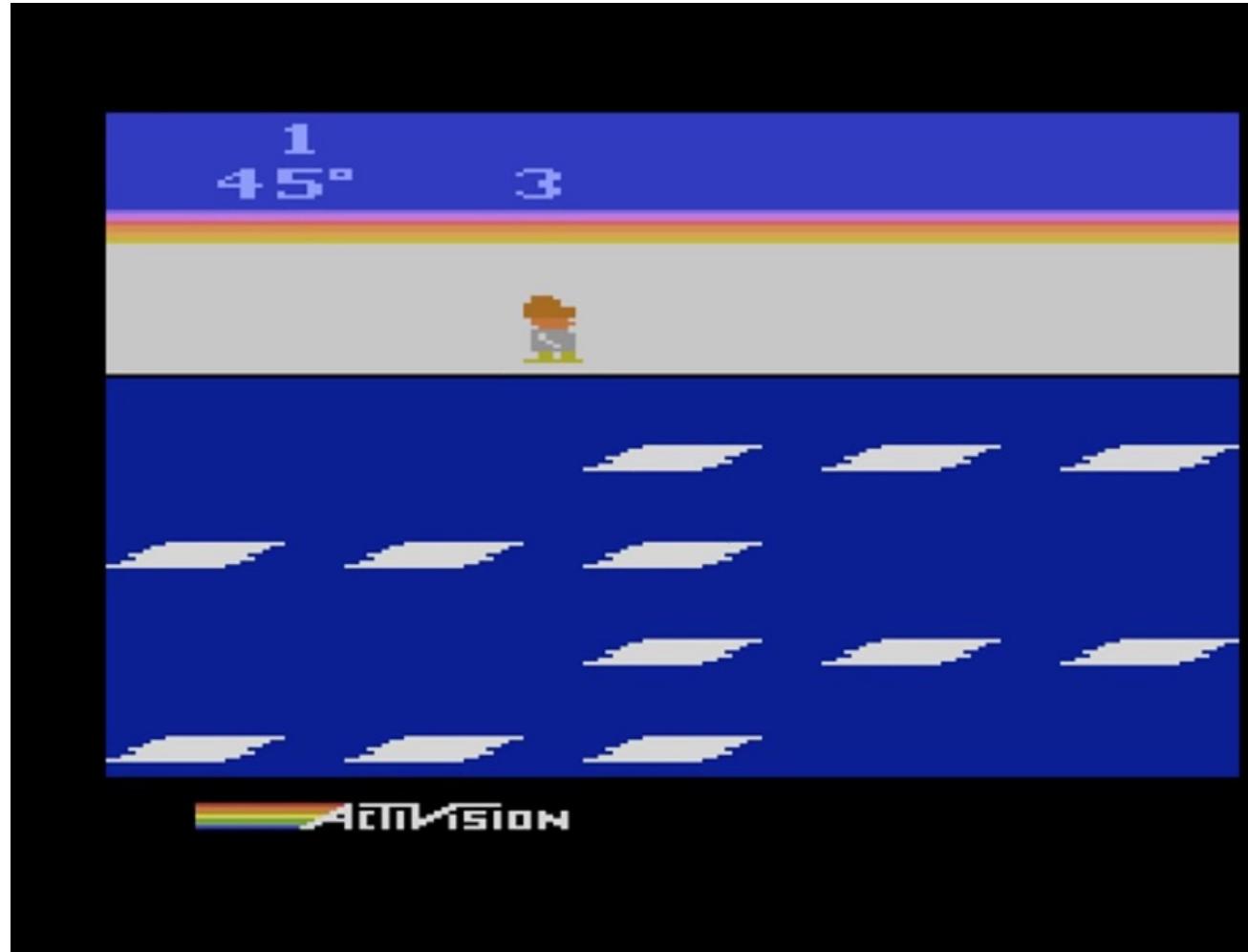


Not every Atari game can be easily beaten by RL

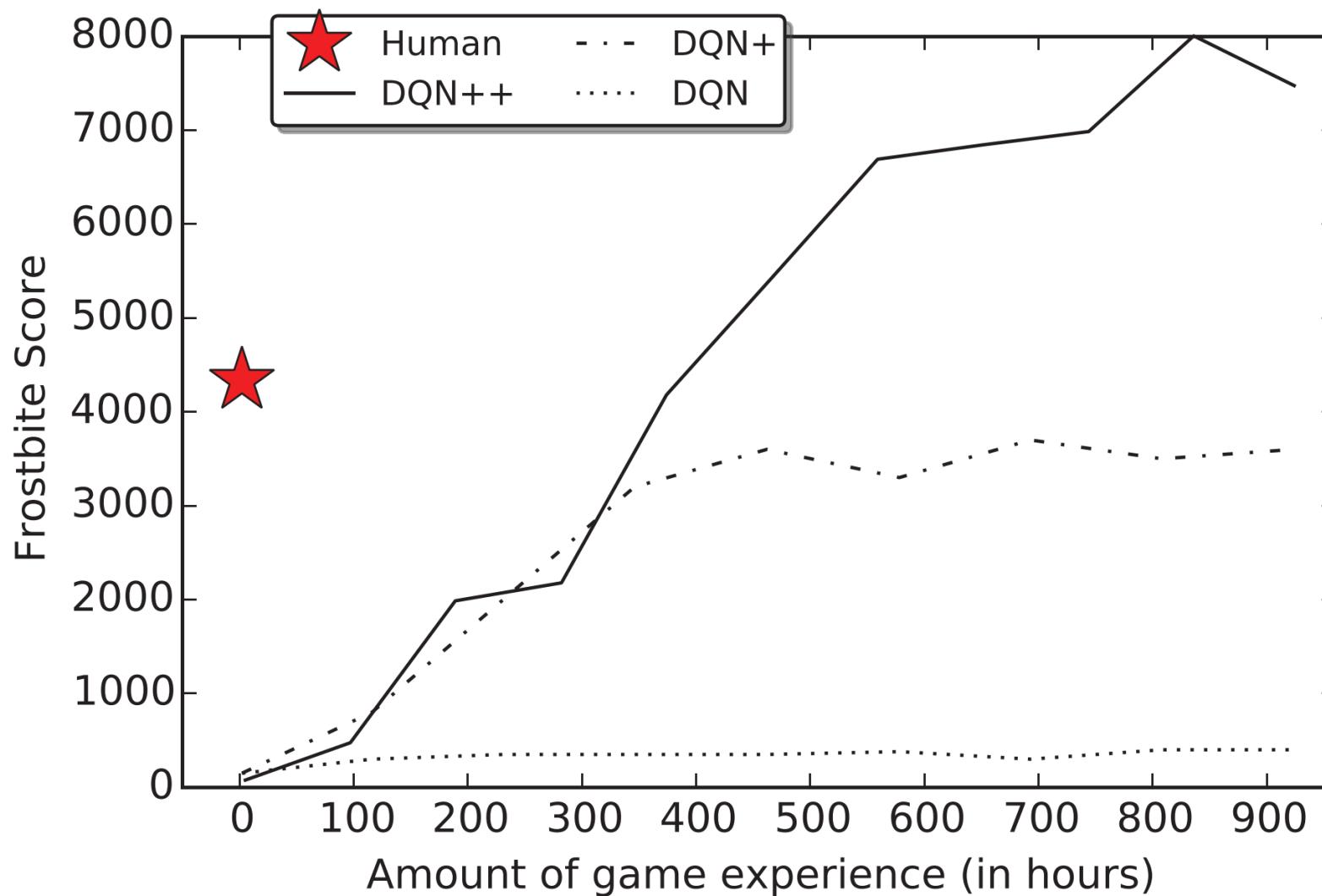


Not every Atari game can be easily beaten by RL

- Frostbite



Human vs RL (Deep Q-learning Network)



Why do humans learn so much faster

- You can try to play it yourself
- For the sake of time, let's watch a successful game play



Why do humans learn so much faster

- For those who have never played this game, what can you learn from this game play?



Why do humans learn so much faster

- Generalization of prior knowledge (objects, physics)
- We can learn abstract knowledge about how this game works
- Exploration/experimentation guided by hypotheses



Not just pixels but abstraction:
Character, ice, birds,
Igloo

Dynamics and rules:
Jump, falling
Touching ice → building igloo
Setting into igloo → next level
Avoid birds

...

Why do humans learn so much faster

- Knowledge, thought → language
- Language → knowledge, thought

Project idea: learning to play a game (or perform a task) faster by learning **necessary knowledge** from **language** guidance

Rule of the game (from Wikipedia):

- The bottom two thirds of the screen is water with four rows of ice blocks floating horizontally. The player moves by jumping from one row to another while avoiding foes. Landing on a row of white ice changes it to blue and adds a piece to the igloo on the shore. When all the ice is blue, it reverts to white. After 15 pieces, the igloo is complete, and the player must enter it to end the level.
- The direction ice is flowing can be changed via the joystick button, but it costs a piece of an incomplete igloo.
- Occasional fish can be collected for points.
- Each level must be completed in 45 seconds (represented as the declining temperature). The faster a level is completed the more bonus points are awarded. Levels alternate between large ice blocks and little ice pieces. The large blocks have gaps that the player can fall in. Starting in the fourth level, a polar bear prowls the shore.

Principles behind human cognition

- Reasoning robustly across a wide range of domains
- Learning so much from so little

From Cognitive Science: A generative model in human minds



Developmental Science
Core knowledge

Elizabeth S. Spelke and Katherine D. Kinzler



Rational quantitative attribution of beliefs, desires and percepts in human mentalizing

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe and Joshua B. Tenenbaum*

Look ahead

- What is a world model?
- What is an agent model?
- How can we conduct model-based reasoning?
- How can we integrate model-free ML models (e.g., LLMs) with model-based reasoning? Robust + efficient + scalable

Logistics

- Canvas site for readings, lecture notes, assignment submissions, and discussions
- No required textbooks. Some recommendations:
 - Kevin Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press (2012).
 - Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Fourth Edition, Pearson Education (2021).
 - Noah D. Goodman, Joshua B. Tenenbaum, and ProbMods Contributors, *Probabilistic Models of Cognition*, <https://probmods.org/> (2016)

Logistics

- 4 Problem sets:
 - Coding, Python + Julia (for a probabilistic programming language, Gen)
 - Experience with cognitive modeling for reasoning tasks
 - You can discuss with your peers but
 - there needs to be an acknowledgment of contributions
 - you **have to write your own code and analysis**
 - DO NOT COPY writeups

Logistics

- Project:
 - Modeling, evaluation, analysis
 - Short write-up (4 – 8 page conference format)
 - Presentation
 - Standards:
 - For undergraduate students, implementing & analyzing an exiting model
 - For graduate students, proposing an original model
 - AI models inspired by principles of human cognition (if unsure, talk to me in advance)
 - Not a standard ML project
 - Cannot be training a standard ML model on a standard benchmark
 - Cannot be *just* evaluating a standard ML model (e.g., GPT-4) – you have to also propose an alternative model using ideas and tools discussed in this course
 - I will suggest some project ideas throughout the lectures
 - Collaboration is allowed, particularly if it is an interdisciplinary collaboration
 - 1 – 3. With more team members, you will be expected to do more.

Logistics

- Problem sets: 40%
- Project: 60%
 - Presentation: 20%
 - Paper: 40%