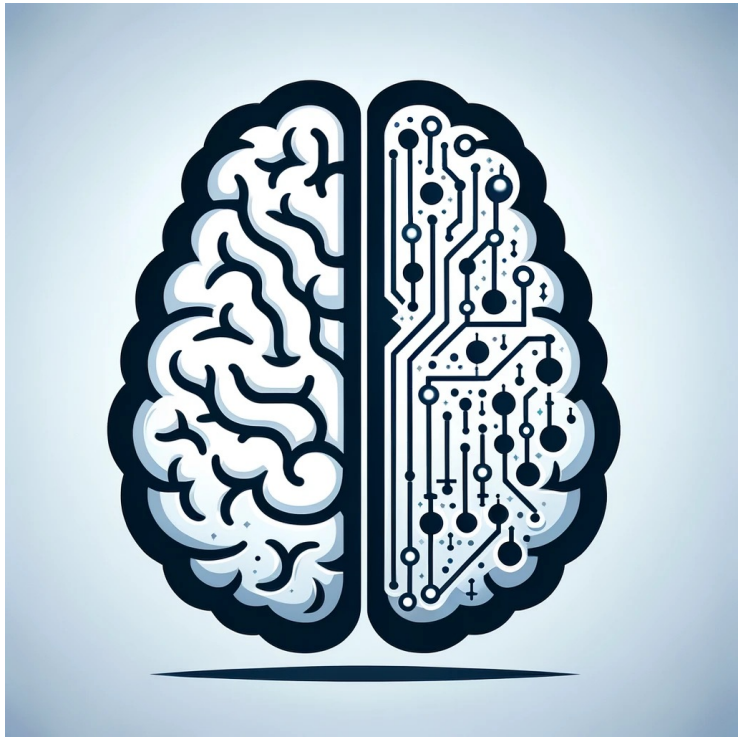
EN 601.473/601.673: Cognitive Artificial Intelligence (CogAI)



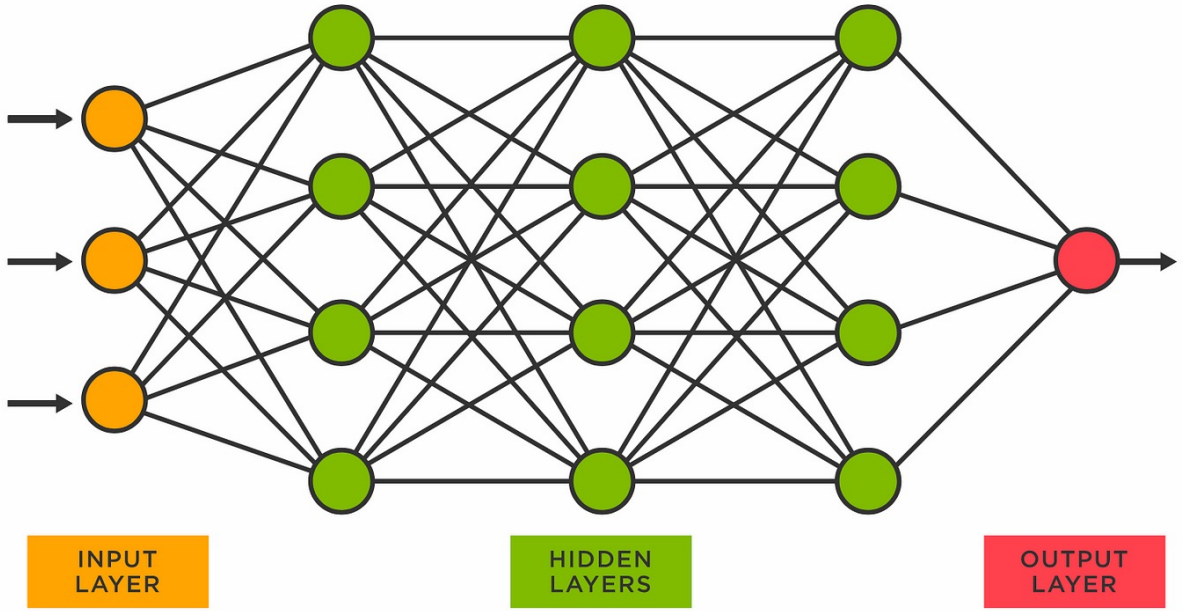
Lecture 2:
World & agent models
The problem of induction

Tianmin Shu

Recap: Limits in model-free ML models

Fuel: large-scale datasets

Engine: deep neural networks



- LLMs
- Self-driving cars
- Model-free RL

- Human reasoning
- Human learning

$$y = f(x)$$

Billions of parameters or even more

Recap: a generative model and model-based reasoning

Developmental Science
Core knowledge

Elizabeth S. Spelke and Katherine D. Kinzler

nature human behaviour

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

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

Clarification on a few items

- Auditing?
  - Yes if there is still space (a few vacant seats even no one drops the class)
  - I will offer this course again in the Fall this year

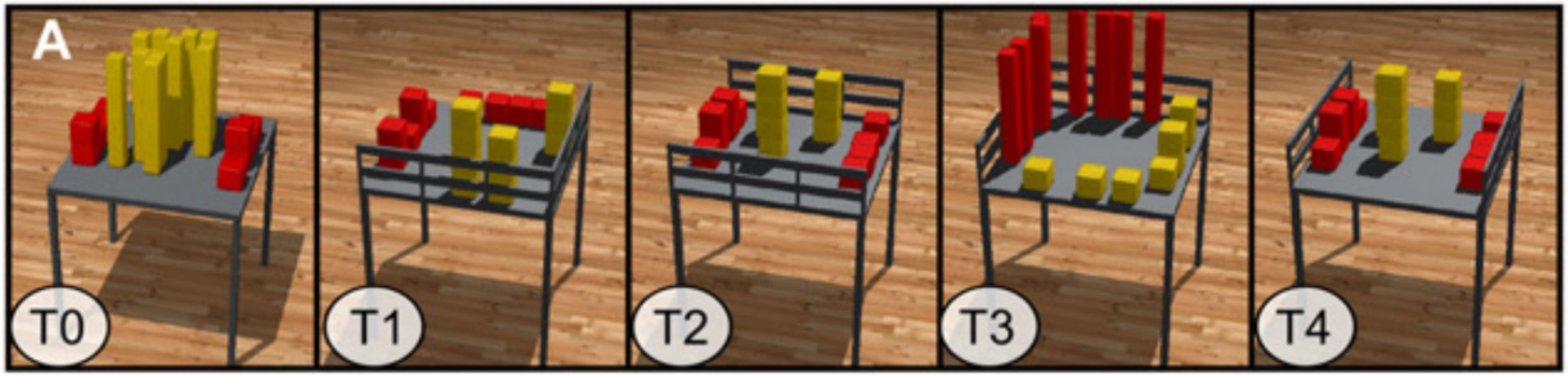
- Example projects?

Example project 1: Intuitive physics

Simulation as an engine of physical scene understanding

Peter W. Battaglia[1], Jessica B. Hamrick, and Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139



Intuitive physics reasoning with language

GROUNDED PHYSICAL LANGUAGE UNDERSTANDING WITH PROBABILISTIC PROGRAMS AND SIMULATED WORLDS

Cedegao E. Zhang[1], **Lionel Wong**[1], **Gabriel Grand**[2] **& Joshua B. Tenenbaum**[1,2]
[1]BCS, MIT [2]CSAIL, MIT
{cedzhang, zyzzyva, grandg, jbt}@mit.edu

Human intuitive physics language benchmark

Scenario: Imagine there is a table with some blocks on it; blocks can be red or yellow.

There is one tall stack of yellow blocks on the left edge of the table, and there are no red blocks on the right edge.

Question: If the table is bumped hard enough to knock at least one of the blocks onto the floor, are there going to be more **red blocks** or **yellow blocks** on the floor?

Intuitive physics reasoning with language

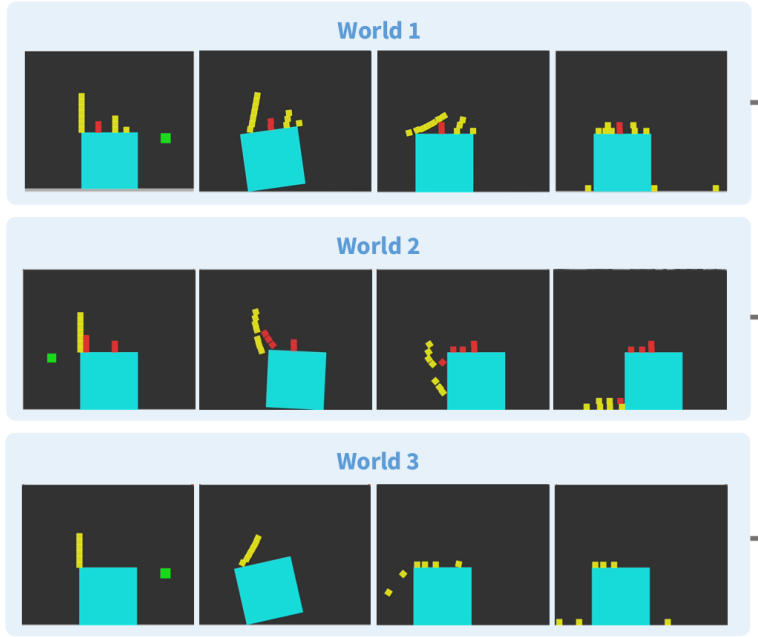
Physics in a Language of Thought (PiLoT)

There is one tall stack of yellow blocks on the left edge of the table.

```
condition(
  filter(isOnEdge,
  filter(isOnLeft,
  filter(isTall,
  filter(isYellow, world.stacks)
))).length == 1)
```

There are no red blocks on the right edge.

```
condition(
  filter(isOnEdge,
  filter(isOnRight,
  filter(isRed, world.stacks)
)).length == 0)
```

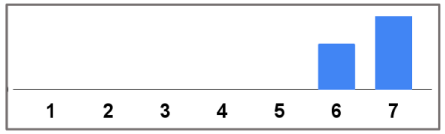


Answer: 6 – Very likely more yellow blocks

Answer: 6 – Very likely more yellow blocks

Answer: 7 – Definitely more yellow blocks

Are there going to be more ***red blocks*** or ***yellow blocks*** on the floor?

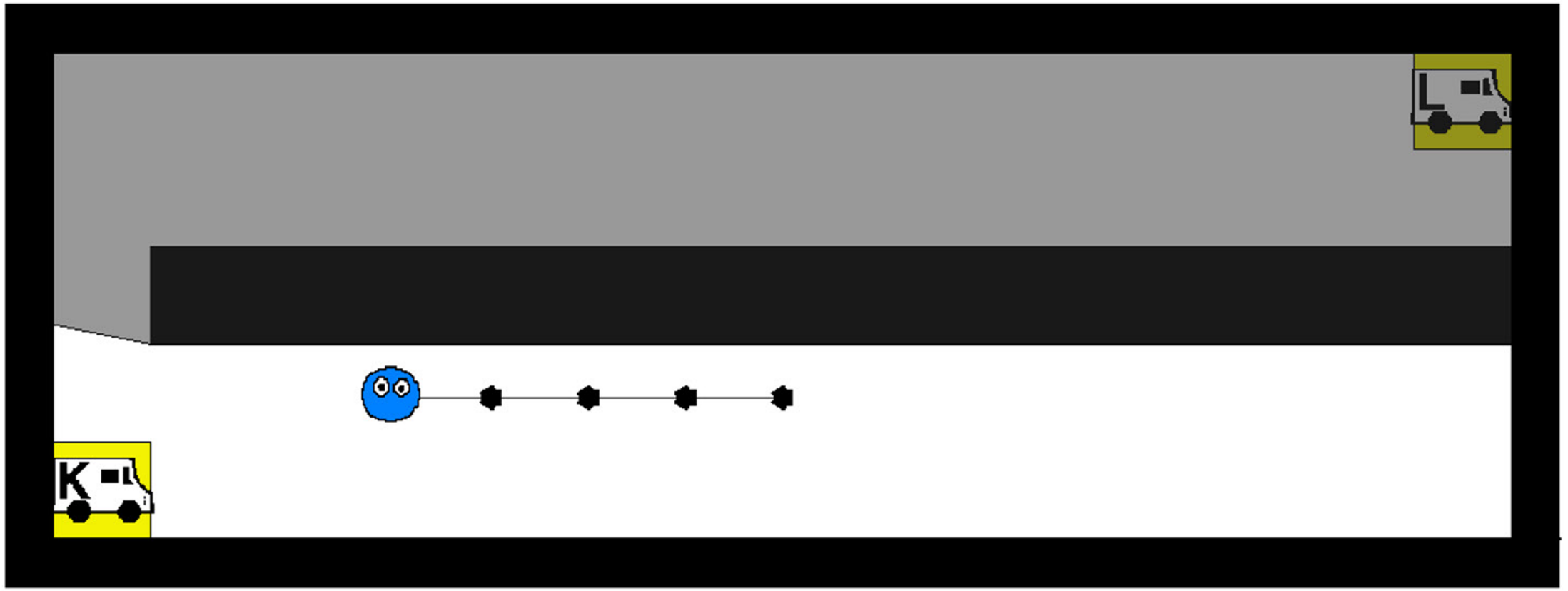


*Distribution over **model answers** reflects probabilistic physical reasoning over possible scenes*

Translate: *Semantics into probabilistic program conditions using a large language model (LLM) trained on code*

Simulate: *Sample possible worlds from conditioned probabilistic program with physics simulation engine*

Example project 2: Theory of Mind from videos



Baker et al. (2017)

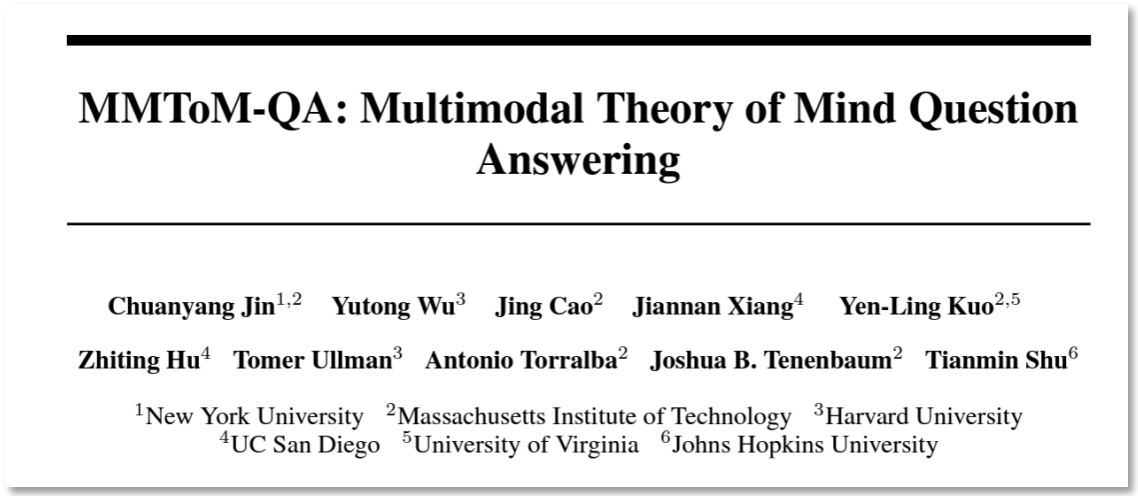
Title: Theory of Mind reasoning from videos
Category: intuitive psychology
Contact: Tianmin Shu (tshu@mit.edu)
How many students: 1

Description:
Prior work on Bayesian Theory of Mind was typically conducted in simple 2D grid worlds, such as the Food Truck experiments in [1]. In this project, you will have a chance to build models that can infer humans' mental states from more complex stimuli. In particular, we can synthesize videos of household activities (such as setting up a dinner table) in a realistic virtual simulation environment, VirutalHome [2]. As shown in Figure 1, we can conduct similar Theory of Mind reasoning tasks based on such videos, bridging the gap between low-level visual perception and high-level psychological reasoning. One of the challenges here is building probabilistic inference methods that can overcome noise and errors from low-level visual perception (such as object detection errors caused by occlusion). Another challenge is to create a dataset to systematically evaluate different kinds of models' capacity to infer humans' mental states from visual inputs in complex and realistic environments (such as homes). In this project, you can have a chance to work on one or both challenges.

Figure 1. An example video sequence where the captions describe the event in each frame. Note that the orange bounding box in the first frame indicates the bottle of wine on the table. After watching this video, we can ask people or models this question – "If Alice has been trying to get a bottle of wine, did she prefer it to be cold?" Similar to the original Food Truck experiments in [1], answering such questions requires a successful inference of Alice's beliefs and desires by watching her actions. However, in this project, you will have a chance to work on a real-world version of the Food Truck experiments, investigating Bayesian Theory of Mind in realistic household environments.

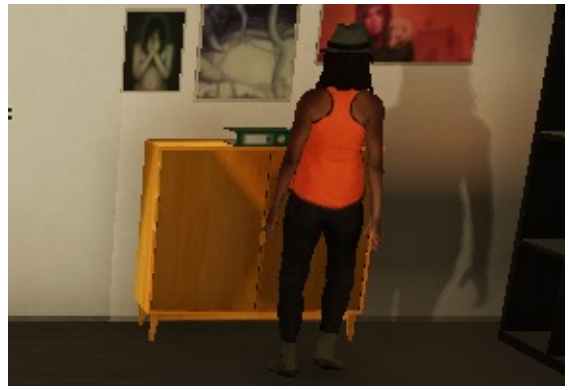
References:
[1] Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour, 1(4), 1-10.
[2] Puig, X., Shu, T., Li, S., Wang, Z., Liao, Y. H., Tenenbaum, J. B., ... & Torralba, A. (2020). Watch-and-help: A challenge for social perception and human-ai collaboration. arXiv preprint arXiv:2010.09890.

MMToM-QA Multimodal Theory of Mind Question Answering



MMToM-QA: Multimodal Theory of Mind Question Answering

Chuanyang Jin[1,2] Yutong Wu[3] Jing Cao[2] Jiannan Xiang[4] Yen-Ling Kuo[2,5]
Zhiting Hu[4] Tomer Ullman[3] Antonio Torralba[2] Joshua B. Tenenbaum[2] Tianmin Shu[6]
[1]New York University [2]Massachusetts Institute of Technology [3]Harvard University
[4]UC San Diego [5]University of Virginia [6]Johns Hopkins University



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.

Outline

- At a high level
  - What is a world model (in humans and machines)?
  - What is an agent model (in humans and machines)?
  - How can we conduct model-based physical and social reasoning?

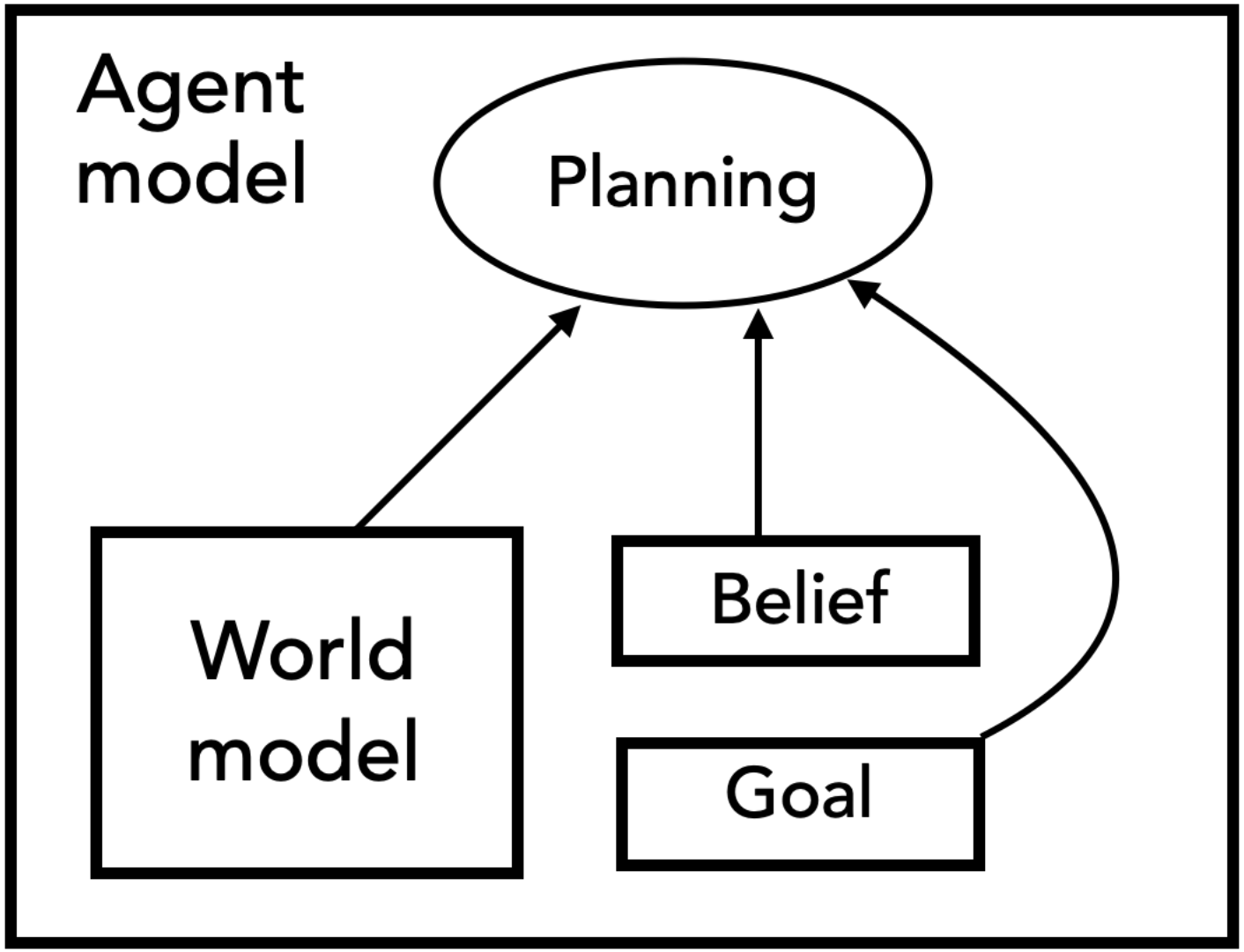
- The problem of induction

Readings

- *Chapter 1, Vision*, Marr
- *Building machines that learn and think like people*, Lake et al.., Behavioral and brain sciences (2017)
- *Probabilistic machine learning and artificial intelligence*, Ghahramani, Nature (2015)

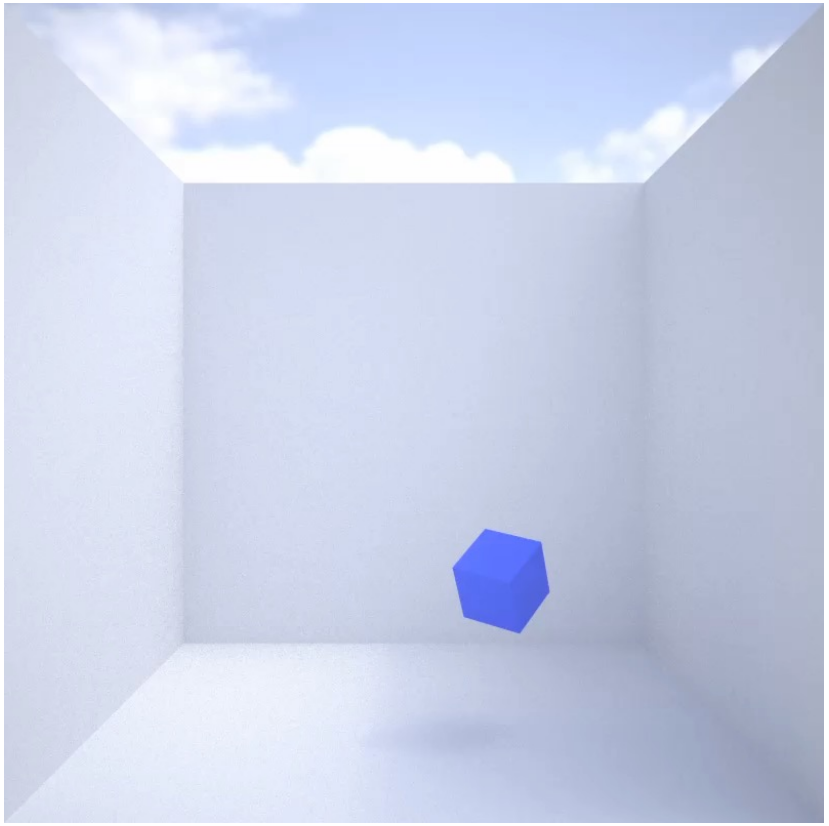
- (Optional) Probability refresher notes

Model-based reasoning w/ world and agent models



World models in humans

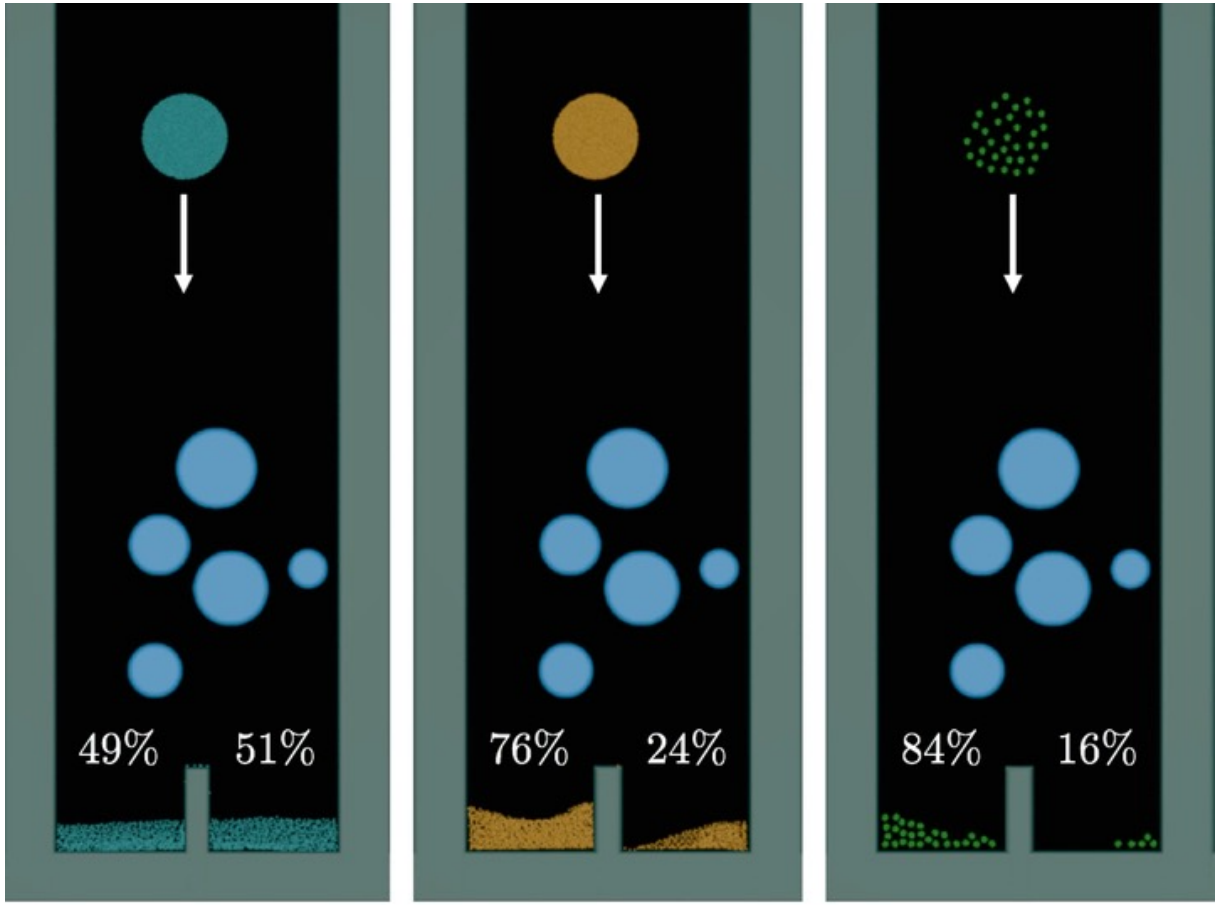
- Perceiving physical properties (e.g., materials, viscosity)



Stimuli from Vivian Paulun

World models in humans

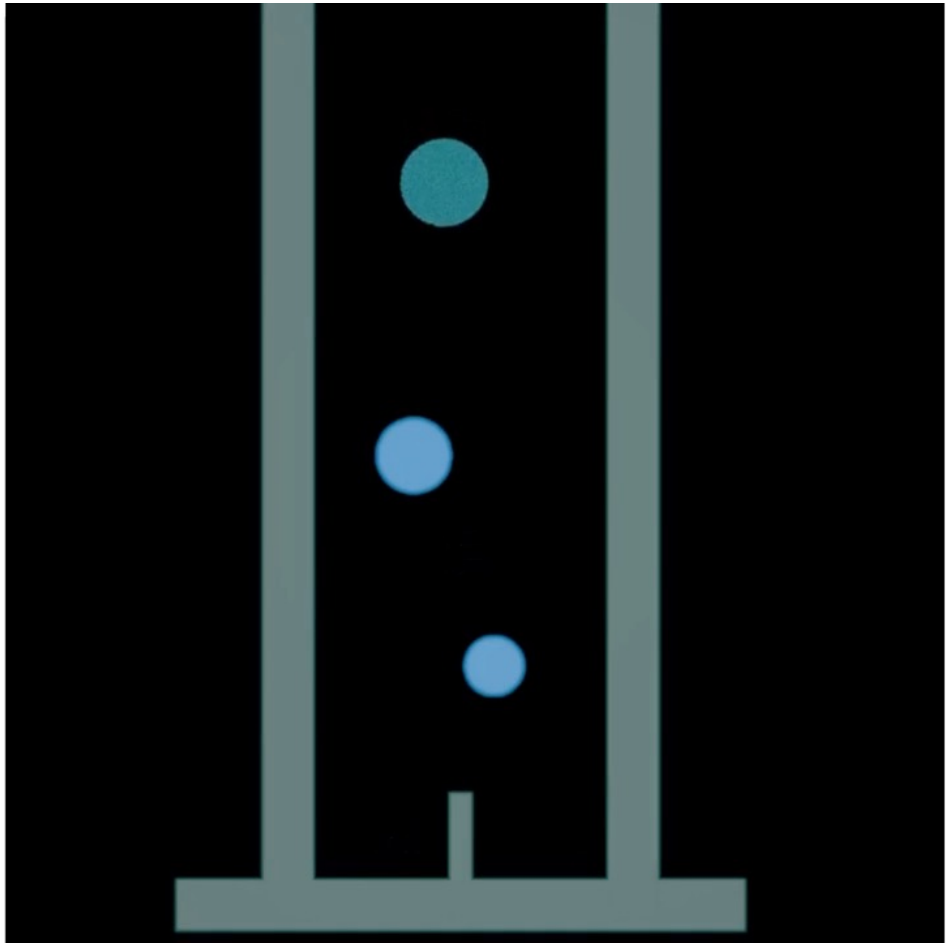
- Predicting dynamics



Kubricht et al. (2017)

World models in humans

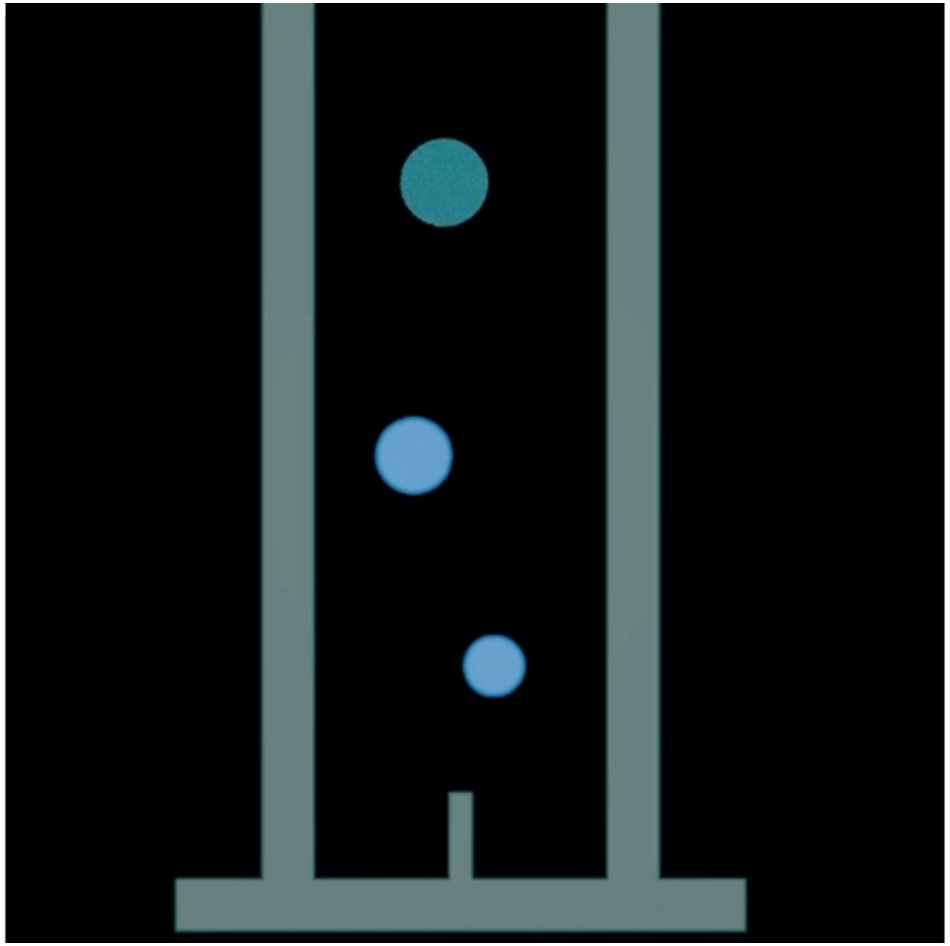
- Predicting dynamics



Kubricht et al. (2017)

World models in humans

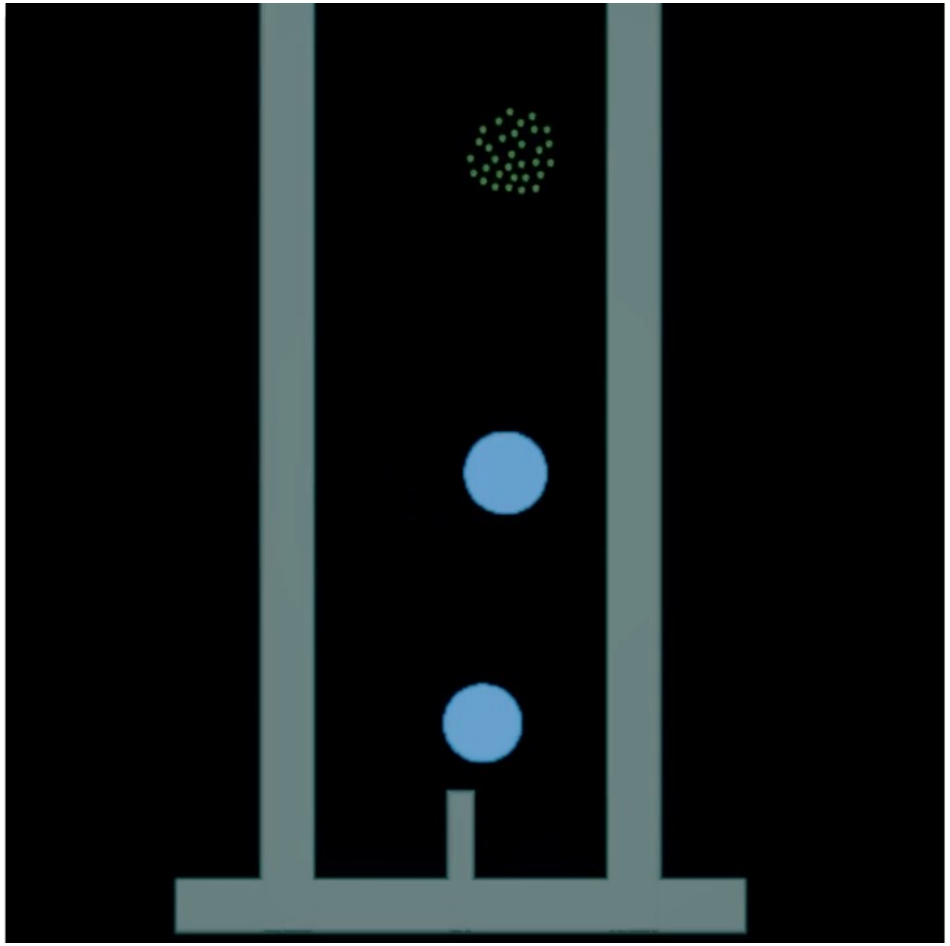
- Predicting dynamics



Kubricht et al. (2017)

World models in humans

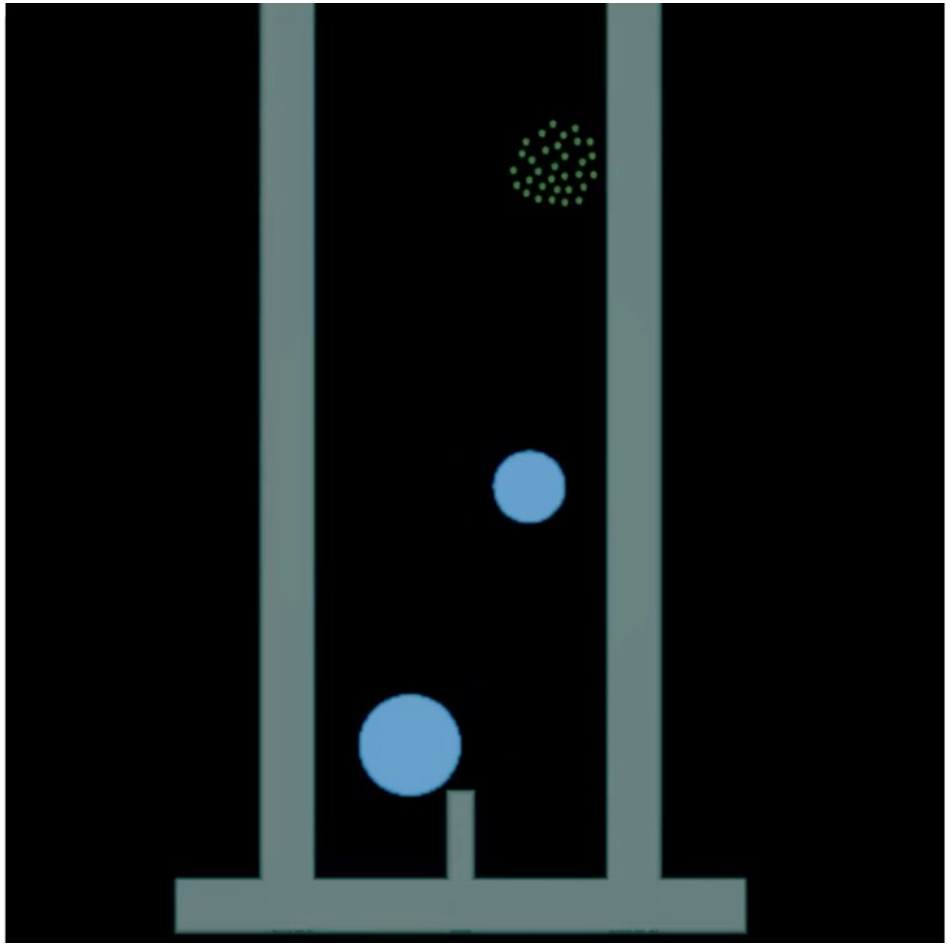
- Predicting dynamics



Kubricht et al. (2017)

World models in humans

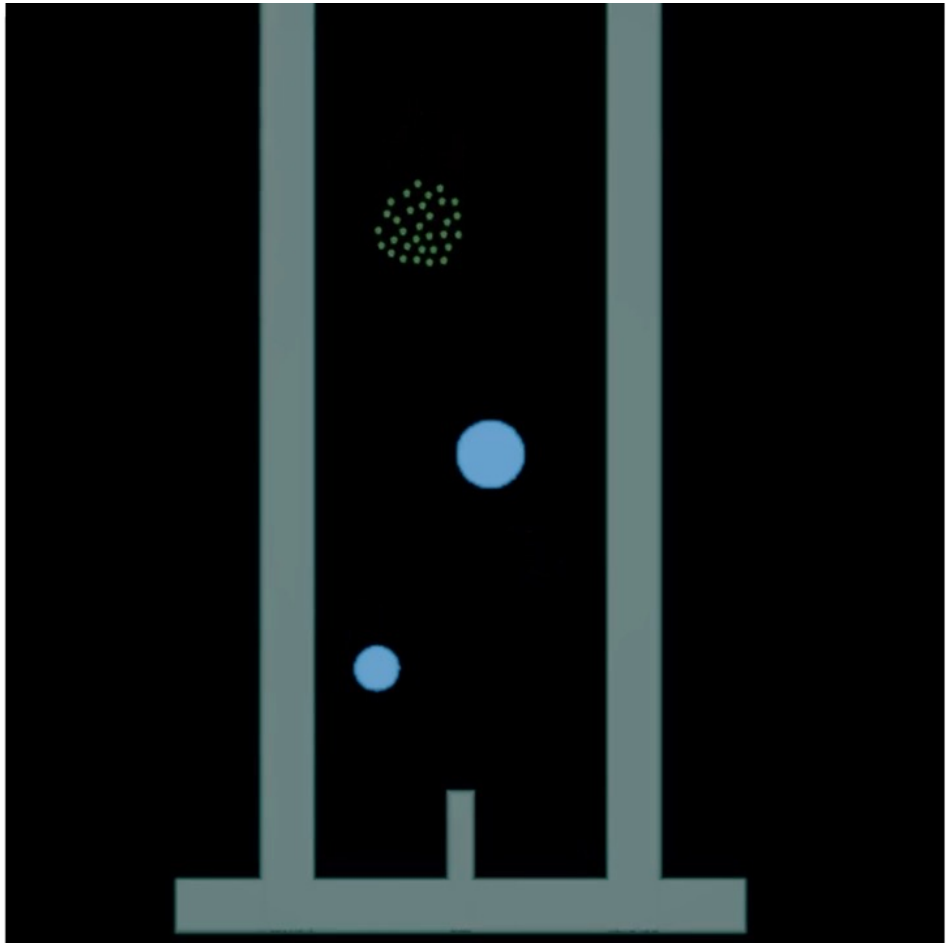
- Predicting dynamics – which container will get more substance?
- Left, right, equal



Kubricht et al. (2017)

World models in humans

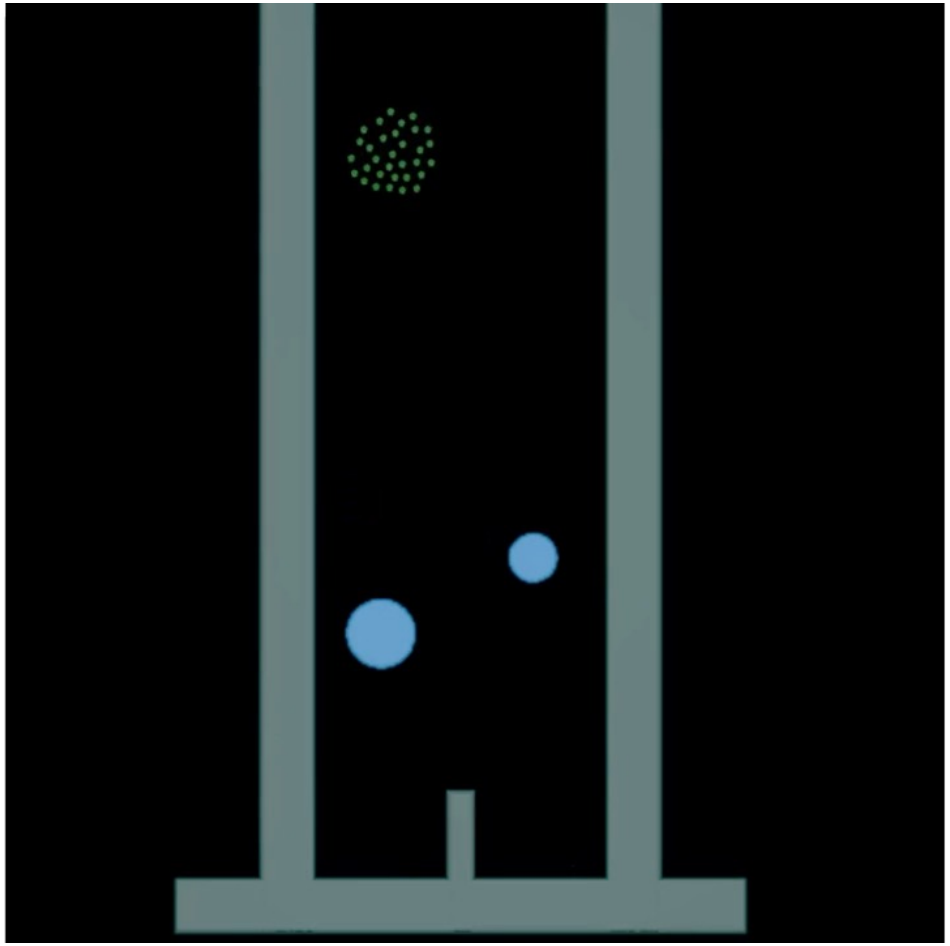
- Predicting dynamics – which container will get more substance?
- Left, right, equal



Kubricht et al. (2017)

World models in humans

- Predicting dynamics – which container will get more substance?
- Left, right, equal



Kubricht et al. (2017)

World models in humans

- Model-based control/planning

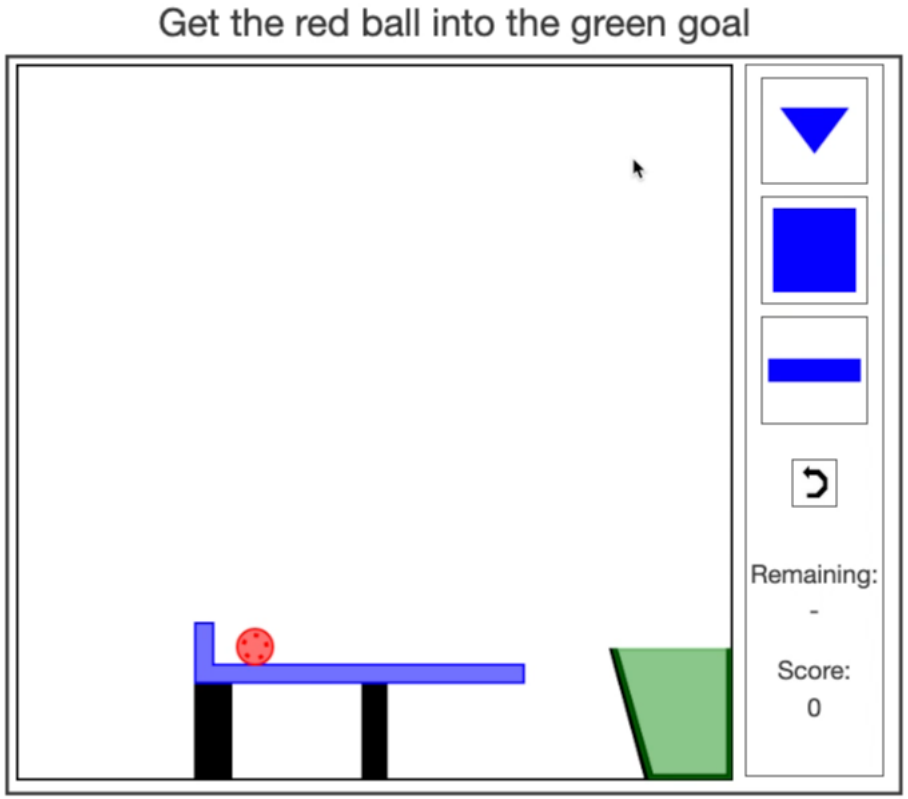
Bates et al. (2015)

World models in humans

- Model-based control/planning

Human tool use

Unlike model-free RL, humans can learn to use tools through just a few trials

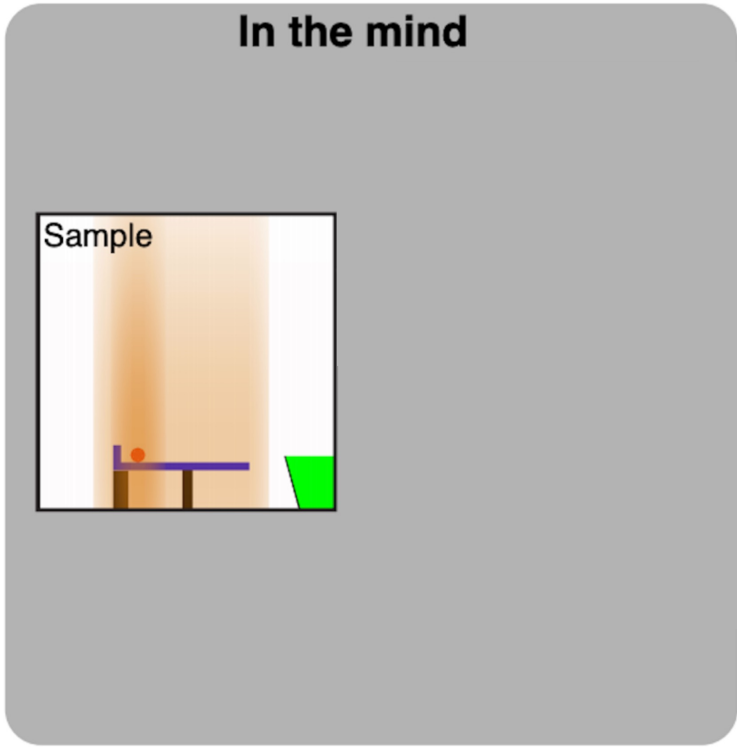


Allen et al. (2020)

World models in humans

- Model-based control/planning

Too use via model-based planning
Key is to use a world model to simulate the outcomes of possible plans

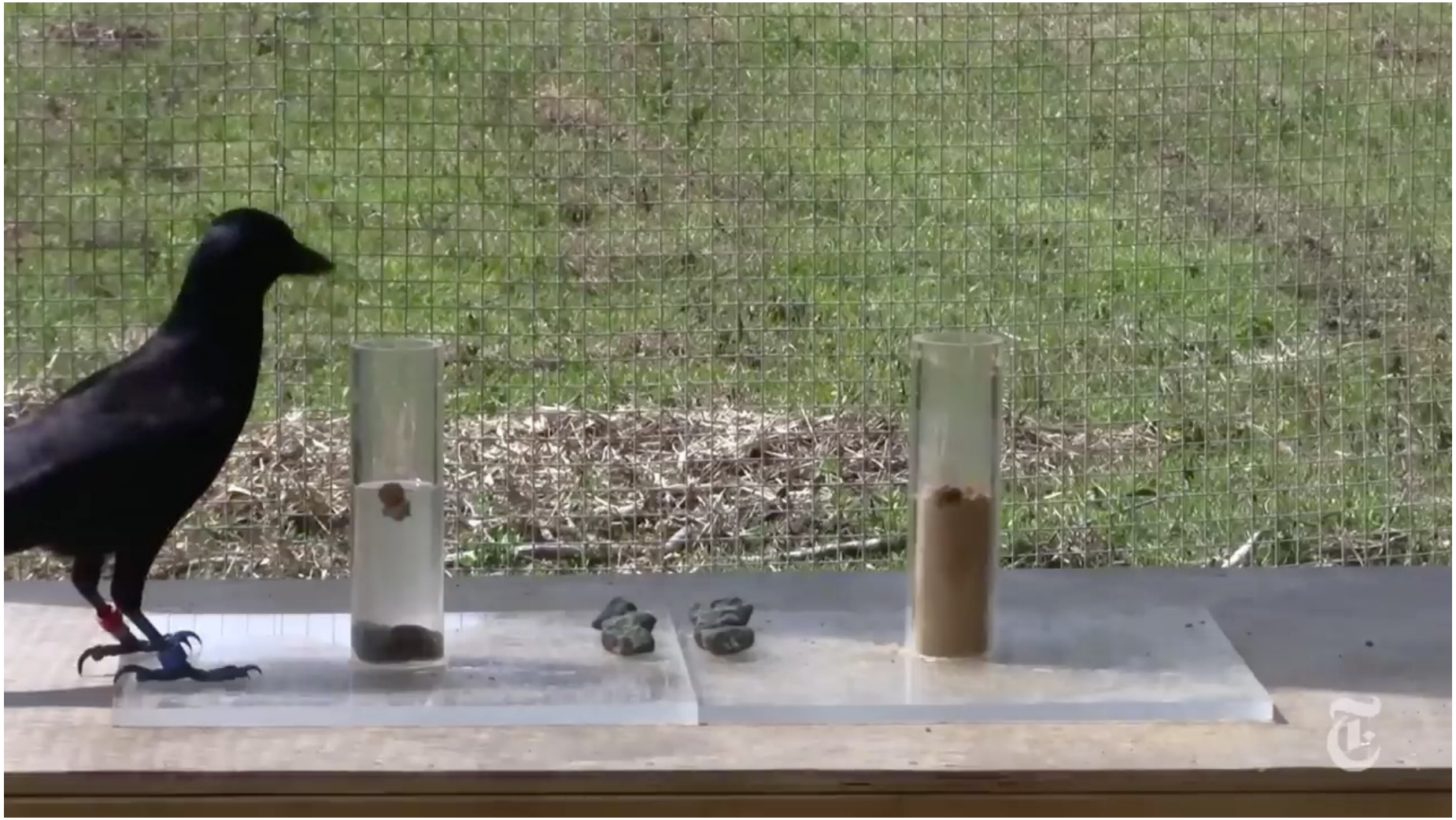


Allen et al. (2020)

World models in ANIMALS

- Model-based control/planning

Learning physical knowledge



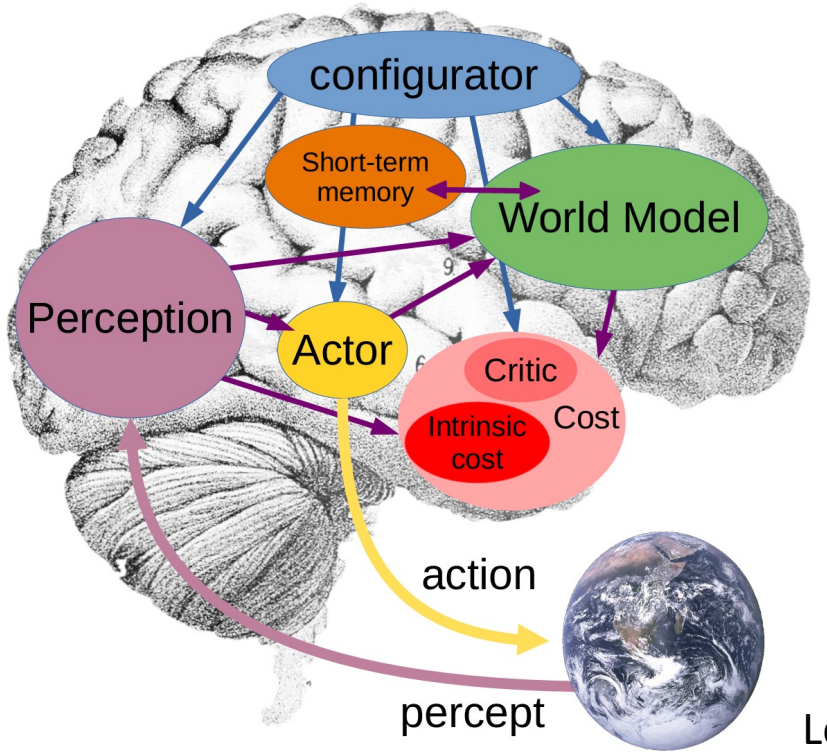
World models in ANIMALS

- Model-based control/planning

Learning to compose plans based on basic skills

World models in robotics and embodied AI

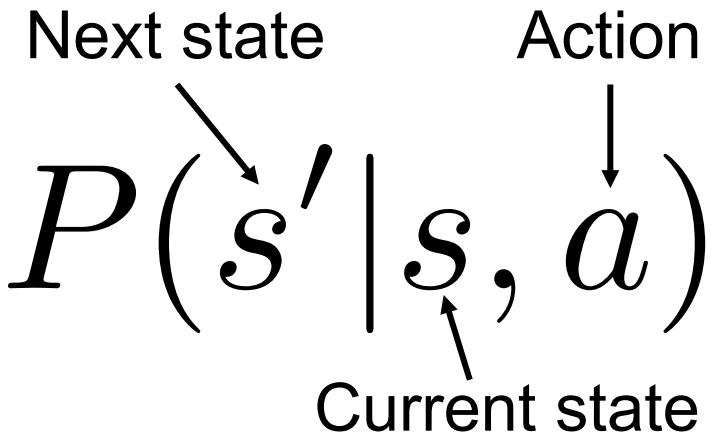
- Model-based planning
- Model-based reinforcement learning



LeCun (2022)

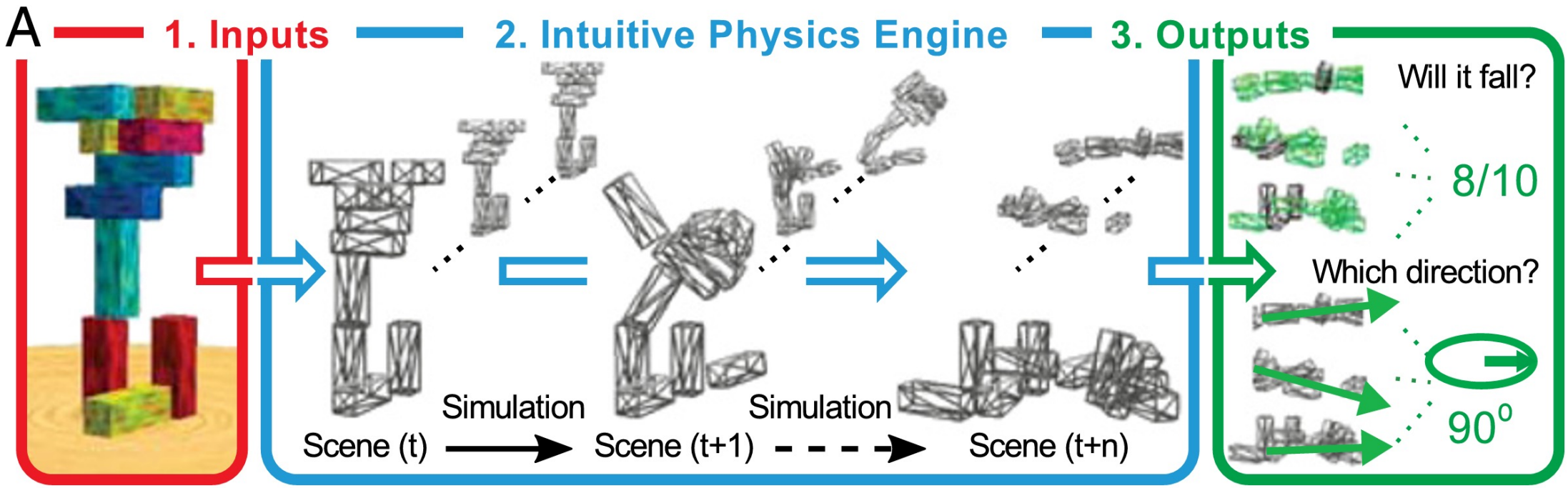
World models in robotics and embodied AI

- World model as state transition probabilities
- Causal relationship between action and state change



Simulators as world models

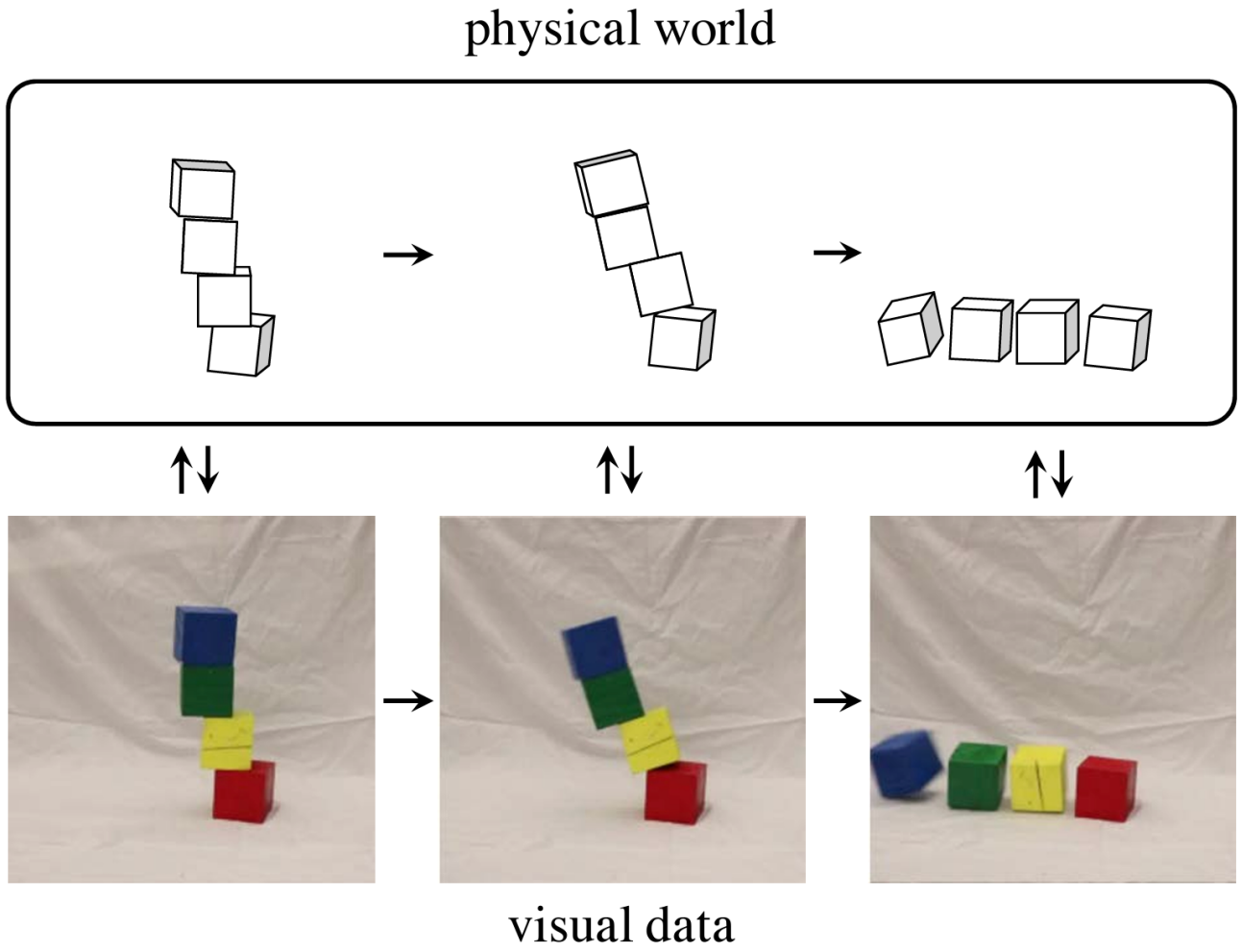
- Intuitive physics engine in human minds



Battaglia et al. (2016)

Simulators as world models

- Computer vision: model-based physical scene understanding

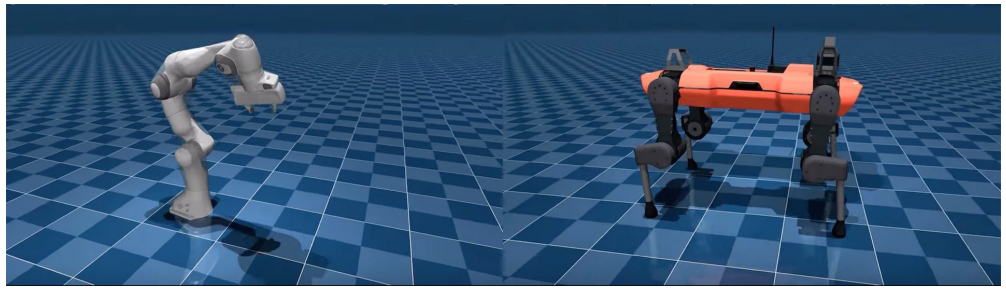


Wu et al. (2017)

Simulators as world models

- Physics engines / embodied simulators

MuJoCo



Todorov et al. (2012)

AI2-THOR

Kolve et al. (2017)

iGibson 2.0

Li et al. (2021)

Habitat 2.0

Szot et al. (2021)

ThreeDWorld

Gan et al. (2020)

Simulators as world models

- Embodied simulators + synthetic humans

VirtualHome 2.0

Puig et al. (2021)

Habitat 3.0

Puig et al. (2023)

Simulators as world models

- Create massive training data that are hard to collect in the real world

Simulator

Real World

Ehasani et al. (2023)

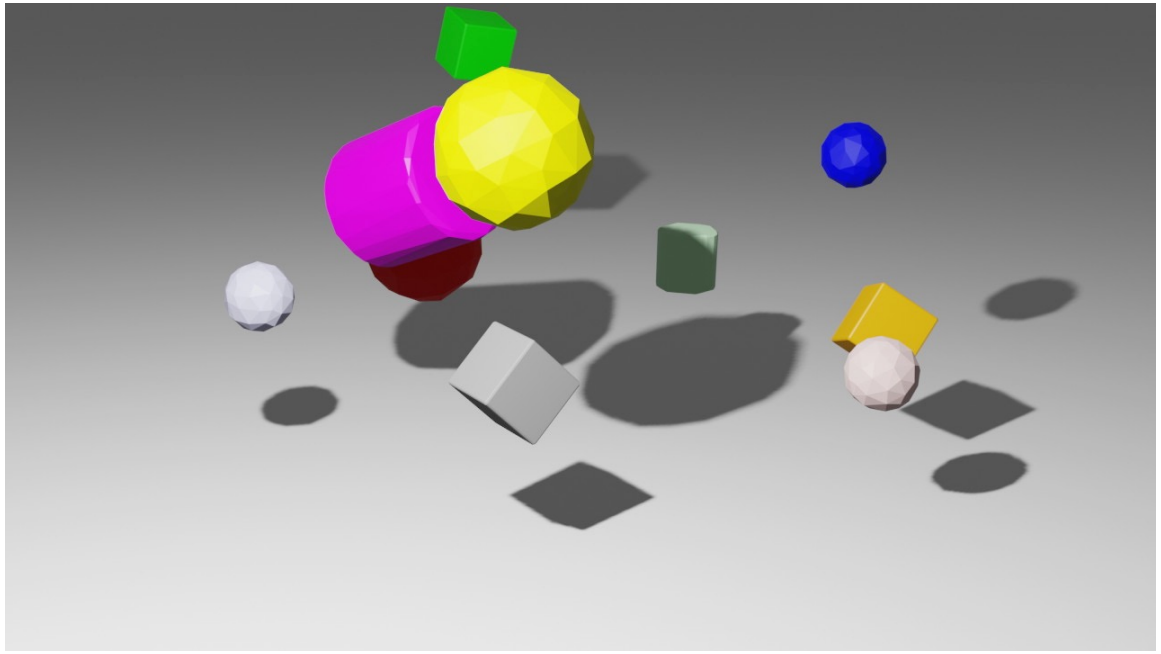
Can we learn world simulators?

- Neural physics engines

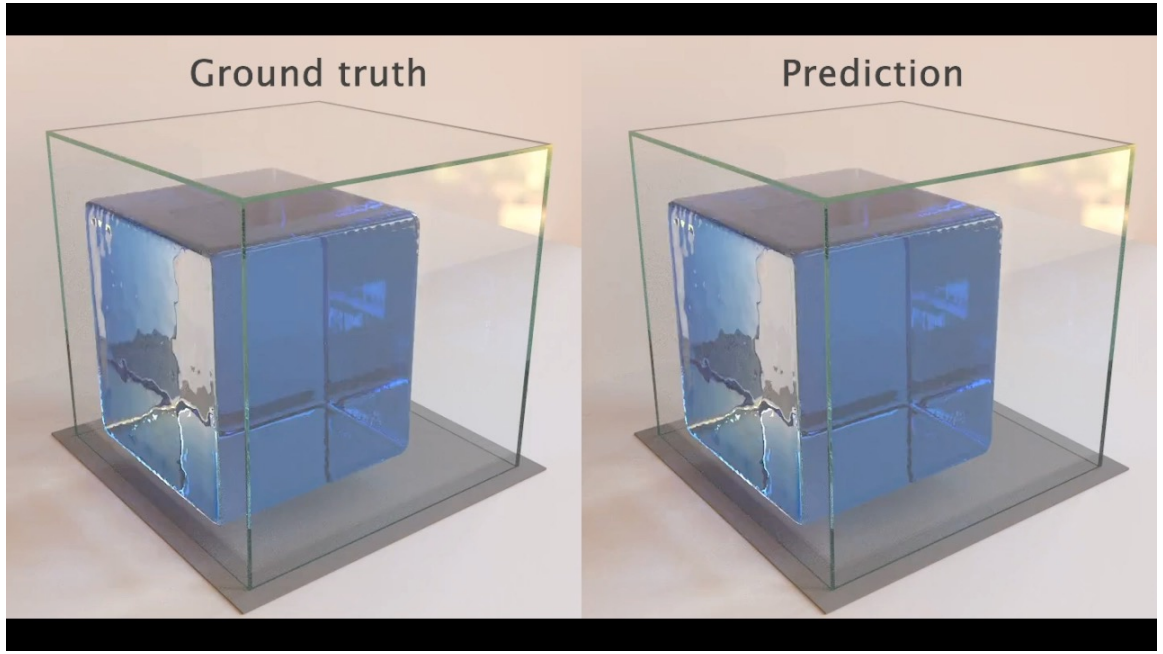
- Video prediction models

Neural physics engines

Represent a world state as objects or particles



Allen et al. (2023)



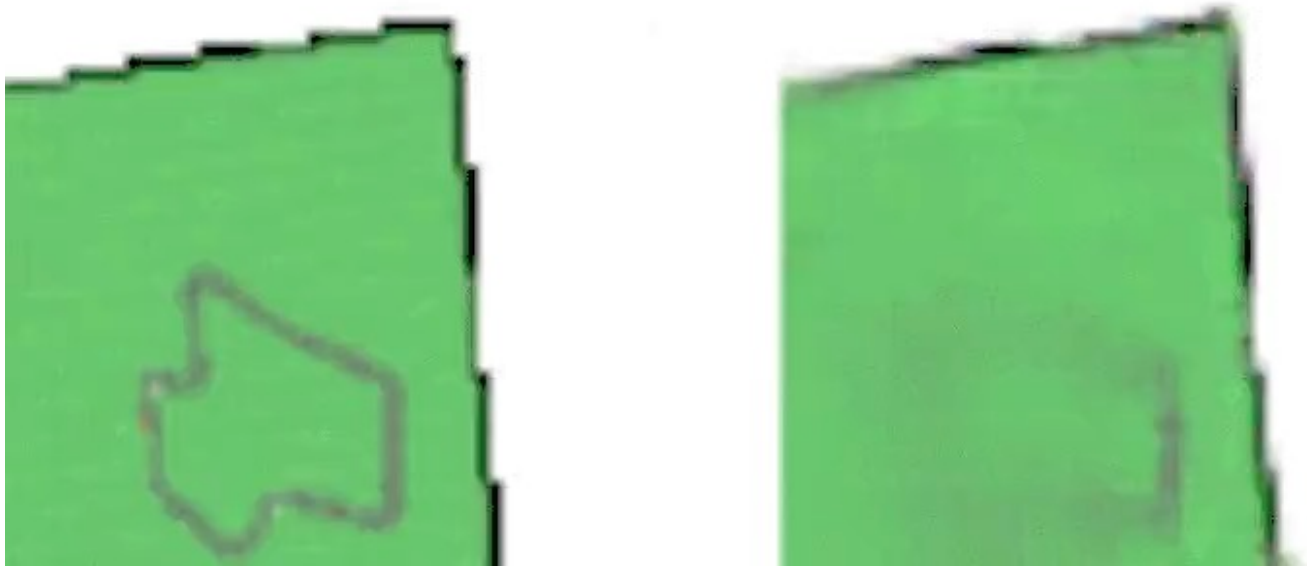
Sanchez-Gonzalez et al. (2020)

Video prediction models

Represent a world state as an image

Ground-truth

Synthesis



Ha & Schmidhuber (2018)

Video prediction models

GAIA-1 from Wayve

Video prediction models

UniSim: Learning Interactive Real-World Simulators

Sherry Yang[12], Yilun Du[3], Kamyar Ghasemipour[2], Jonathan Tompson[2], Dale Schuurmans[2], Pieter Abbeel[1]

[1] UC Berkeley [2] Google DeepMind [3] MIT

Video prediction for robot planning

Simulating long sequence of robot executions.

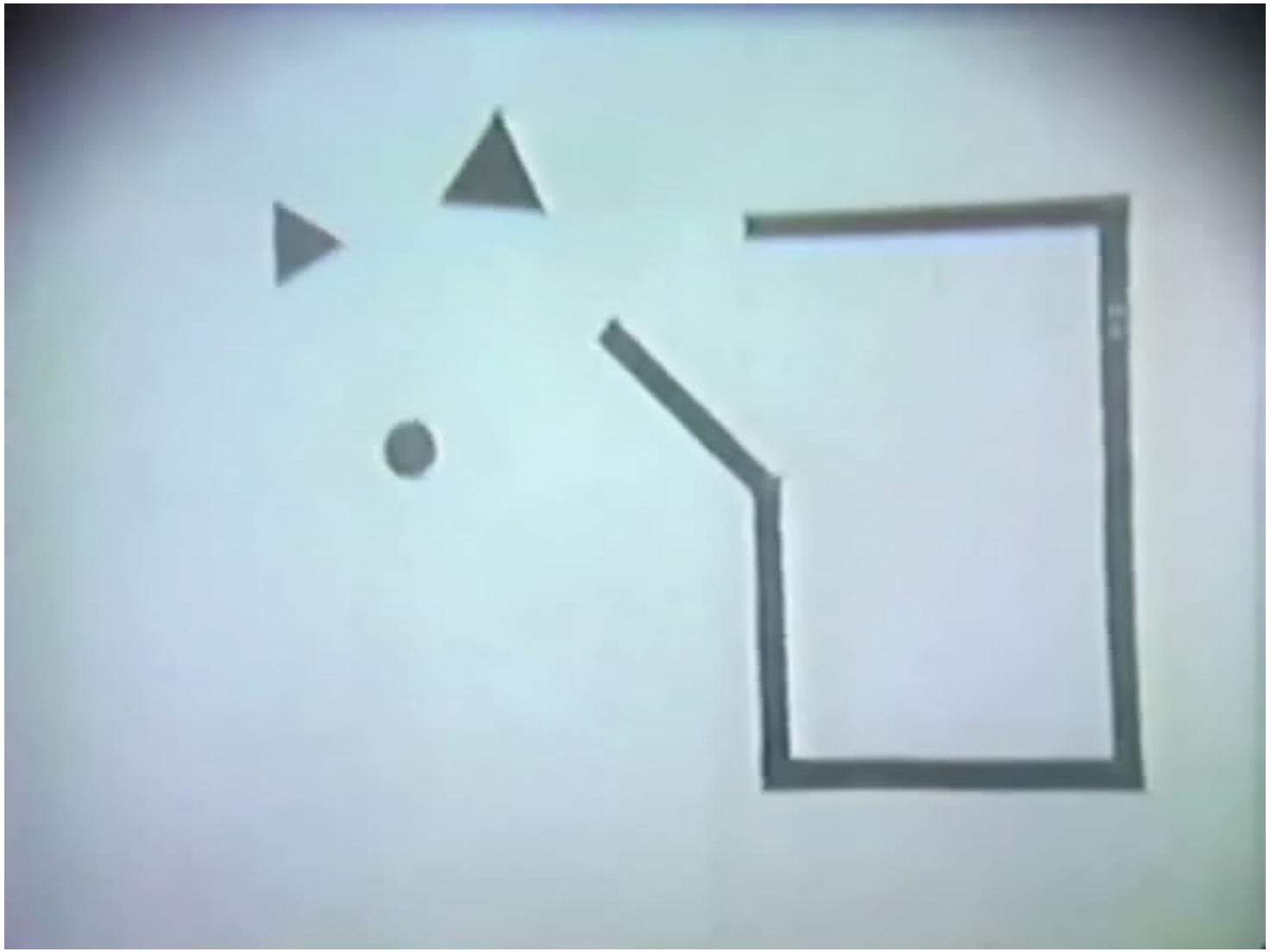
Step 1:

Yang et al. (2023)

Current world models are typically domain specific

- A physics engine for fluid simulation
- A simulator of the household environments
- A video prediction model for driving
- A video prediction model for table top object manipulation
- …

Humans represent agents differently from objects



Heider & Simmel (1944)

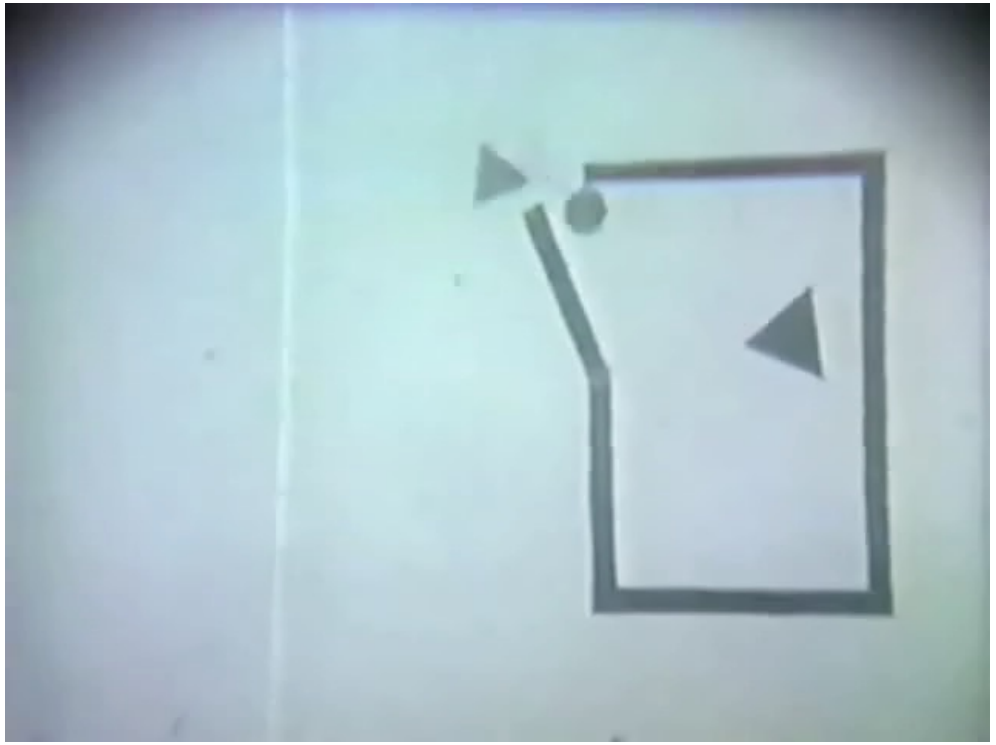
Humans represent agents differently from objects

Strengths
strong, weak

Goals
helping, hurting, escaping

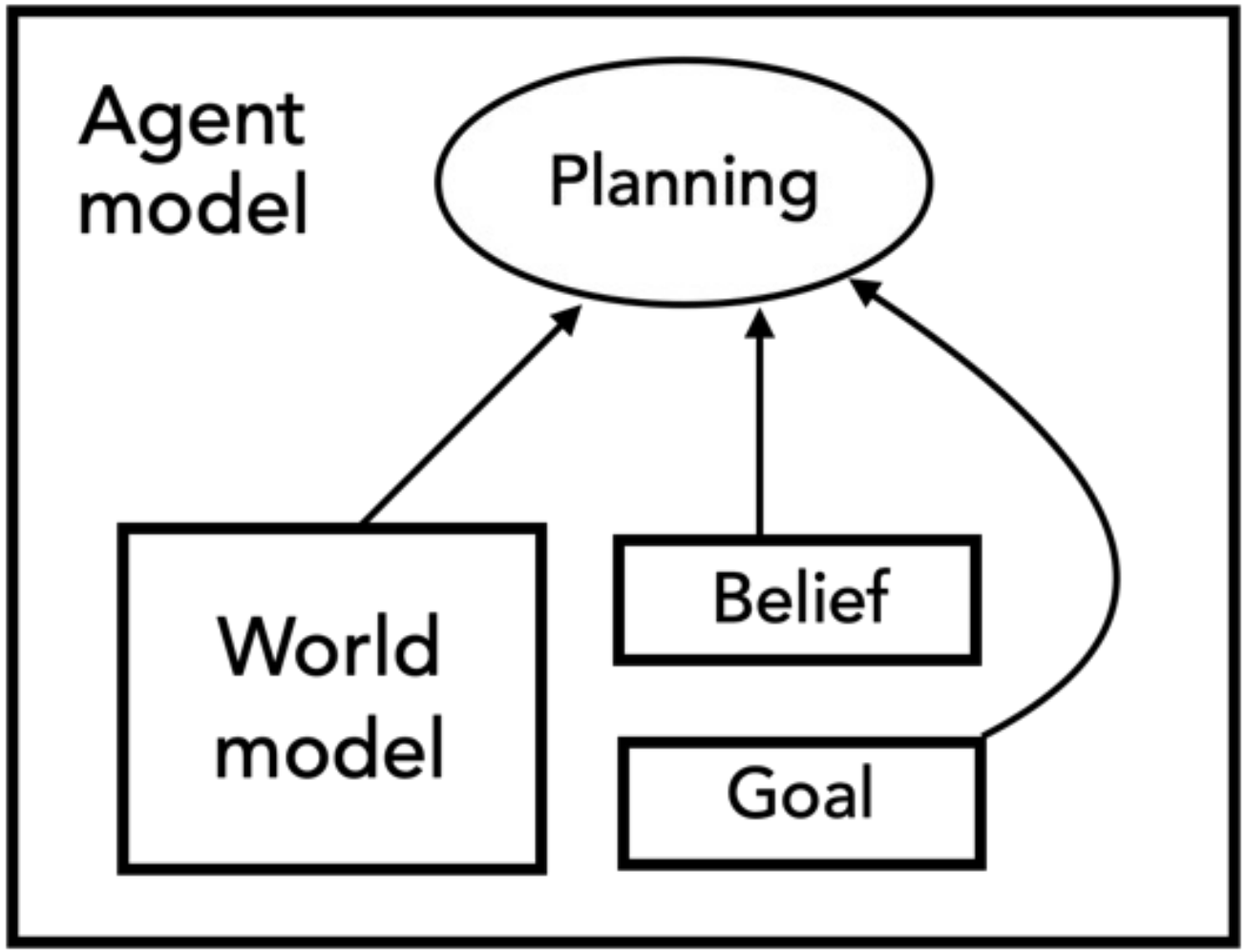
Relationships
friends, enemies

Moral judgment
good guy, bully



(size / velocity / angle…)
A big triangle moves back and forth, while a small triangle and a small circle rotate 360°…

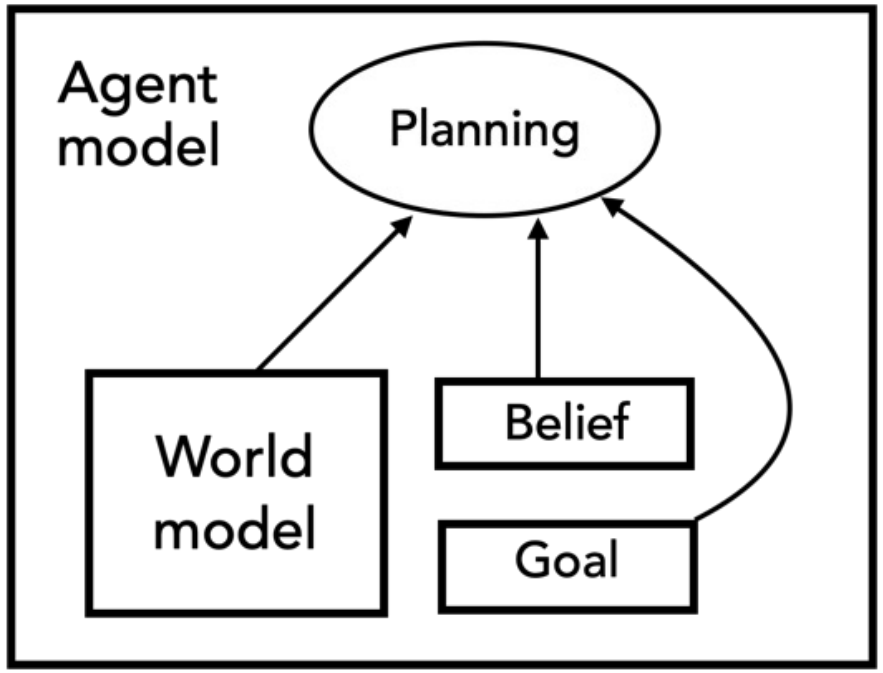
The minimum definition of an agent model

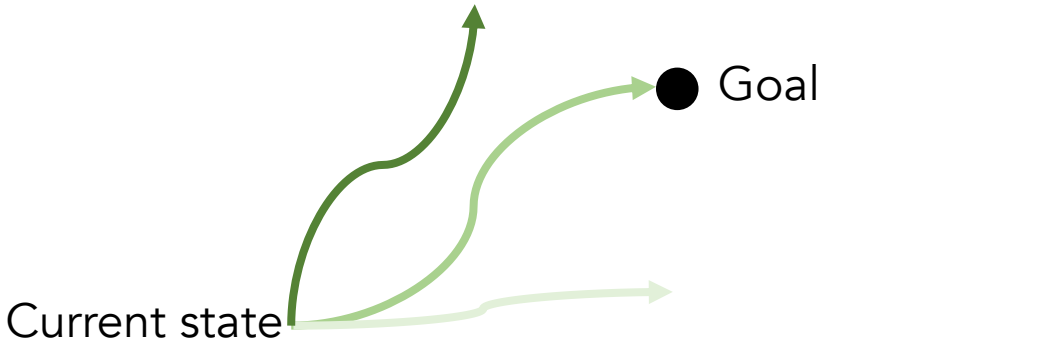


Formulating human decision making

- Game theory
- Behavioral economics
- Computational cognitive science

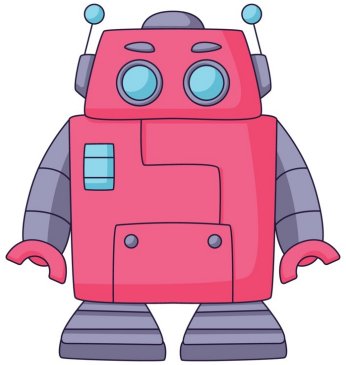
Level-0 agent models for embodied tasks



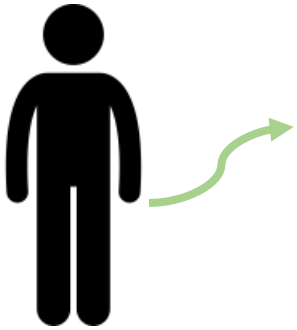


Simulate plans via world model

Choose the best plan

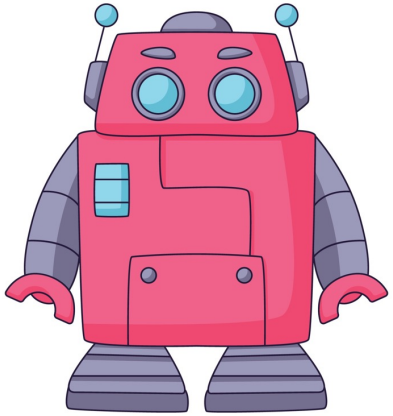


Level-1 agent models for social reasoning

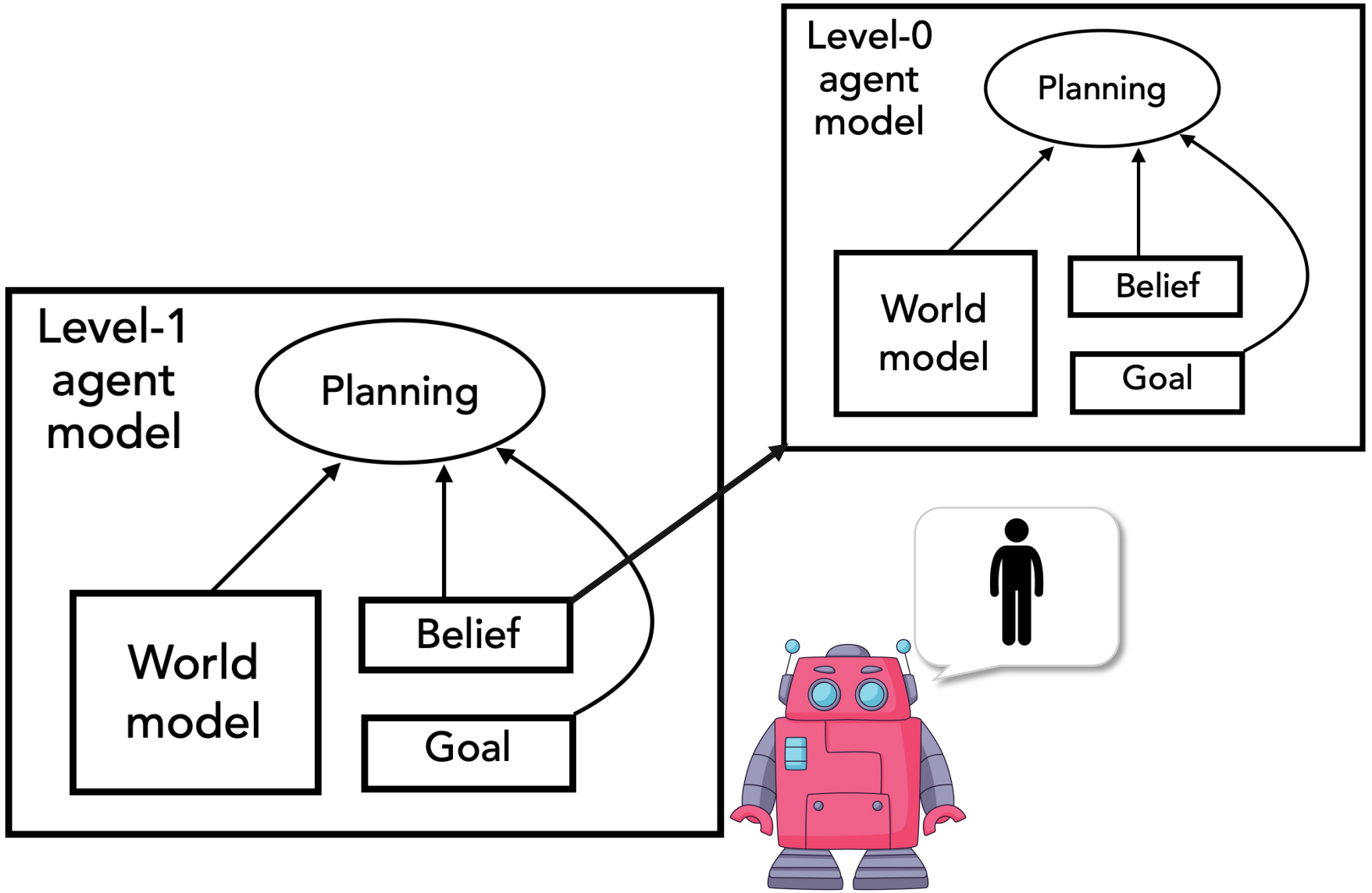


Goal: Office or coffee shop?

An observer



Level-1 agent models for social reasoning



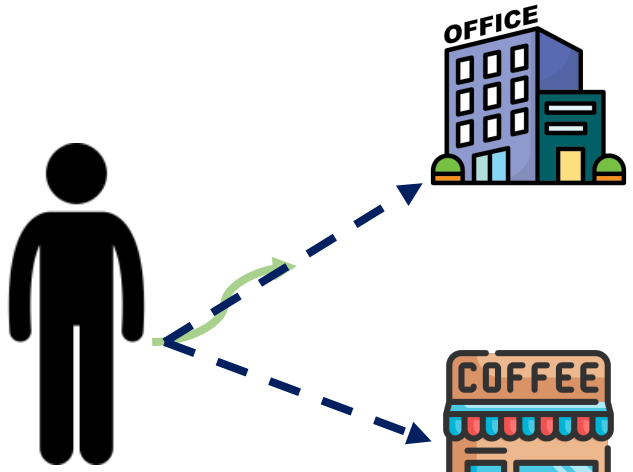
Level-1 agent models for social reasoning

Model-based Theory of Mind

$$P(\text{mind}|\text{state}, \text{actions}) \propto P(\text{actions}|\text{state}, \text{mind})P(\text{mind})$$

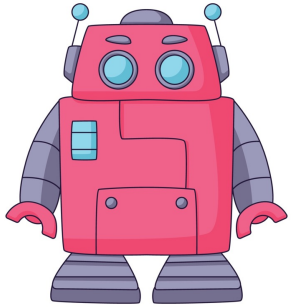
Bayesian inference

Level-0 agent model



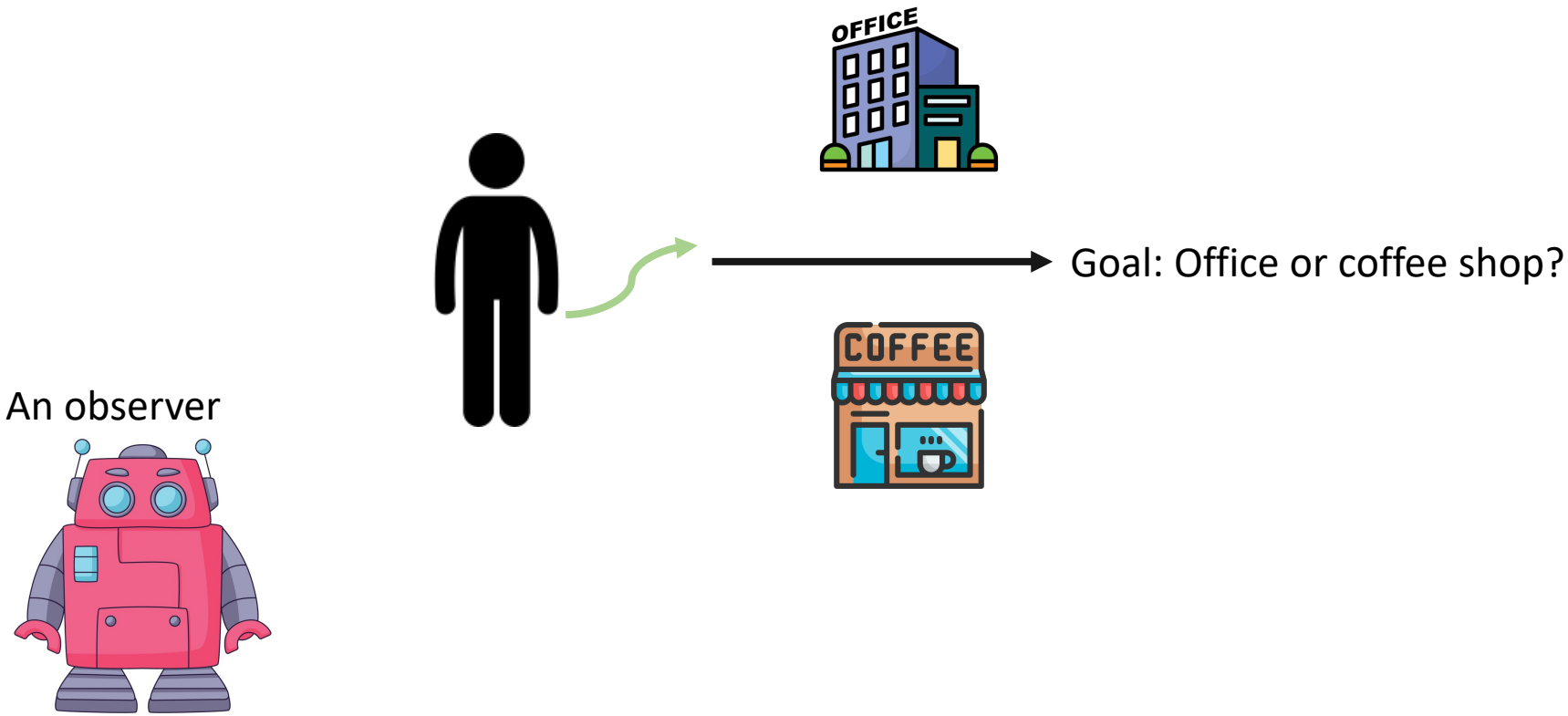
Goal: Office or coffee shop?

An observer

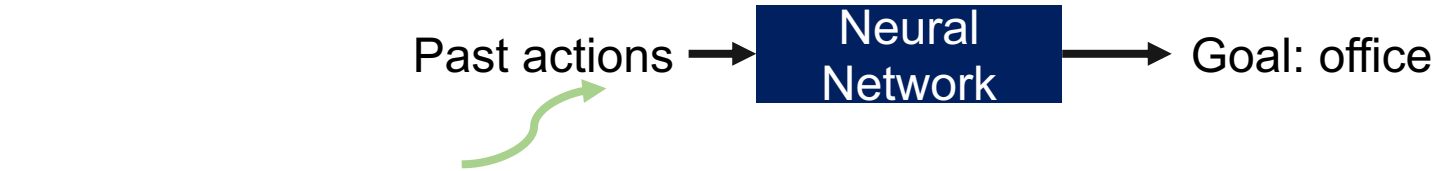


Social reasoning without agent models

Model-free Theory of Mind: directly maps actions to mental state

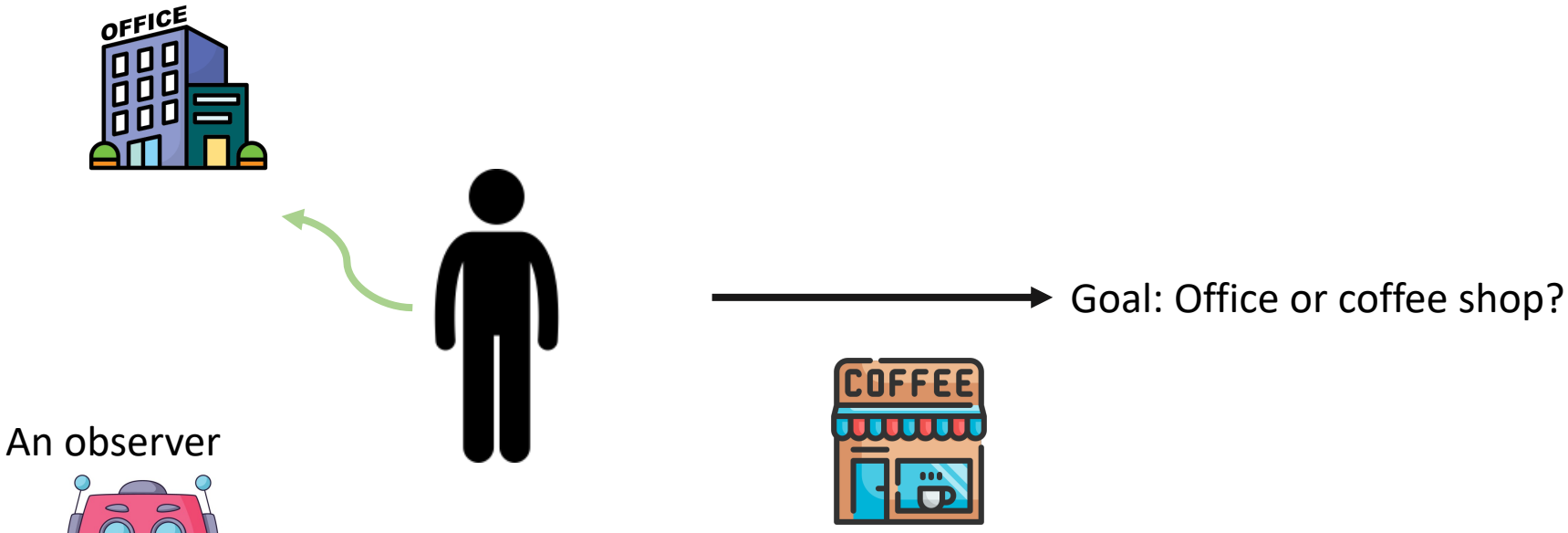


Social reasoning without agent models



Model-free Theory of Mind: directly maps actions to mental state

Difficult to generalize to new scenarios

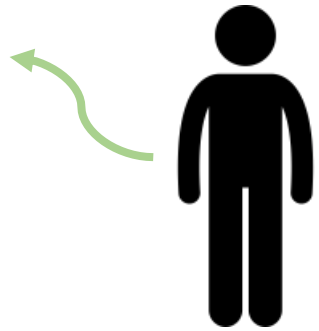


Netanyahu et al. (2021)
Shu et al. (2021)

Social reasoning without agent models

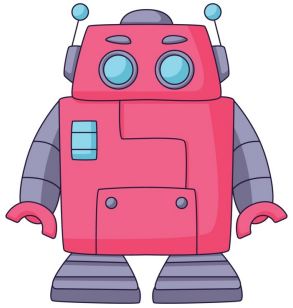
Model-based Theory of Mind would still work

$$P(\text{mind}|\text{state}, \text{actions}) \propto P(\text{actions}|\text{state}, \text{mind})P(\text{mind})$$



Goal: Office or coffee shop?

An observer



Netanyahu et al. (2021)
Shu et al. (2021)

Level-1 agent models for social reasoning

Model-based Theory of Mind

$$P(\text{mind}|\text{state}, \text{actions}) \propto P(\text{actions}|\text{state}, \text{mind})P(\text{mind})$$

Level-0 agent model

Human Behavior Prediction

$$P(\text{future actions}|\text{state}, \text{mind})$$

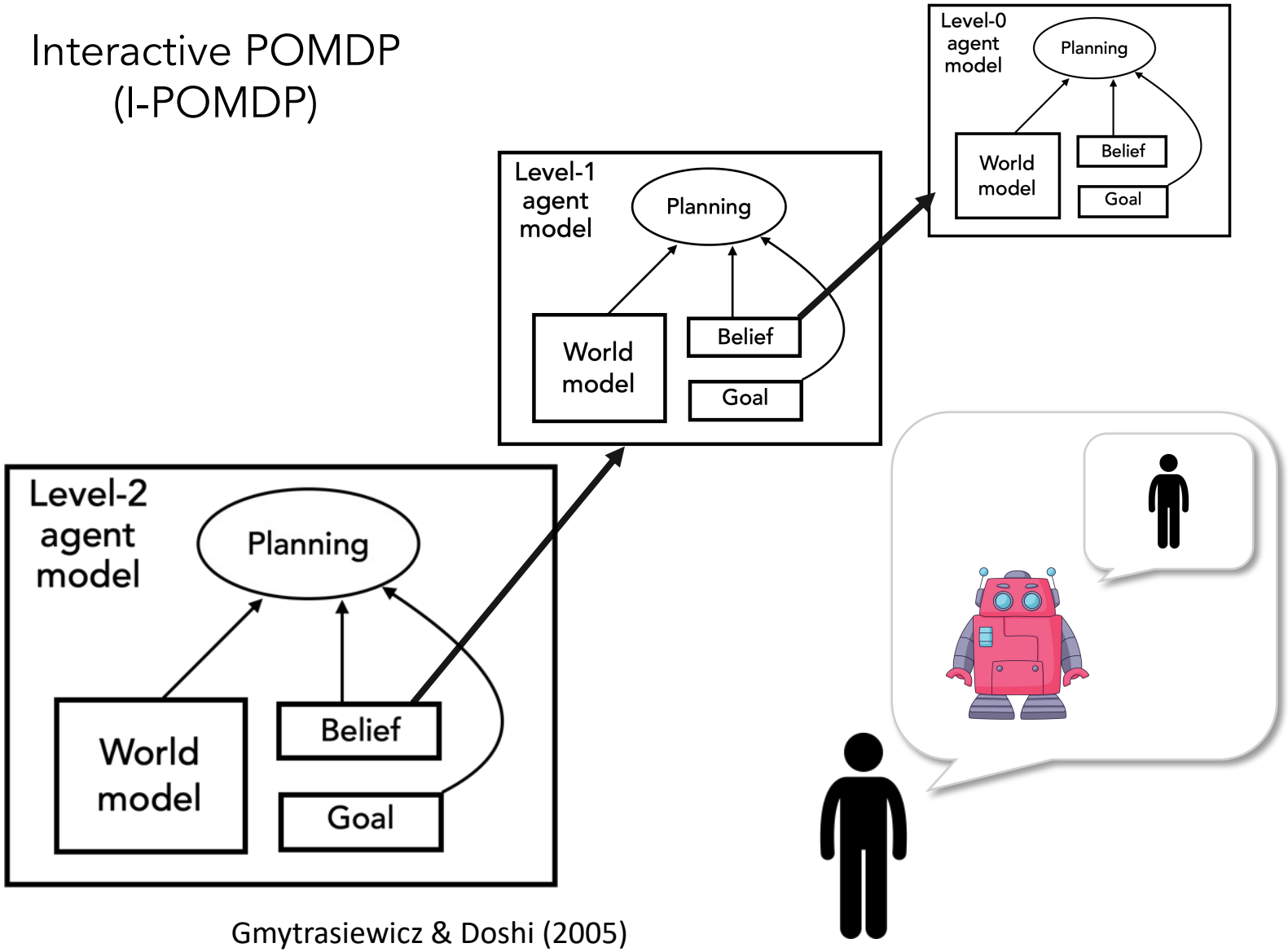
OR

Human-AI Interaction

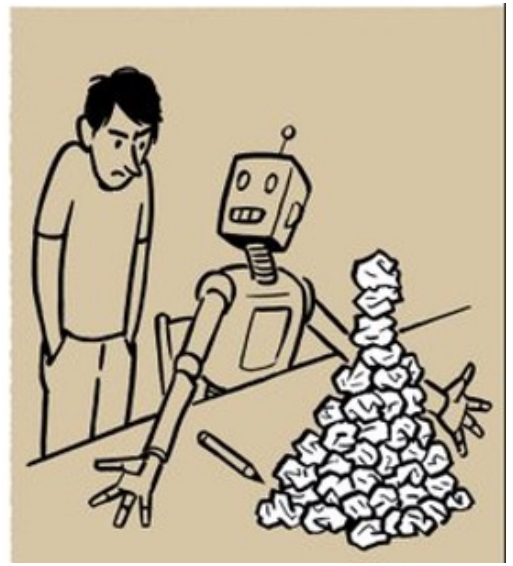
$$\pi(\text{action}_{\text{AI}}|\text{state}, \text{mind}_{\text{AI}}, \text{mind}_{\text{human}})$$

Higher-order agent models for *recursive* social reasoning

Interactive POMDP
(I-POMDP)



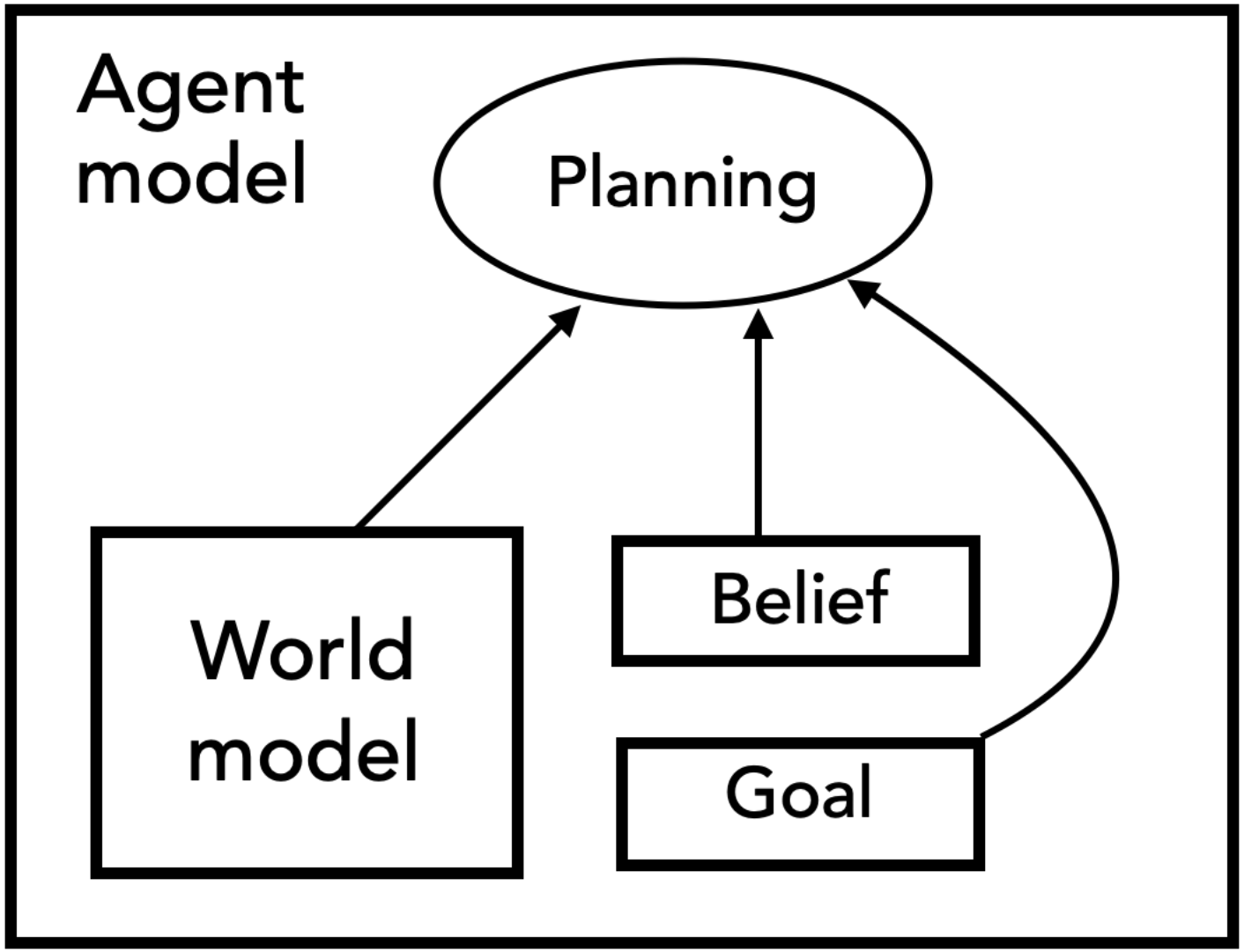
Gmytrasiewicz & Doshi (2005)



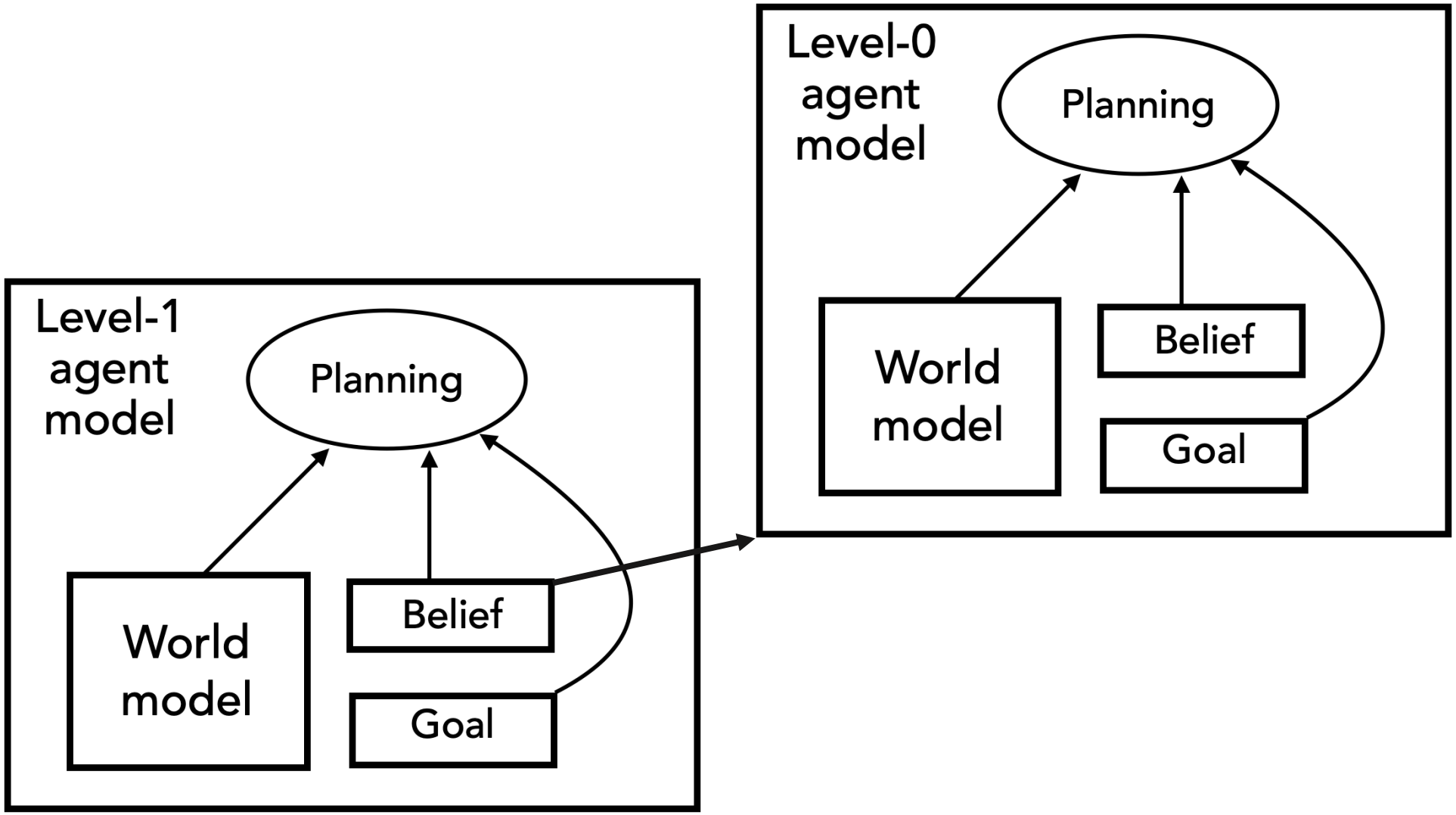
GPT-4V:

… The final panel reveals the punchline: the robot has merely produced a pile of crumpled paper, just like the human did, suggesting that the robot also suffers from writer's block … highlighting a situation where the human and the AI are equally challenged

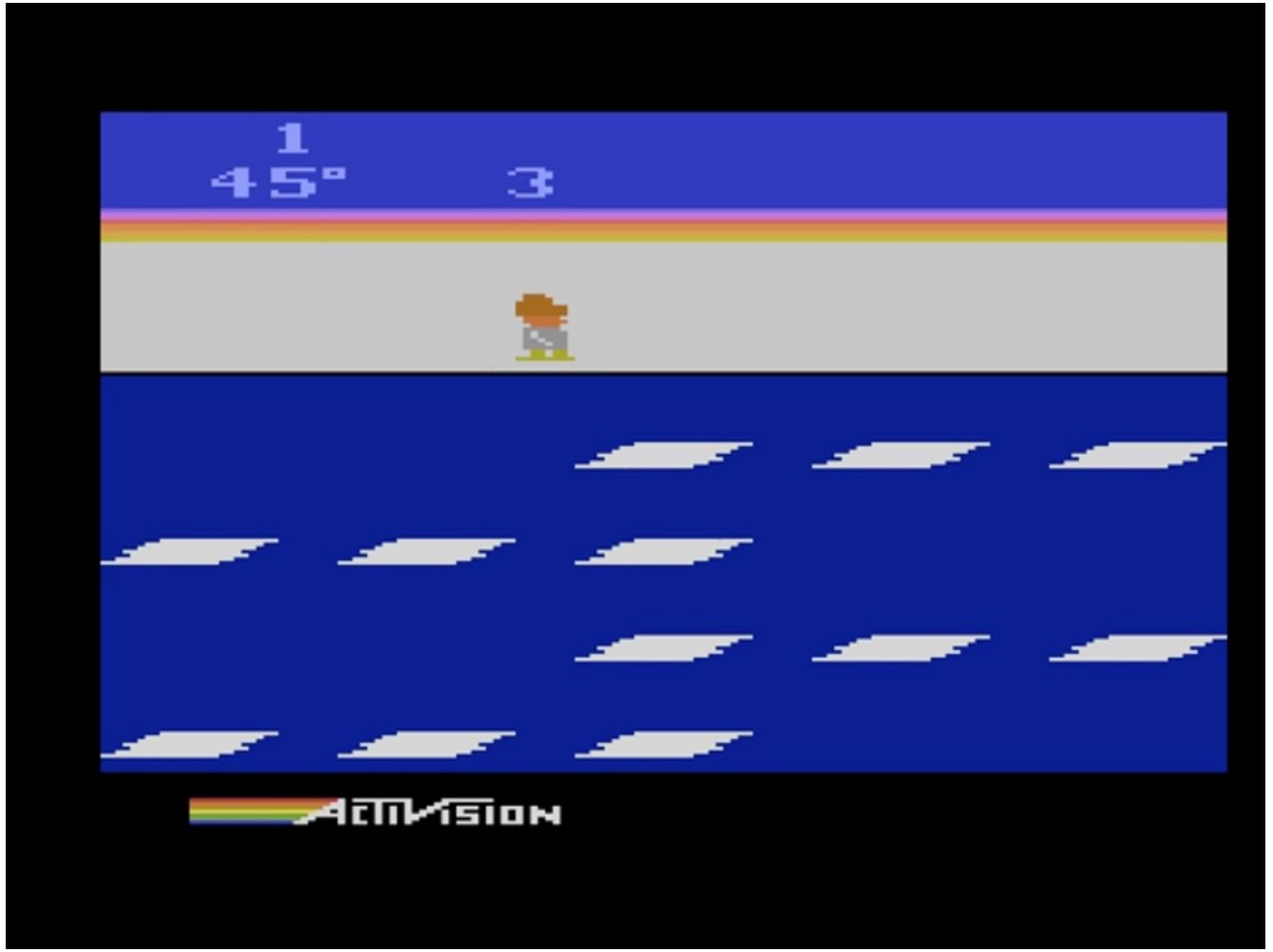
Summary



Summary


Level-0
agent
model
Planning
World
model
Belief
Goal
Level-1
agent
model
Planning
World
model
Belief
Goal

Why do humans (or animals) learn world and agent models from little data?



Why do humans (or animals) learn world and agent models from little data?

- Learning about novel concepts and cause-effect relations between events, from one or few experiences
  - Learning about rules behind frostbite
- Perceiving objects, agents, and events from all kinds of modalities
  - Seeing a character as an agent, ice, igloo, birds, lighting up ice, building the igloo
- Exploring, being curious, experimenting
  - Jumping on the ice to see what will happen, checking what's inside the igloo
- Forming goals, sub-goals, and plans
  - Building the igloo to go the next level
- Learning from others
  - Watching a video of gameplay or reading a game manual

How do we reverse engineer a cognitive system?

- From the “problem of induction” to the problem of inductive inference
- Three levels of analysis
- A toolkit for solving these problems

Modeling the world: the problem of induction

- How do learn so much from so little (sample efficient)? How do we go beyond the data given (strong generalization)?

Data → abstract knowledge → generalization in unseen scenarios

A common (model-free) machine learning paradigm
Data → patterns

Learning words for objects

“segway”

Learning words for objects

“cam”

Learning words for objects

From wikipedia

Learning words for objects

Learning words for objects

"tufa"

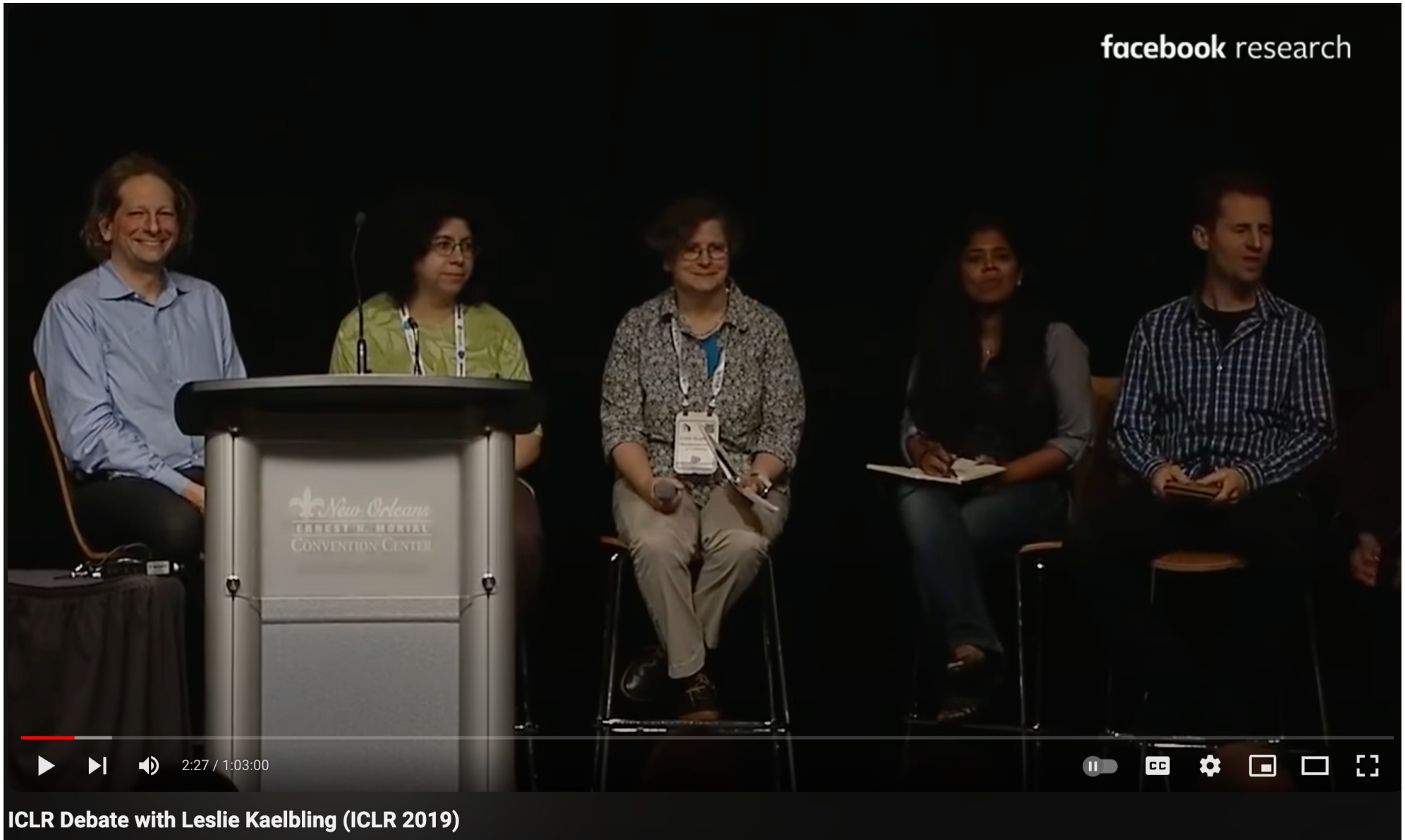
Modeling the world: the problem of induction

- Abstract knowledge.
- Constraints / inductive biases / priors

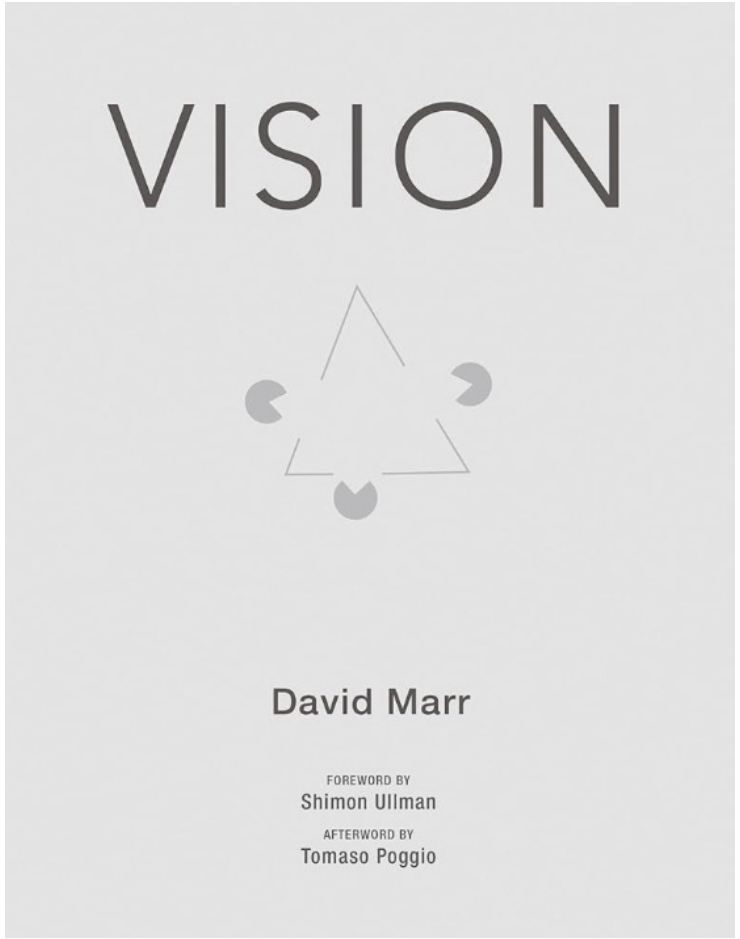
- How does abstract knowledge guide learning and inference from sparse data?
- What form does abstract knowledge take, across different domains, and tasks?
- How is abstract knowledge itself constructed, from some combination of innate specifications and experience?

ICLR 2019 Debate: structure, learning, and the role of priors

(At the peak of the interests in RL)



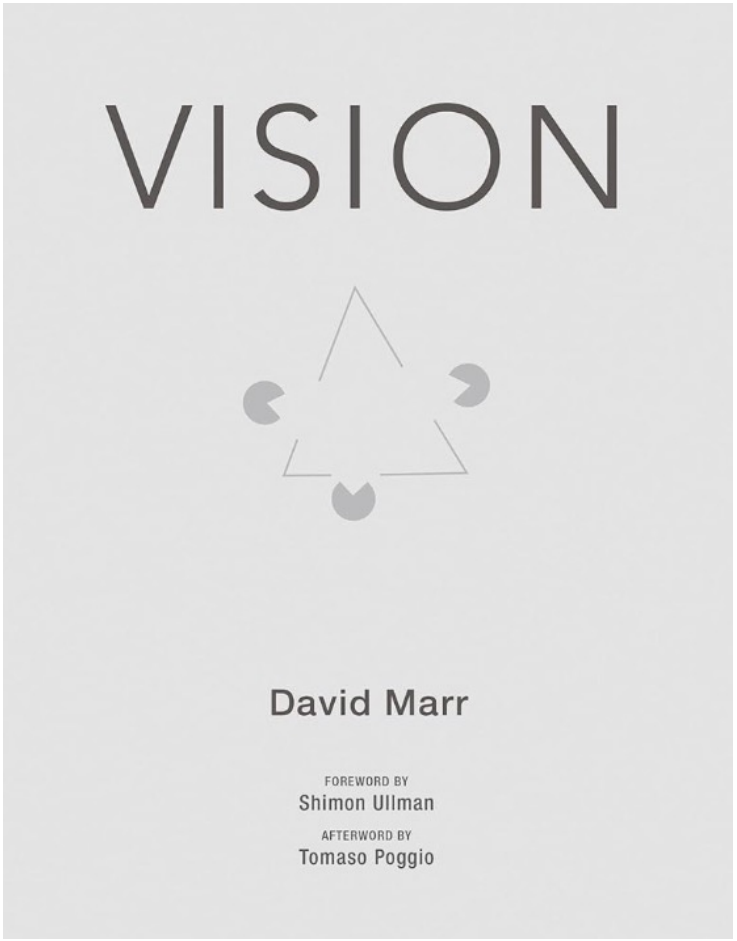
How to systematically think about a cognitive system?



"When I started vision in 1982 there was a dream –articulated by David Marr in his book "Vision" – that Computer Vision (CV) and Biological Vision (BV) could be studied together in a complimentary manner."

-- Alan Yuille

Three levels of analysis for reverse engineering a cognitive system



- **Level 1: Computational theory**
  - What are the inputs and outputs to the computation, what is its goal, and what is the logic by which it is carried out?
- **Level 2: Representation and algorithm**
  - How is information represented and processed to achieve the computational goal?
- **Level 3: Hardware implementation**
  - How is the computation realized in physical or biological hardware?

A toolkit for solving the problem of induction

- 1. How does abstract knowledge guide learning and inference given sparse data?

  Bayesian inference in probabilistic generative models.

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$$

Basics of Bayesian inference

- Bayes' Rule

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$$

- An example
  - Data *d*: John is coughing
  - Some hypotheses *h*:
    - 1. John has a cold
    - 2. John has lung cancer
    - 3. John has a stomach flu
  - What is your best guess?
  - Likelihood $P(d \mid h)$ favors 1 and 2 over 3
  - Prior probability $P(h)$ favors 1 and 3 over 2
  - Posterior probability $P(h|d)$ favors 1 over 2 and 3

Word learning as Bayesian inference

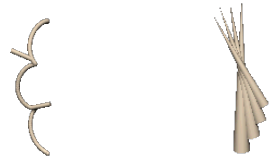
Word Learning as Bayesian Inference

Fei Xu
University of British Columbia

Joshua B. Tenenbaum
Massachusetts Institute of Technology

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$$

h: knowledge, i.e., word meaning / categorization
d: examples/images of an object category



What is the right prior?
What is the right hypothesis space?
How do learners acquire that background knowledge?

A toolkit for solving the problem of induction

- 1. How does abstract knowledge guide learning and inference sparse data?

  Bayesian inference in probabilistic generative models.

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$$

- 2. What form does that knowledge take, across different domains and tasks? Probabilities defined richly structured symbolic representations: spaces, graphs, grammars, logical predicates, schemas...
- 3. How is the knowledge itself constructed, from some combination of innate specifications and experience?

  Hierarchical models, with inference at multiple levels.

  In machine learning terms: learning models as probabilistic inference, “learning to learn”, transfer learning, learning representations and learning inductive biases

Word learning as Bayesian inference

Plants?
Fungi?

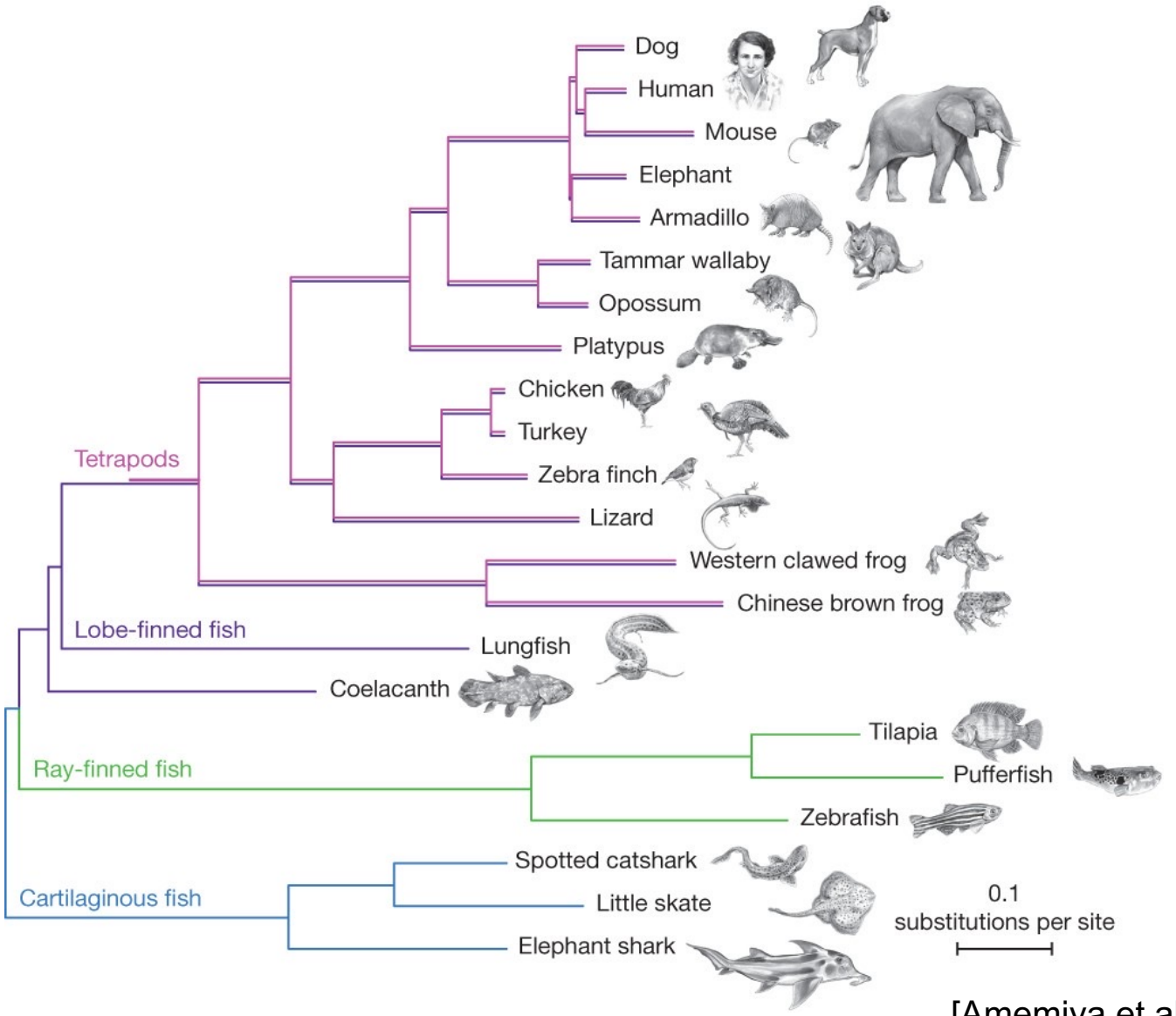
Bacteria?
Worms?

Shellfish?

Informed by our prior knowledge of biology

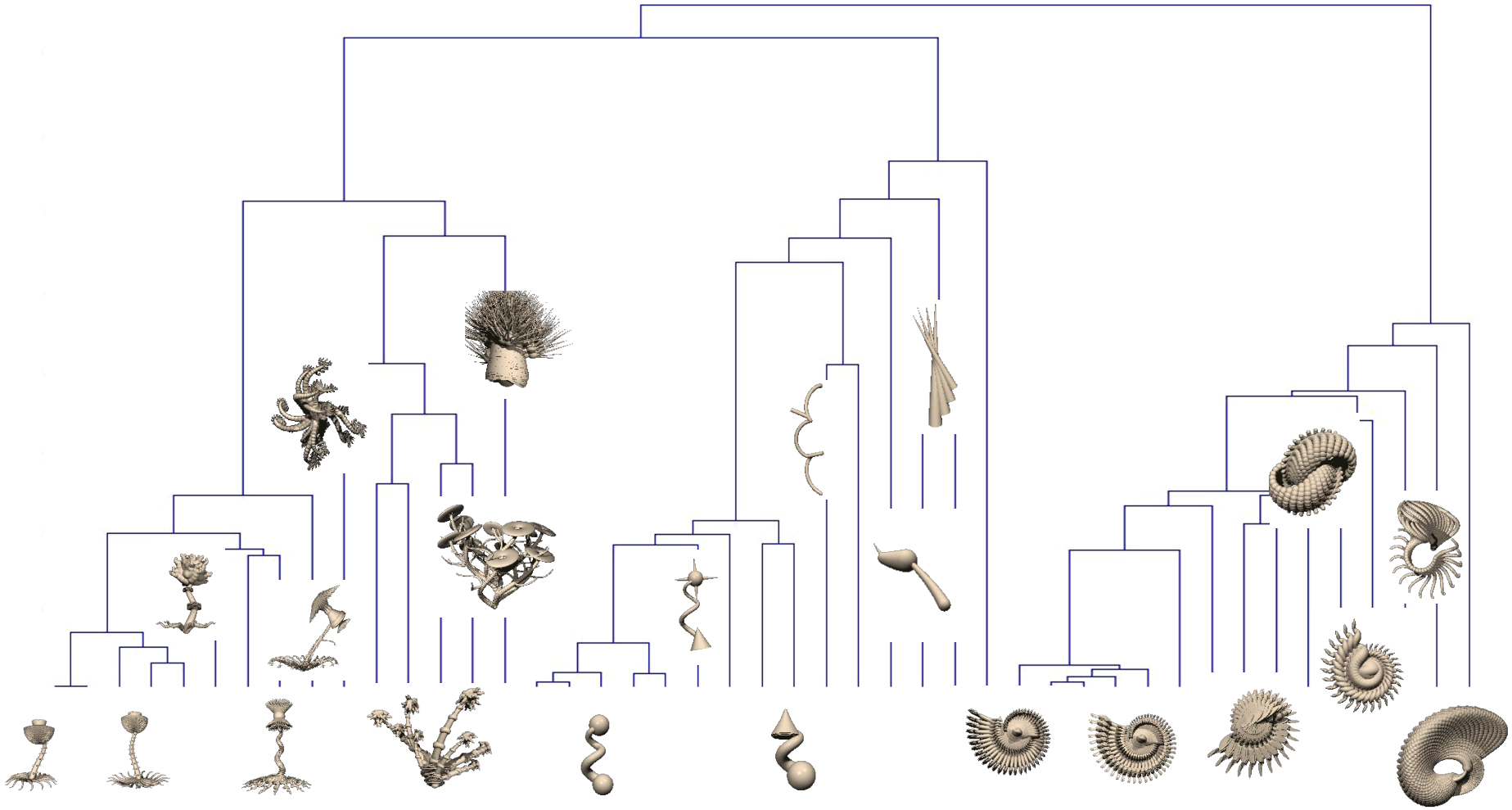
Word learning as Bayesian inference

From biology: **tree** structure to represent evolutionary or genetic lineage

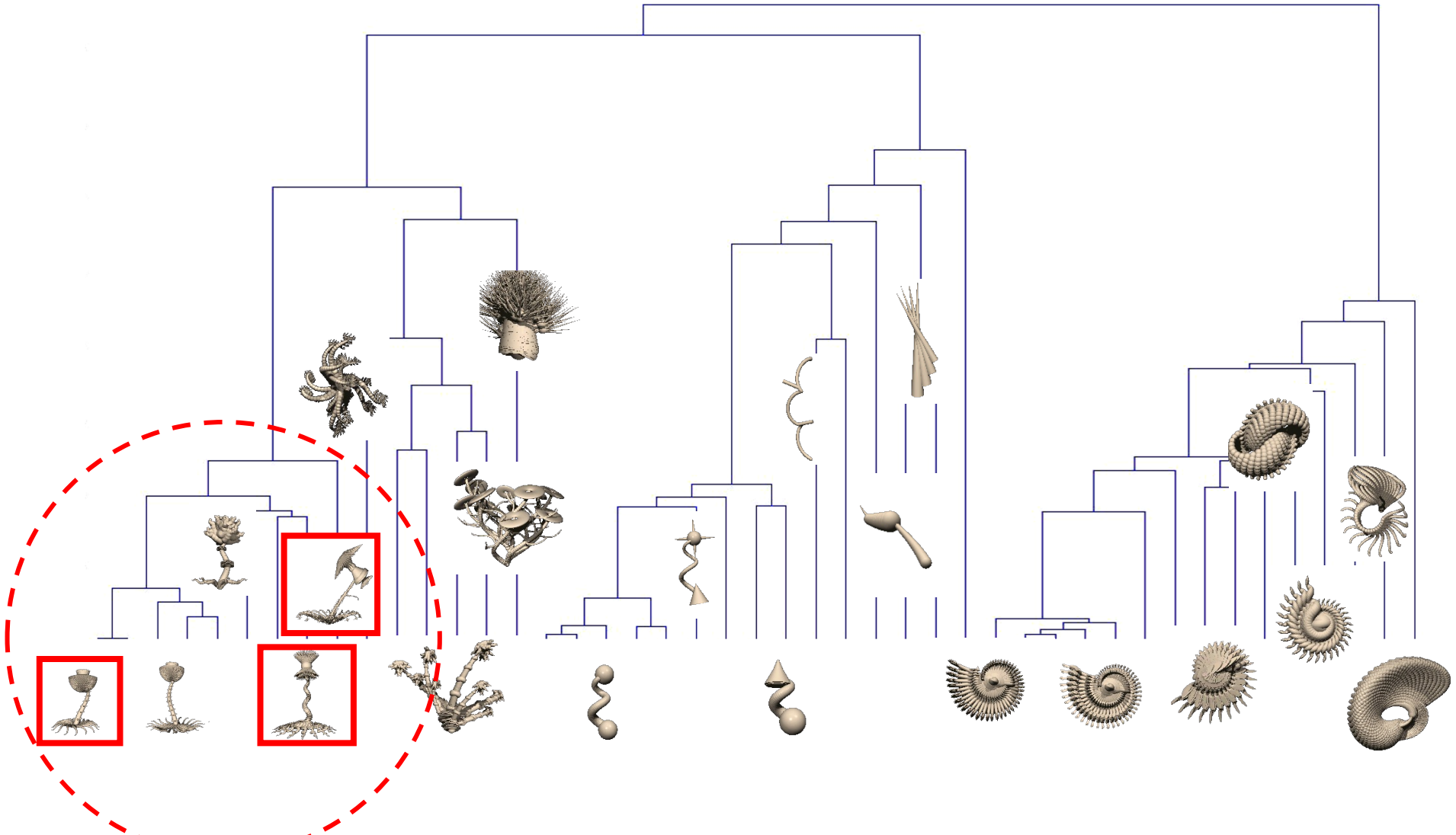


[Amemiya et al. 2013: 312, fig. 1]

Word learning as Bayesian inference



Word learning as Bayesian inference



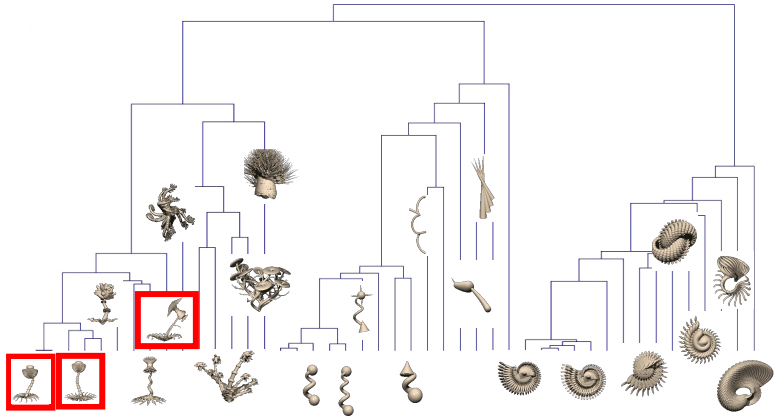
Origins of the hypothesis space

$P(\text{form})$

Hierarchically construct a hypothesis space

F: form

Tree

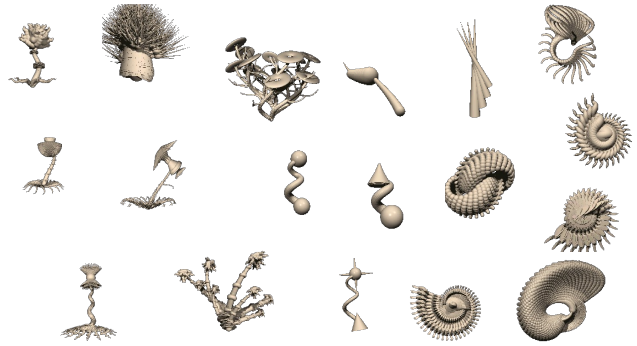


$P(\text{structure} \mid \text{form})$

S: structure

$P(\text{data} \mid \text{structure})$

D: data



A toolkit for solving the problem of induction

- 1. How does abstract knowledge guide learning and inference sparse data?

  Bayesian inference in probabilistic generative models.

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{\sum_{h_i \in H} P(d \mid h_i)P(h_i)}$$

- 2. What form does that knowledge take, across different domains and tasks? Probabilities defined richly structured symbolic representations: spaces, graphs, grammars, logical predicates, schemas...
- 3. How is that knowledge itself constructed, from some combination of innate specifications and experience?

  Hierarchical models, with inference at multiple levels.

  In machine learning terms: learning models as probabilistic inference, “learning to learn”, transfer learning, learning representations and learning inductive biases

A toolkit for solving the problem of induction

- 4. How can learning and inference proceed efficiently and accurately, even with very complex hypothesis spaces?

  Sampling-based algorithms for approximate inference, e.g., MCMC, sequential Monte Carlo (particle filtering), fast initialization with bottom-up recognition models (neural networks).

- 5. How can probabilistic inference be used to drive action?

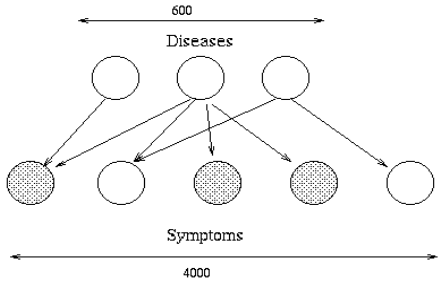
Utility-based frameworks for decision and planning under uncertainty and risk.

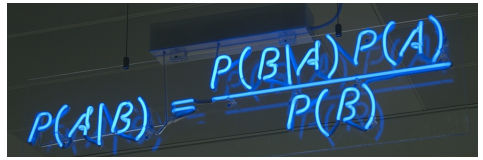
- 6. How could these computations be implemented in hardware?
  Brain, GPU

Probabilistic programming languages (PPLs)

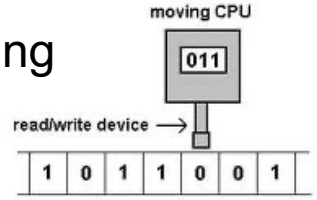
A unifying framework integrating our best computational ideas on intelligence, across multiple eras of cognitive science and AI:

- **Probabilistic (Bayesian) inference** for reasoning about likely unobserved causes from observed effects, and decision making under uncertainty.

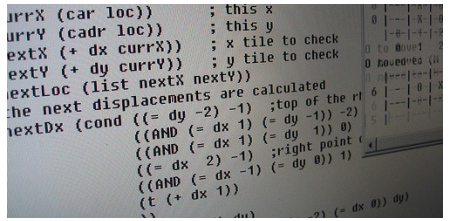




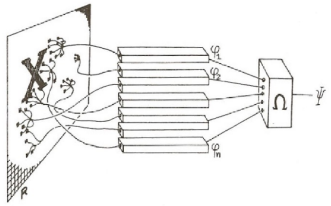
- **Symbolic programs** for representing and reasoning with abstract knowledge

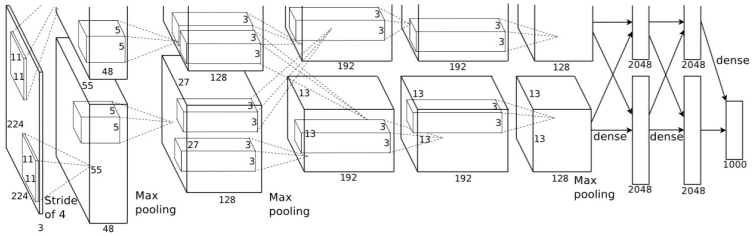


$$x \in V \Rightarrow x \in \Lambda,$$
$$M, N \in \Lambda \Rightarrow (MN) \in \Lambda,$$
$$\in \Lambda, x \in V \Rightarrow (\lambda x M) \in \Lambda.$$



- **Neural networks** for pattern recognition and function approximation.





Look ahead

- Bayesian cognition and Bayesian modeling
- Bayesian concept learning
  - As an example of Bayesian modeling
  - Foundation for world modeling and agent modeling
- Bayesian networks
  - Complex Bayesian modeling
  - Efficient inference algorithms
- Probabilistic programming languages
- Physical reasoning
  - Intuitive physics in humans and machines
  - Model-based physical scene understanding
- Social reasoning
  - Intuitive psychology in humans and machines
  - Single/multi-agent decision making
  - Inverse decision making
  - Moral judgment (a guest lecture)