DSC250: Advanced Data Mining

Text (Multi-modal) Mining:
Large Language (Multi-modal) Models

Zhiting Hu

Lecture 3, Jan 14, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

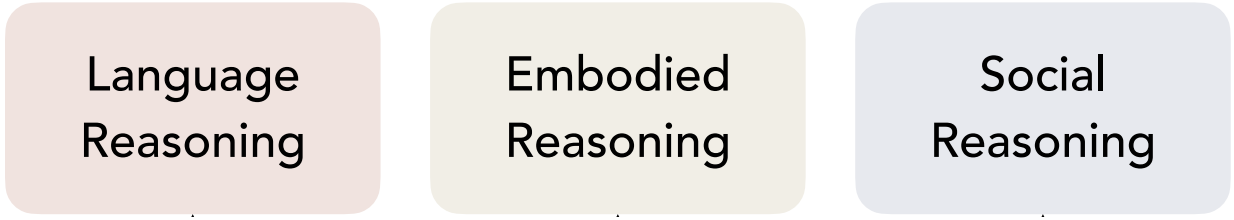
Recap: Existing Multi-Modal Models and Limitations

Summary of existing works

- **Multi-modal LMs (I)**
  - Can understand images
  - Can not generate images for describing a world state
- **Multi-modal LMs (II)**
  - Can do interleaved generation of image and text
  - Not describing the same world consistently
- **Video Simulation Models**
  - Generate videos given actions
  - Not (yet) generalist models: domain-specific states/actions
  - Reasoning only in pixel space
- Text-to-video Models
  - Generate videos given text prompts
  - Limited length of reasoning (60s)
  - Limited control with actions
  - Reasoning only in pixel space

Recap: Limitations

- LLMs/LMMs have limited language, embodied, and social reasoning abilities; not human-level yet



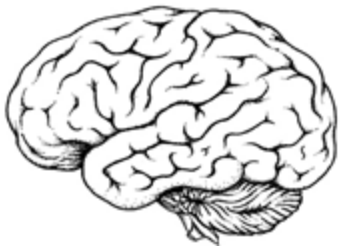
- Humans conduct model-based reasoning based on models of the **world** and **agents**

Large language models

- Next word prediction
  - Do LMs “truly understand”, or is it just string manipulation?
  - Do LMs have “internal world models”?

> “Large language models have no idea of the underlying reality that language describes.”
>
> **(Yann Lecun)**

> “On the surface, it may look like we are just learning statistical correlations in text. But, it turns out that to just learn it to compress them really well, what the neural network learns is some representation of the process that produced the text. This text is actually a projection. This text is actually a projection of the world; there is a world out there and it’s as a projection on this text.”
>
> **(Ilya Sutskever)**



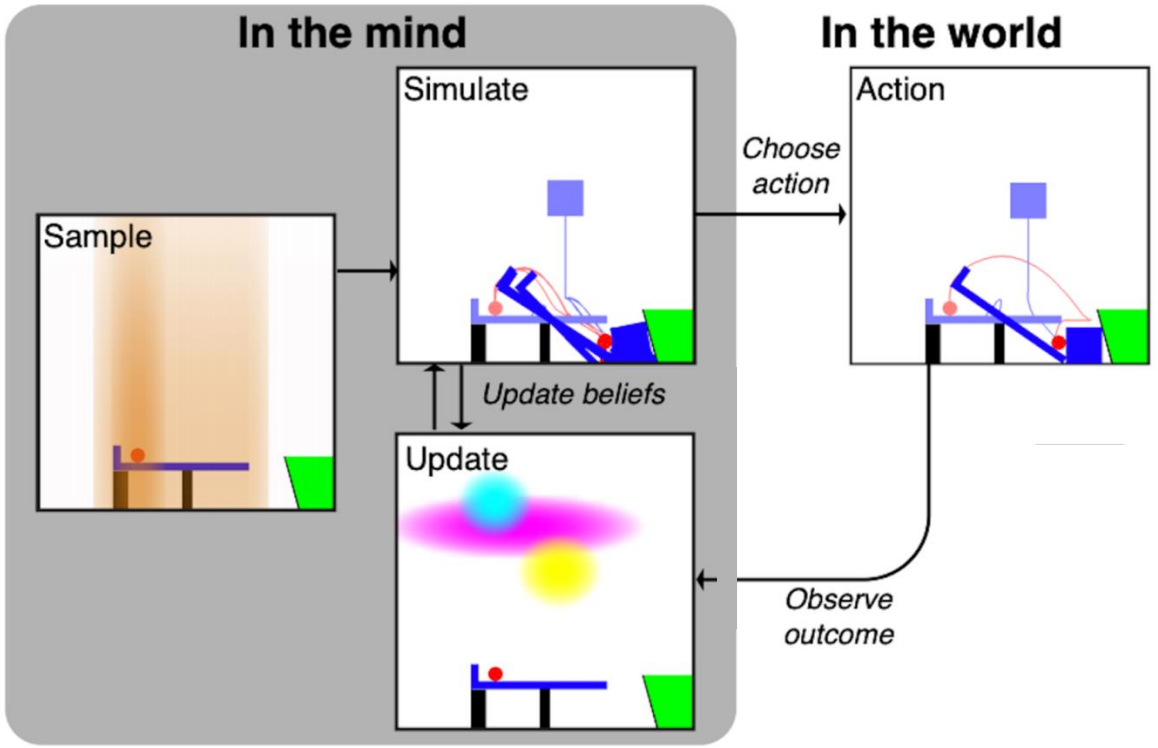
Recap: World models in humans

- Understanding the world
- Predicting the world
- Model-based control/planning

Human tool use

Humans can learn to use tools through just a few trials

Key is to use a world model to simulate the outcomes of possible plans



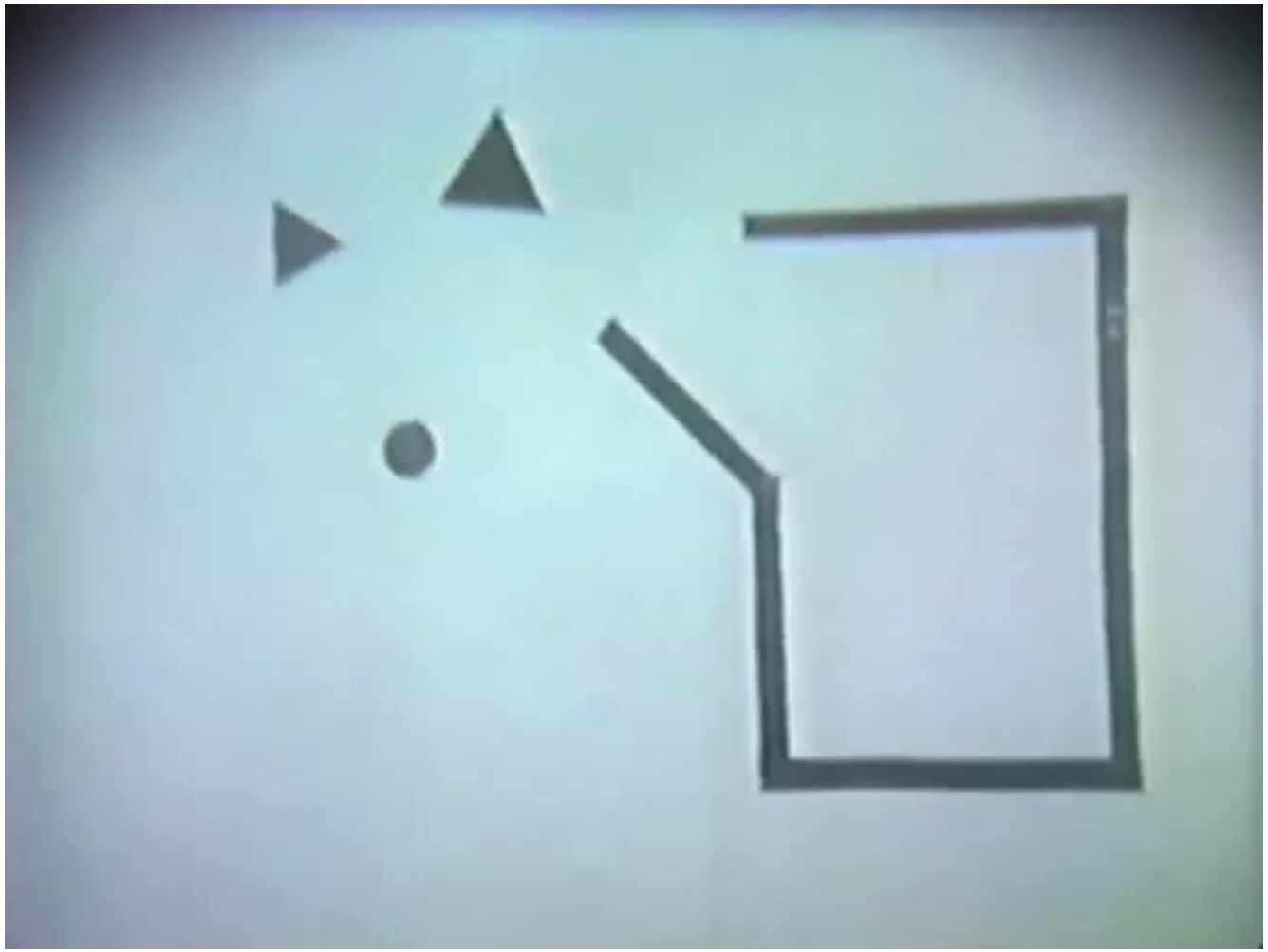
Recap: Simulative reasoning based on world models

- Next "world" prediction $P(s'|s,a)$
- The paradigm of "simulative reasoning"

- Simulate plans with world model
- Choose the best plan

Agent models

- An agent is more than just an object and actions



Heider & Simmel (1944)

Agent models

- An agent is more than just an object and actions

Strengths

strong, weak

Goals

helping, hurting, escaping

Relationships

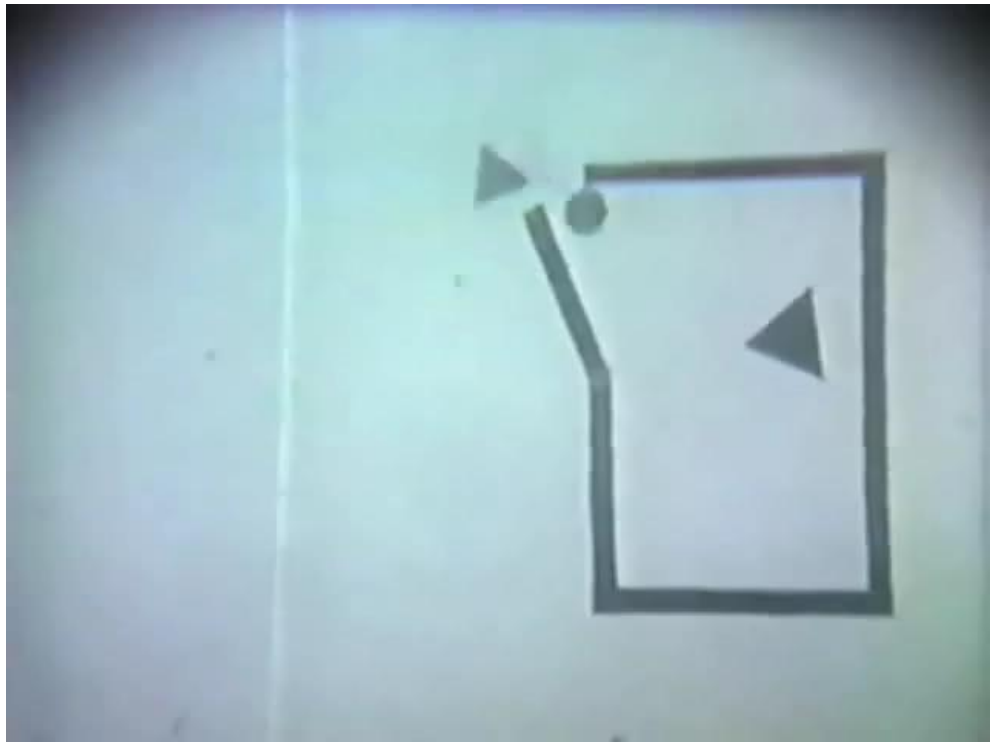
friends, enemies

Moral judgment

good guy, bully

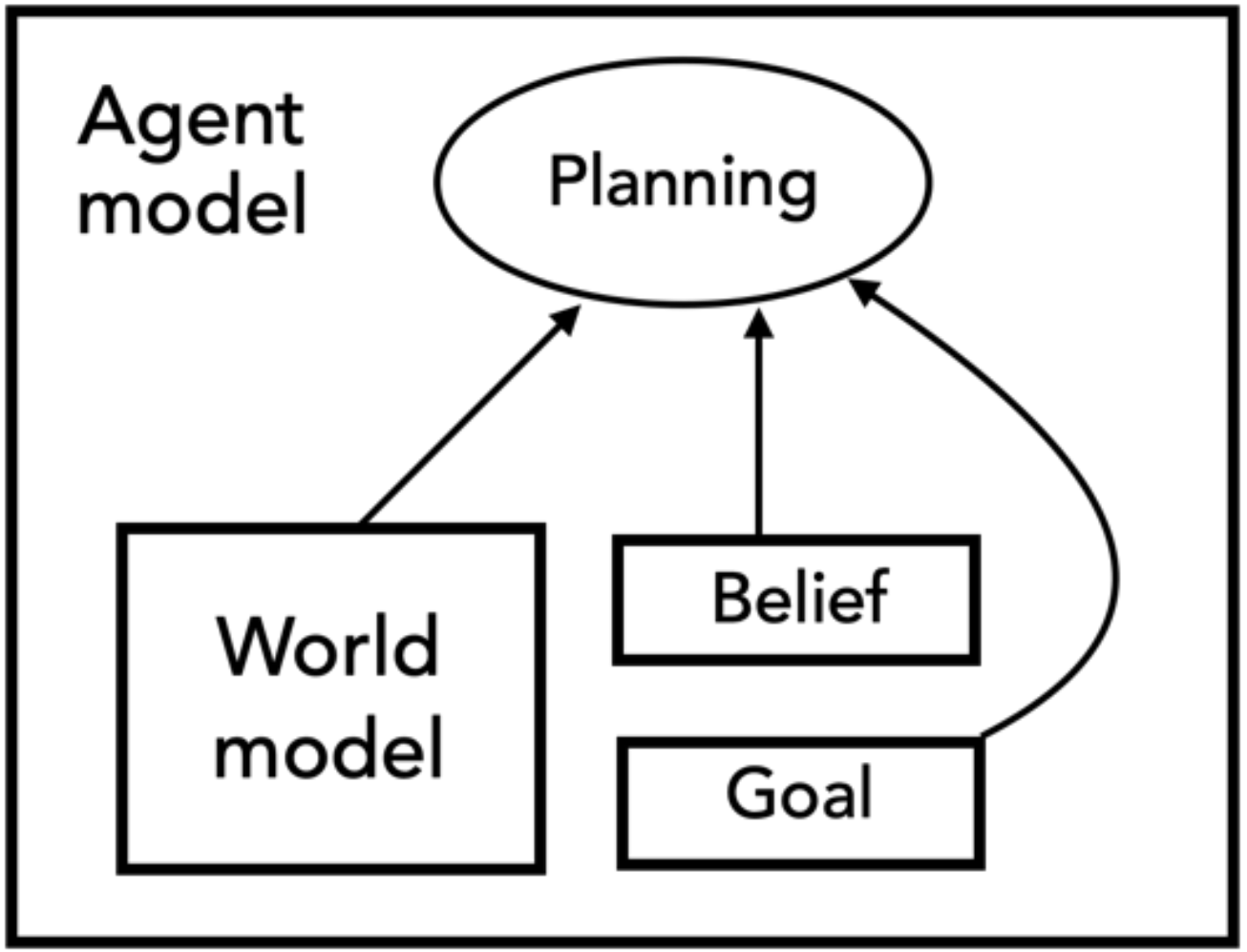
Beliefs

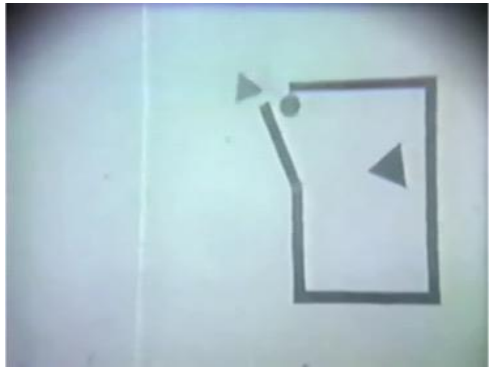
he is locked, i am safe



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

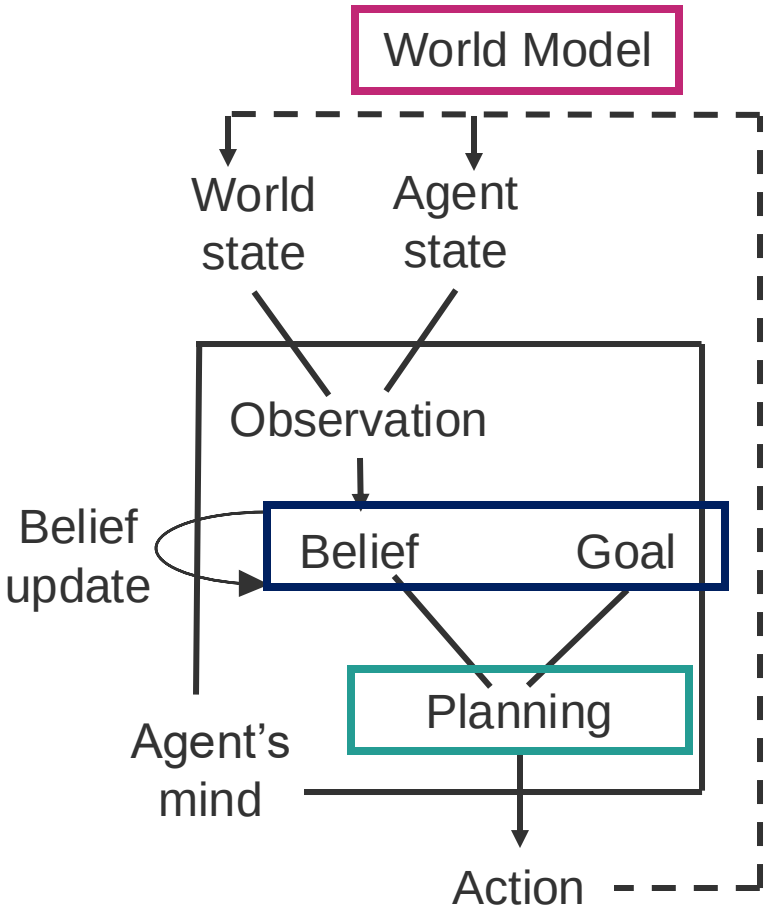
The minimum definition of an agent model





Formulation (not focus of this course)

Partially observable Markov decision process (POMDP)



State $s \in \mathcal{S}$

Action $a \in \mathcal{A}$

State transition probabilities $P(s'|s,a)$

Observation probabilities $O(o|s)$

Belief $b(s)$

Belief update $b'(s') \propto O(o'|s')P(s'|s,a)b(s)$

Goal $g \in \mathcal{G}$

Reward function $R(s,a,g) = R(s,g) - C(a)$

Discounted factor $\gamma \in [0,1]$

Planning $\max_{a^0,a^1,\ldots} E\left[\sum_{t=0}^{\infty} \gamma^t R(s^t,a^t,g)\right]$

Baker et al. (2017)

World Models

Recap: Simulative reasoning based on world models

- Next "world" prediction $P(s'|s,a)$
- Prior research built domain-specific world models
  - Primarily in robotics and embodied AI

(iv) Video prediction models

GAIA-1

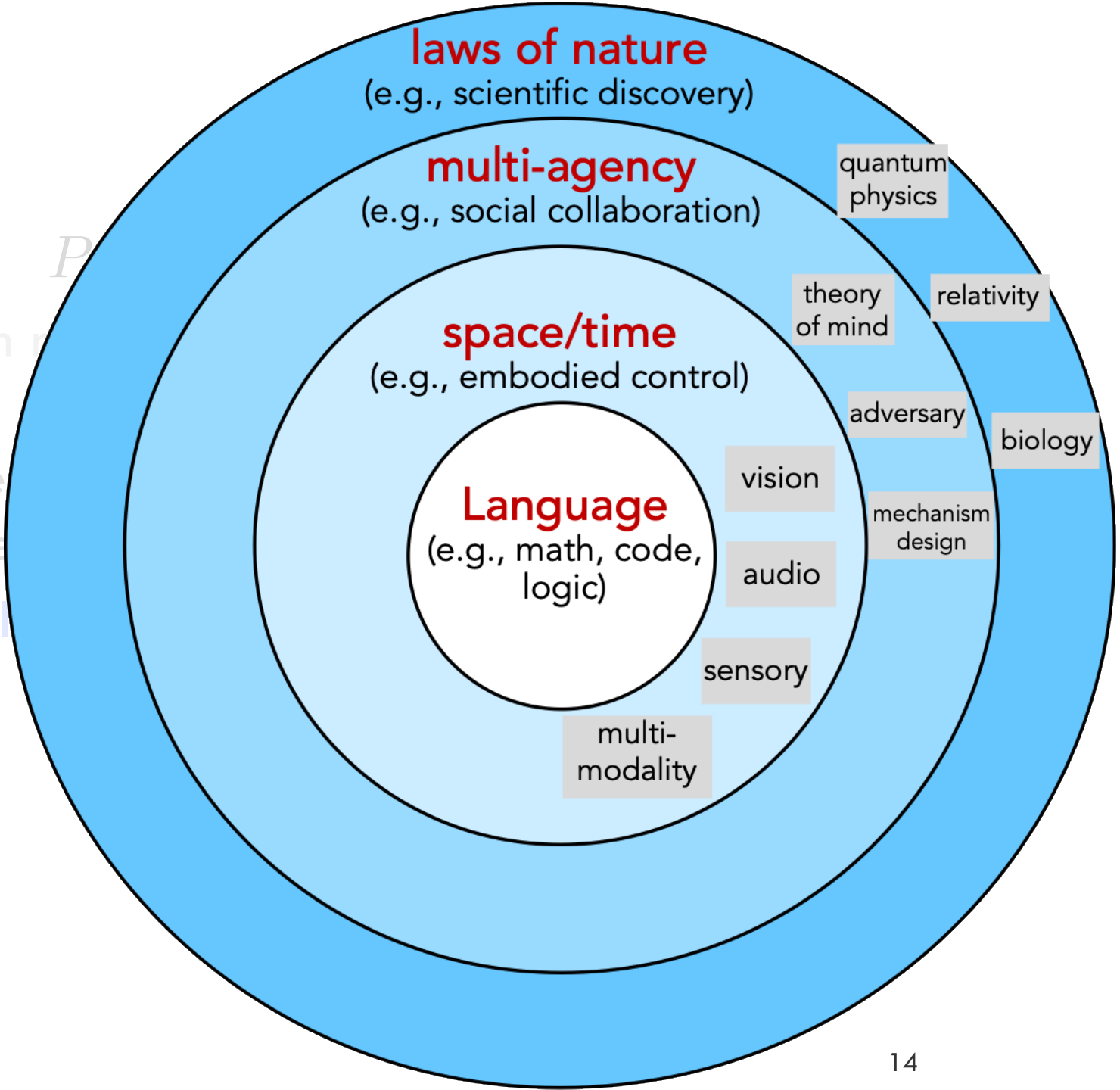
Simulative reasoning based on world models

- Next "world" prediction $P(s'|s,a)$
- Prior research built domain-specific world models
  - Primarily in robotics and embodied AI
- The scope of simulation defines the capability of reasoning
  - "More simulation, more intelligence"
- Can we build general world models?

Simulative reasoning

- Next "world" prediction
- Prior research (primarily in r
  specific world models

The scope of simulation defines the capability of reasoning

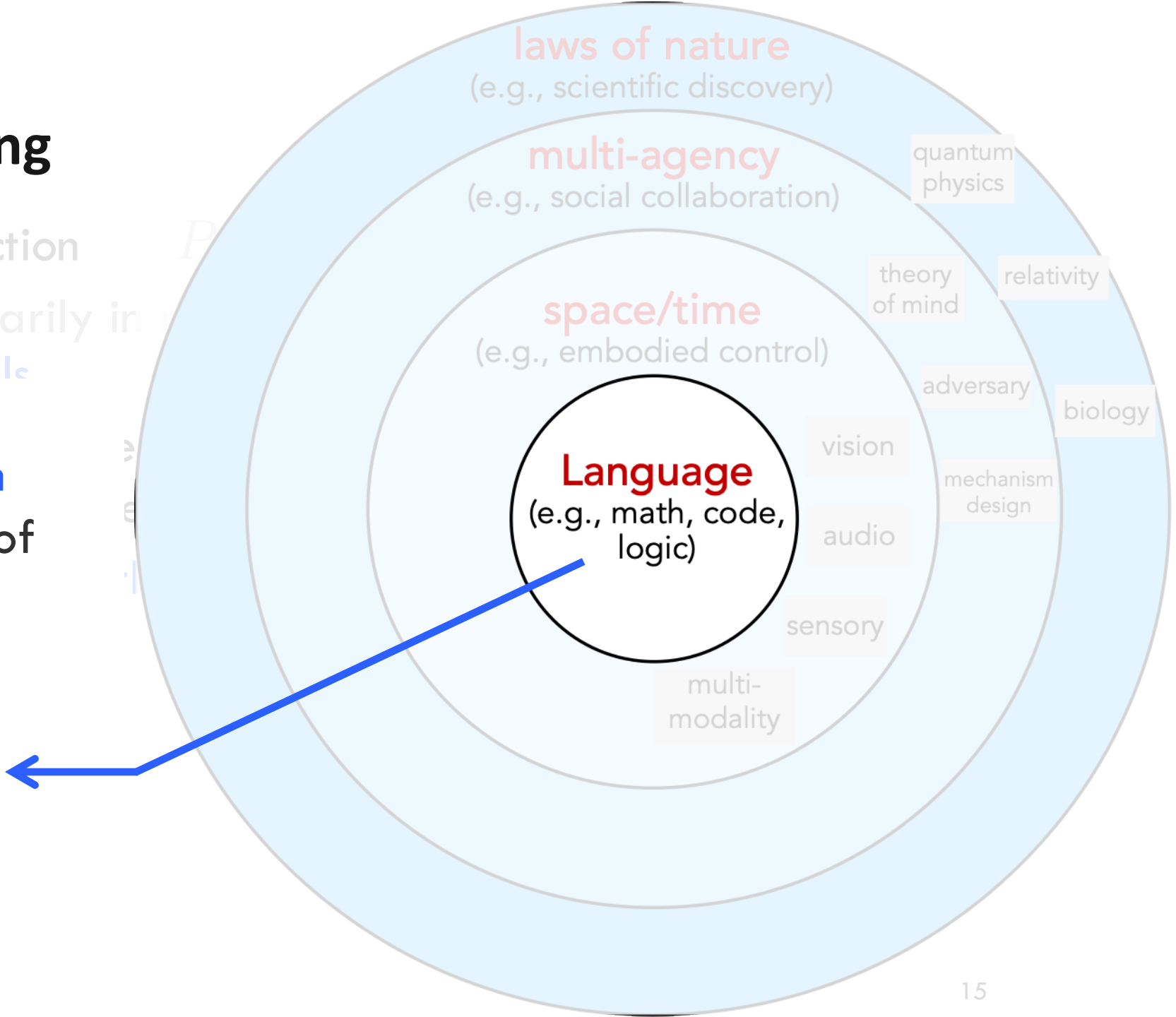


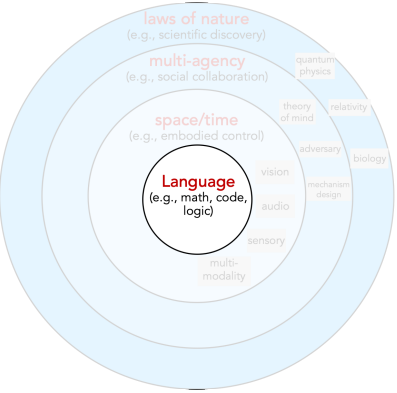
Simulative reasoning

- Next "world" prediction
- Prior research (primarily in specific world models

The scope of simulation defines the capability of reasoning

Language models as world models

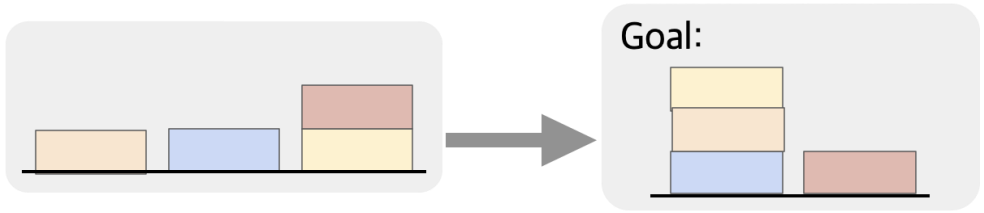


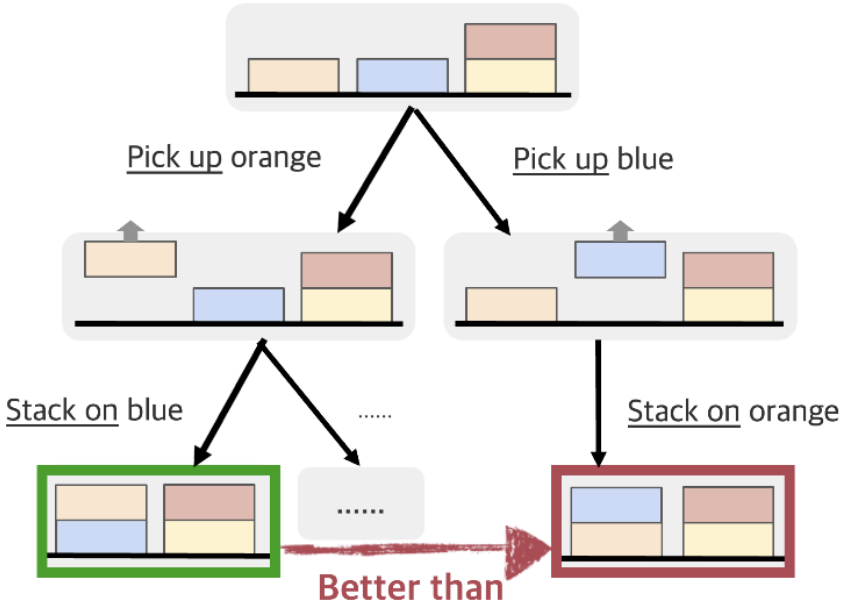


Simulative reasoning

Language models as world models

How to move the blocks to the goal state?





Human: model-based planning

- Internal world model to predict states
- Simulation of alternative plans

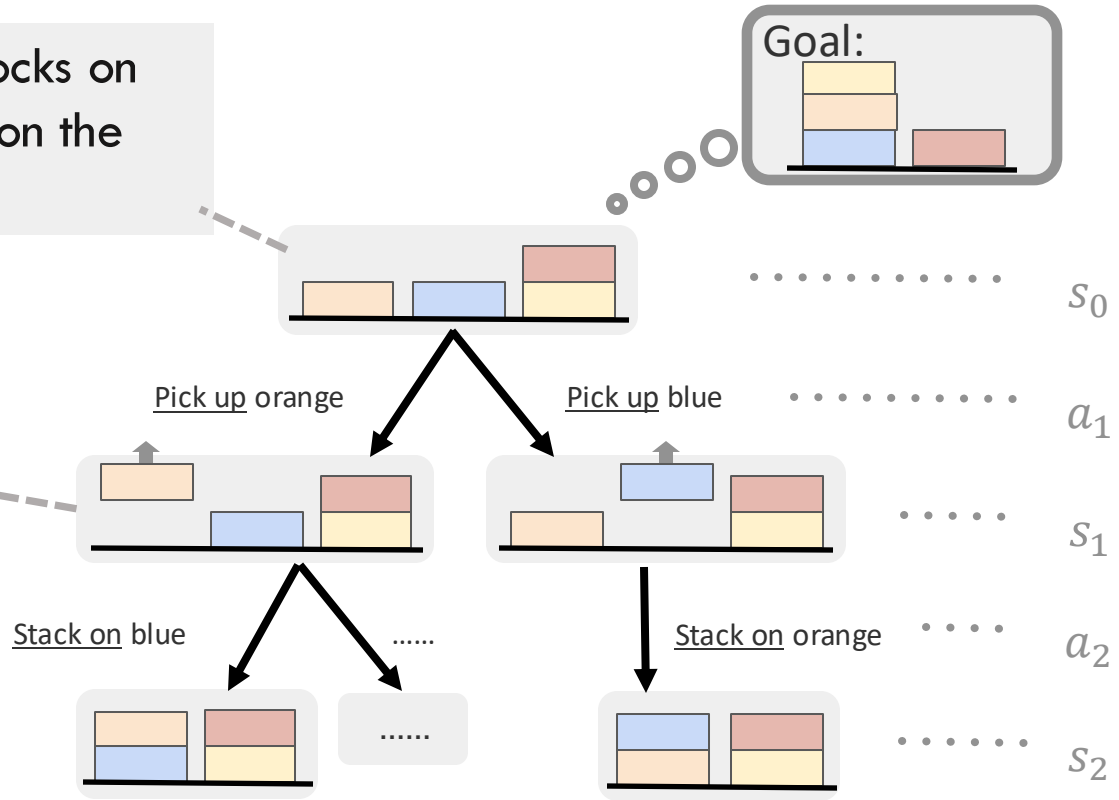
[Hao et al., 2023. Reasoning with language model is planning with world model]

Simulative reasoning

Orange and blue blocks on the table; red block on the yellow block

Language Model

Orange in the hand; blue block on the table; red block on the yellow block



Language Model as World Model

- Describe states with text.
- LM generates the description of next state

$$P(s'|s,a)$$

[Hao et al., 2023. Reasoning with language model is planning with world model]

AlphaGo-like reasoning

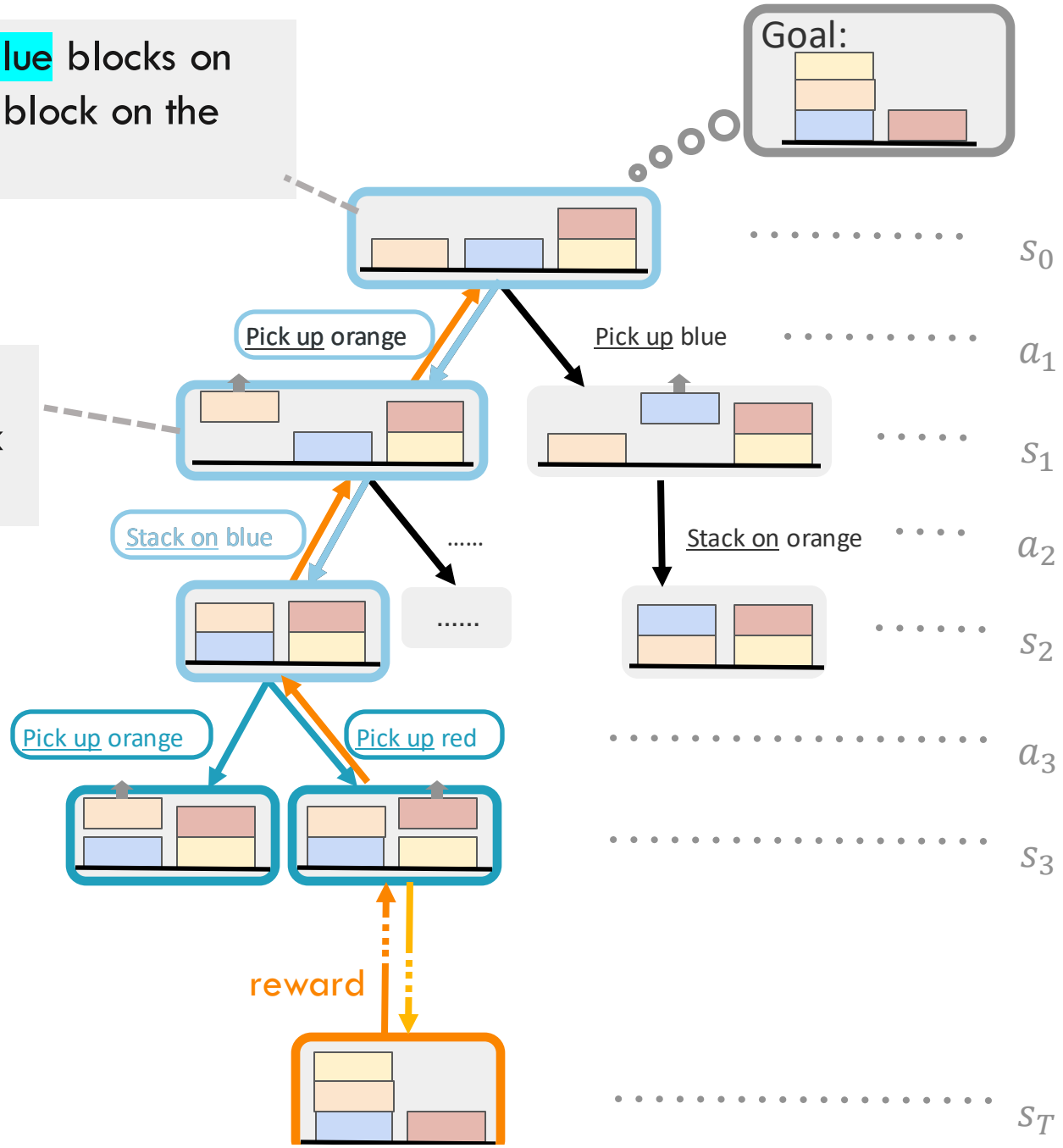
1. Selection
2. Expansion
3. Simulation
4. Back-propagation

reasoning

Orange and blue blocks on the table; red block on the yellow block

Language Model

Orange in the hand; blue block on the table; red block on the yellow block



Language Model as World Model

- Describe states with text.
- LM generates the description of next state

$$P(s'|s,a)$$

[Hao et al., 2023. Reasoning with language model is planning with world model]

AlphaGo-like reasoning

1. Selection
2. Expansion
3. Simulation
4. Back-propagation

Major improvement over conventional LLM word-by-word plan generation

Method	Success Rate
CoT	0.05
ToT (BFS)	0.09
ToT (DFS)	0.08
RAP	**0.51**

Language Model as World Model

- Describe states with text.
- LM generates the description of next state

$$P(s'|s,a)$$

[Hao et al., 2023. Reasoning with language model is planning with world model]

Simulative reasoning ba

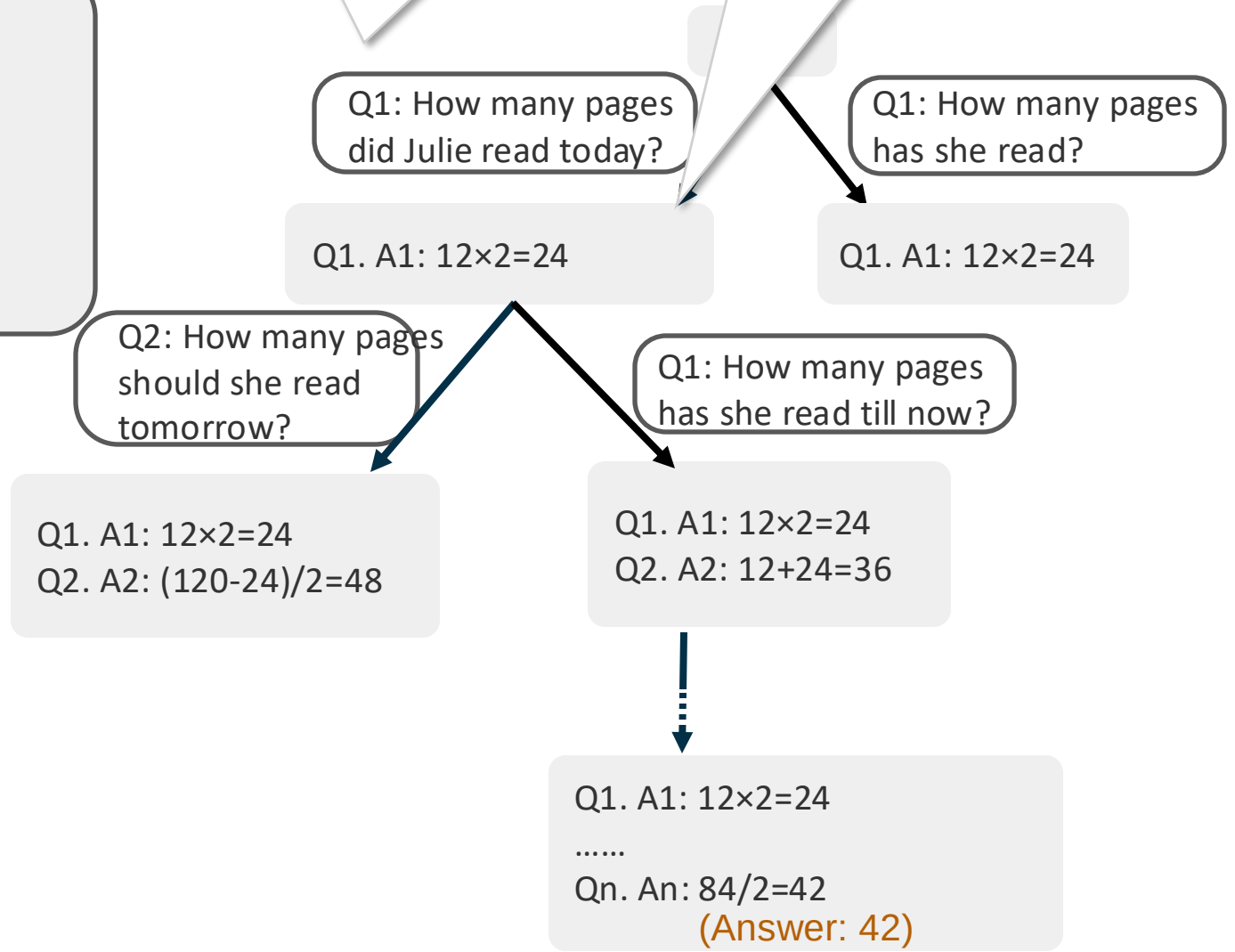
Question (Math):
Julie is reading a 120-page book.
Yesterday she read 12 pages
Today she read twice as many pages as yesterday
If she wants to read half of the remaining pages tmr,
how many pages should she read?

LM as World Model:
Prompt LM to generate the next state given the current state and action

$$P(s'|s,a)$$

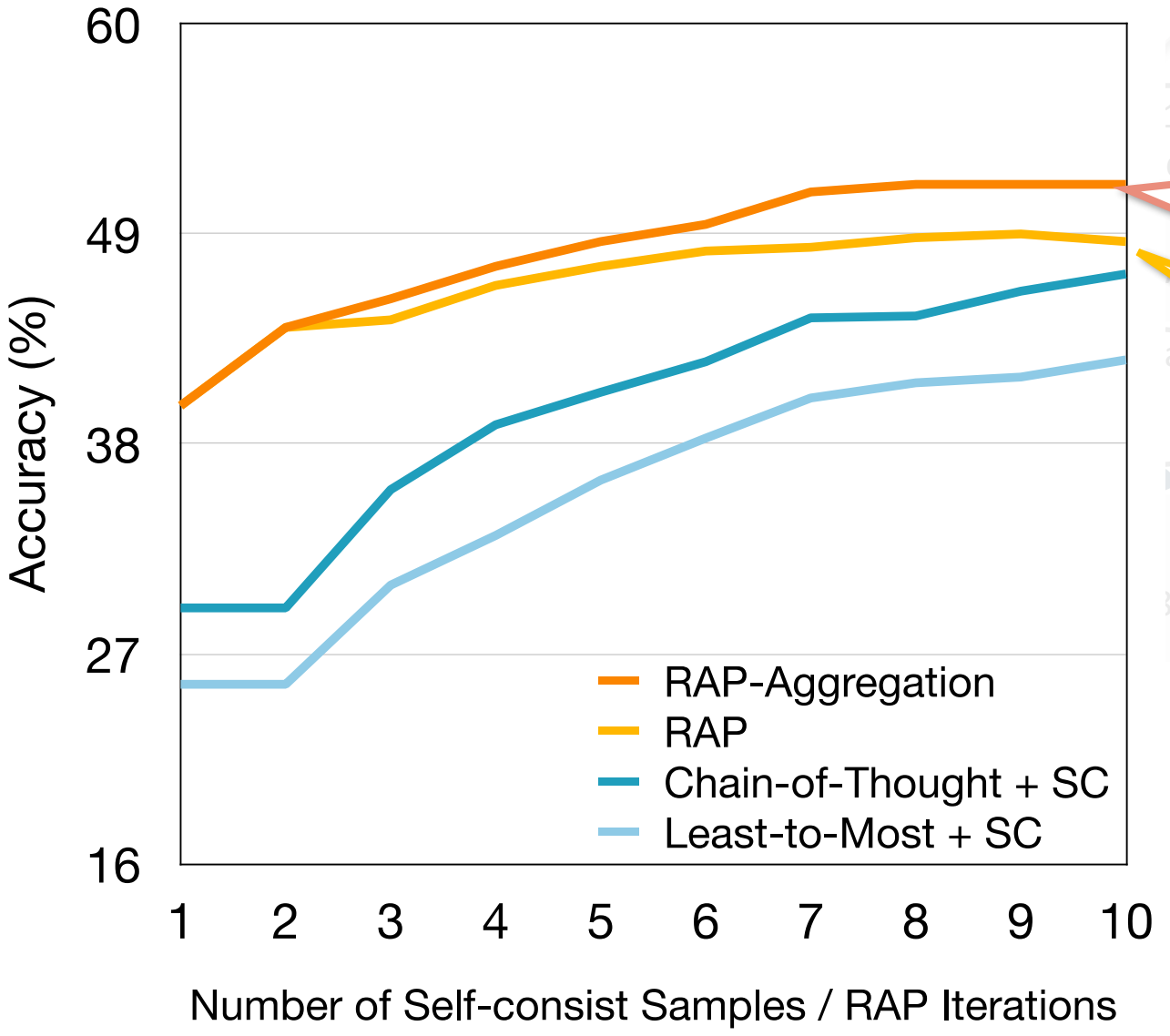
action: a sub-question for an unknown variable

state: intermediate values of variables



[Hao et al., 2023. Reasoning with language model is planning with world model]

Simulative reasoning



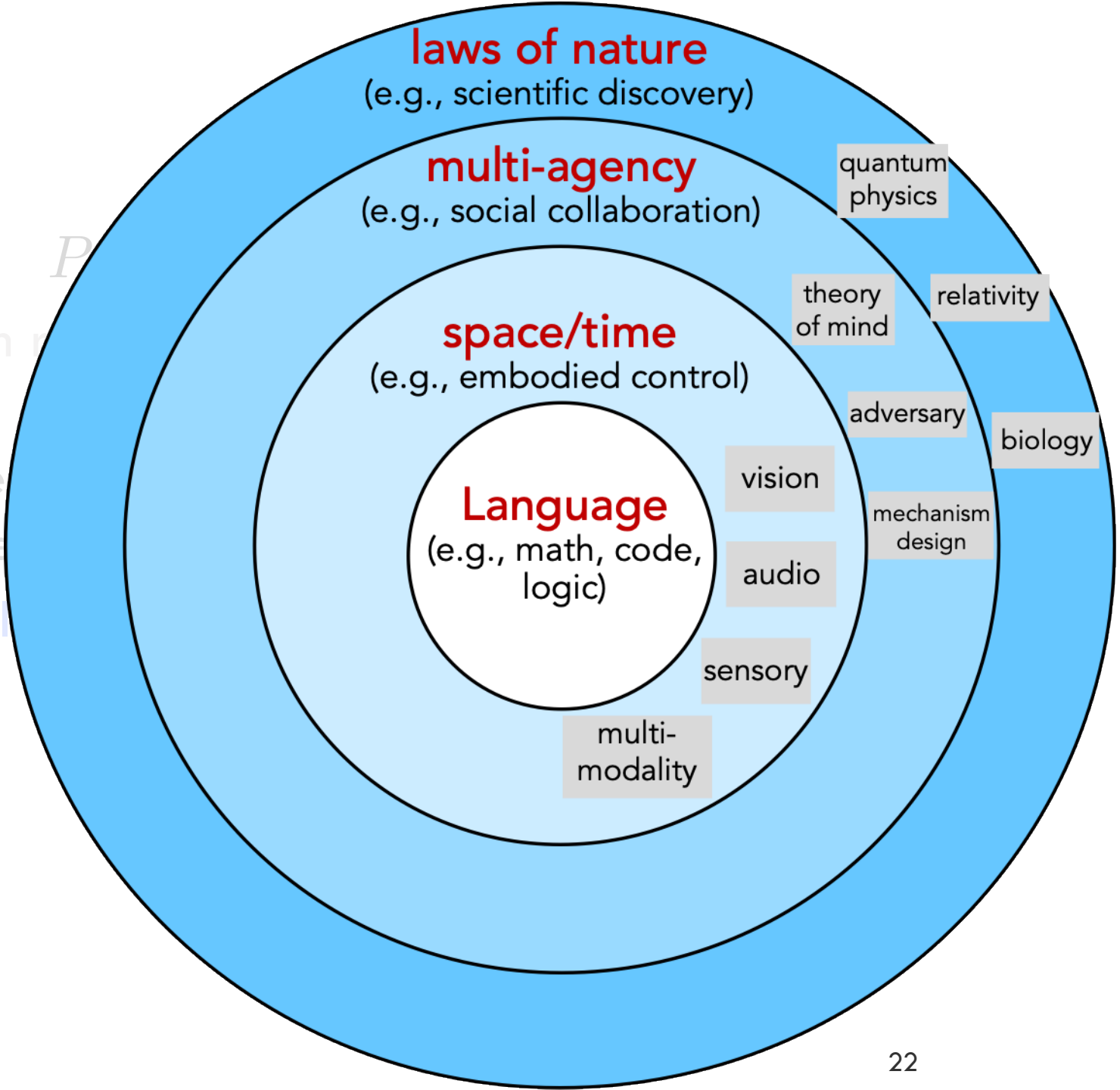
Aggregating multiple plans improves even further

Simulative reasoning outperforms *autoregressive* planning (CoT + Self-consistency)

Simulative reasoning

- Next "world" prediction
- Prior research (primarily in r
  specific world models

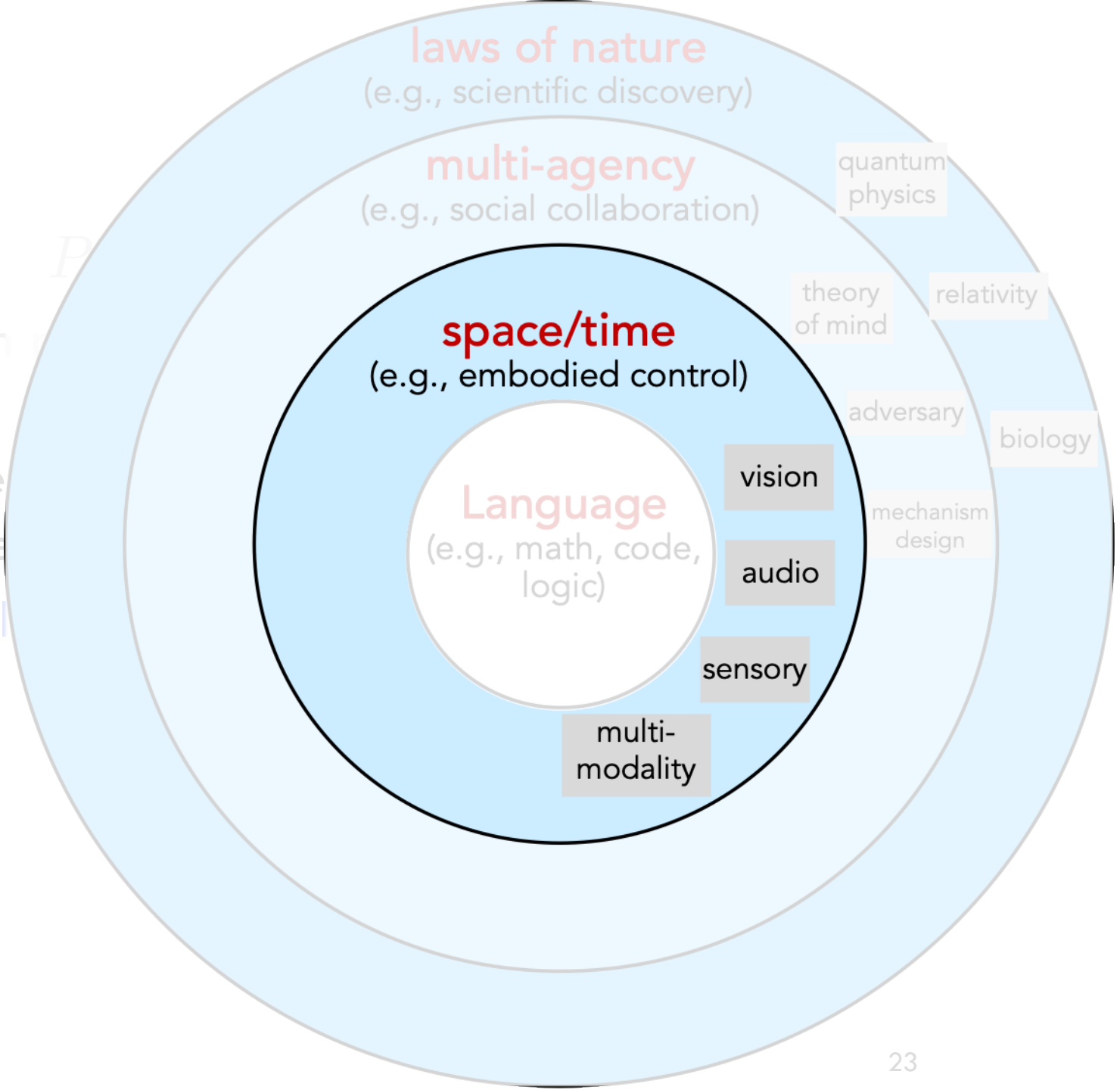
The scope of simulation defines the capability of reasoning

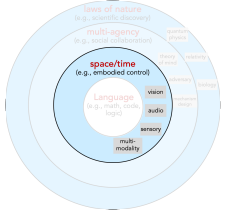


Simulative reasoning

- Next "world" prediction
- Prior research (primarily in specific world models

The scope of simulation defines the capability of reasoning





Simulative reasoning beyond LM-based world models

- Language is often **not** the most efficient medium to describe all information during reasoning
- Other modalities (e.g., videos) can be more efficient

In auto-driving: describe the street state

- Vehicles' locations & movements

Pour liquid into a glass without spilling

- Viscosity & volume of the fluid
- shape & position of the container

Simulative reasoning beyond LM-based world models

What's needed for a more general world model:

1) Integrating different spaces for simulation / reasoning: text, video, ...
2) **Generalist** language capability (like LLMs) + **generalist** vision capability (video pretraining)
3) Real-time control of the simulation through action inputs $P(s'|s,a)$
   - Controllability allows to simulate many counterfactual worlds, and pick the best to actualize
   - Existing video-generation models (e.g., Sora) are not for this

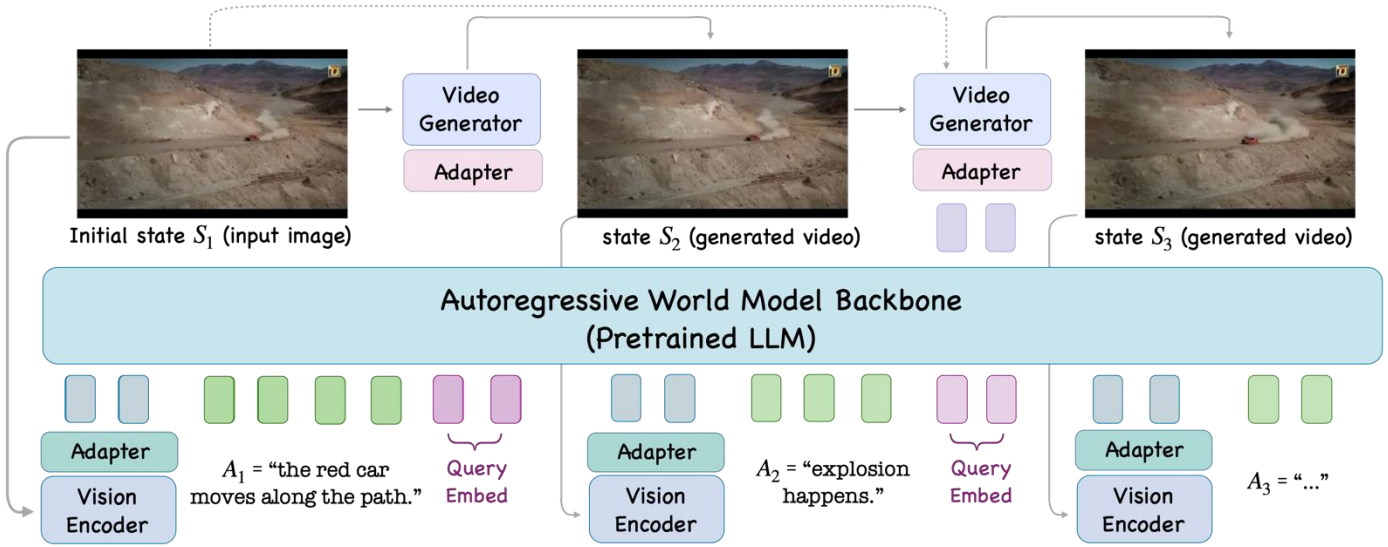
Simulative reasoning beyond LM-based world models

What's needed for a more general world model:

1) Integrating different spaces for simulation / reasoning: text, video, ...
2) **Generalist** language capability (like LLMs) + **generalist** vision capability (video pretraining)
3) Real-time control of the simulation through action inputs $P(s'|s,a)$

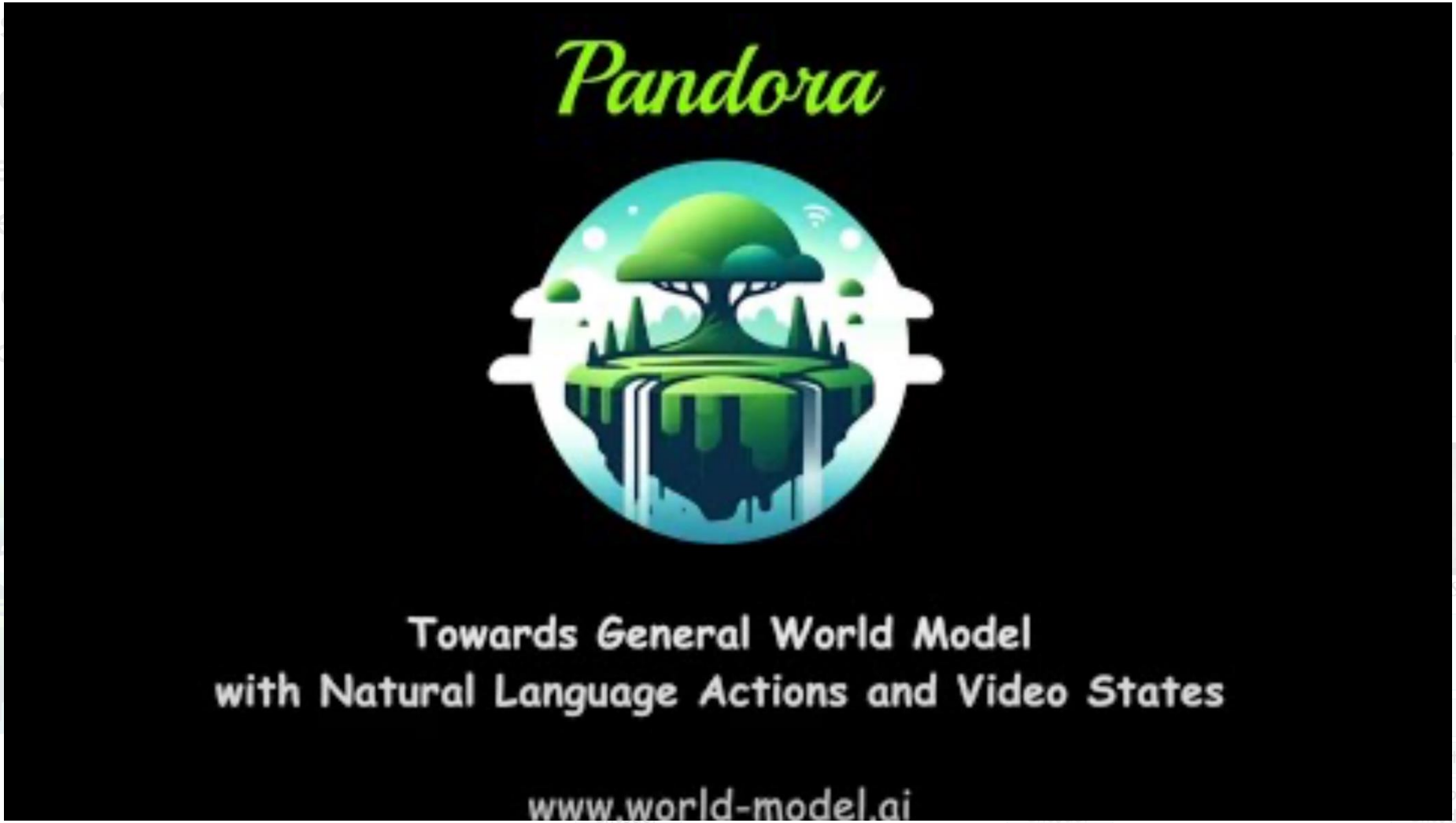
Pandora

www.world-model.ai



[Xiang*, Gu*, Liu*, et al., 2024]

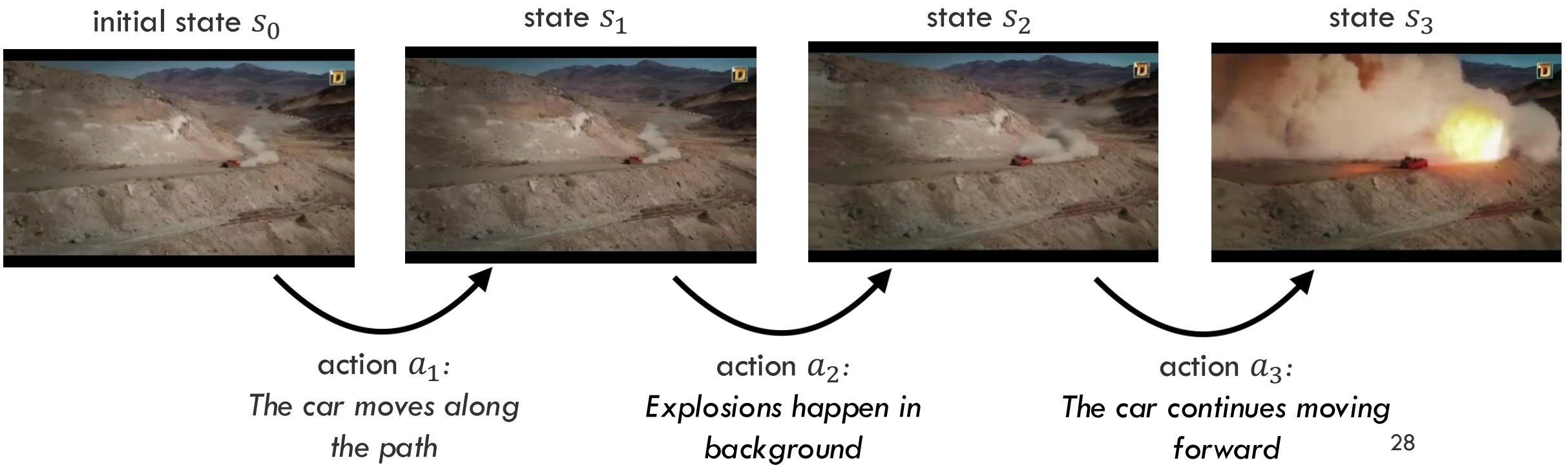
Simulative reasoning beyond LM-based world models



[Xiang*, Gu*, Liu*, et al., 2024]

Pandora: stepping towards more general world models

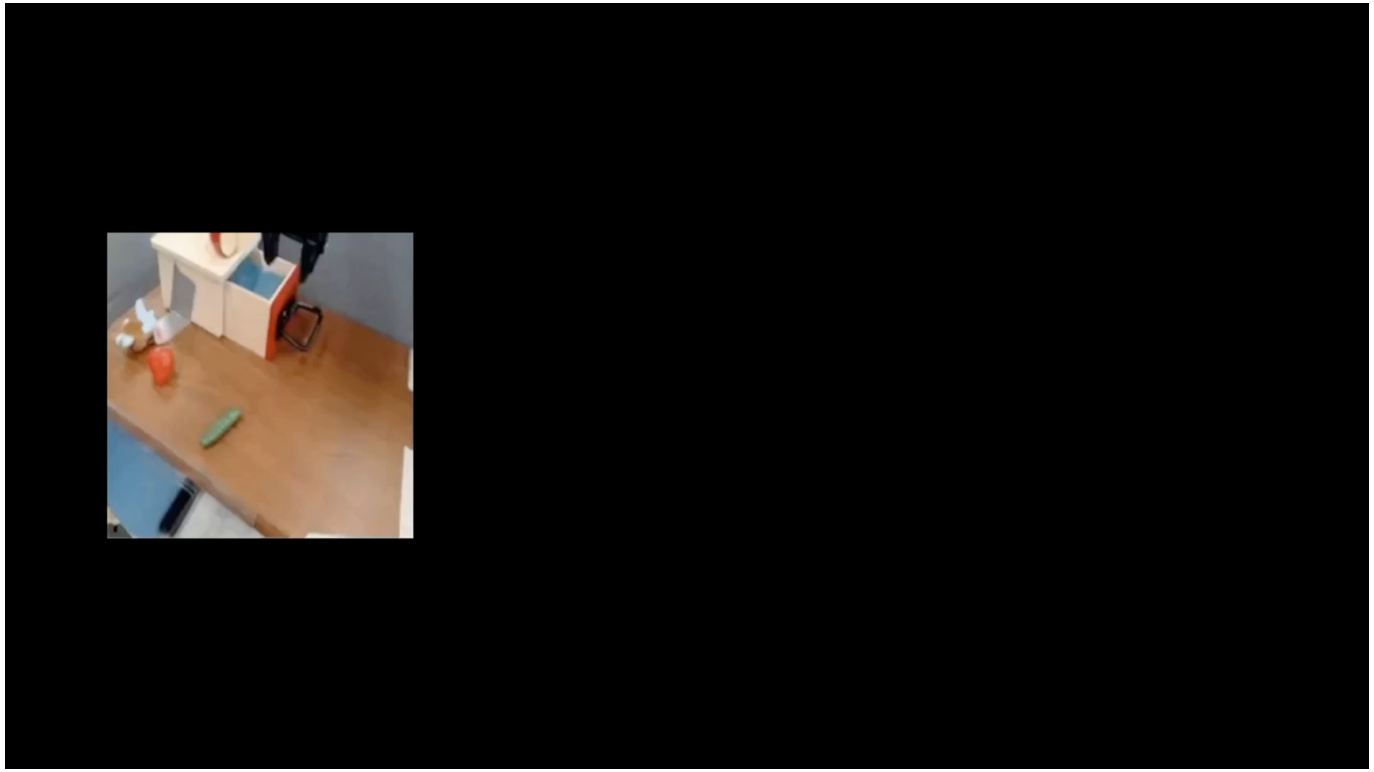
1) Integrating different spaces for simulation / reasoning: text, video, ...
3) Real-time control of the simulation through action inputs



Pandora: stepping towards more general world models

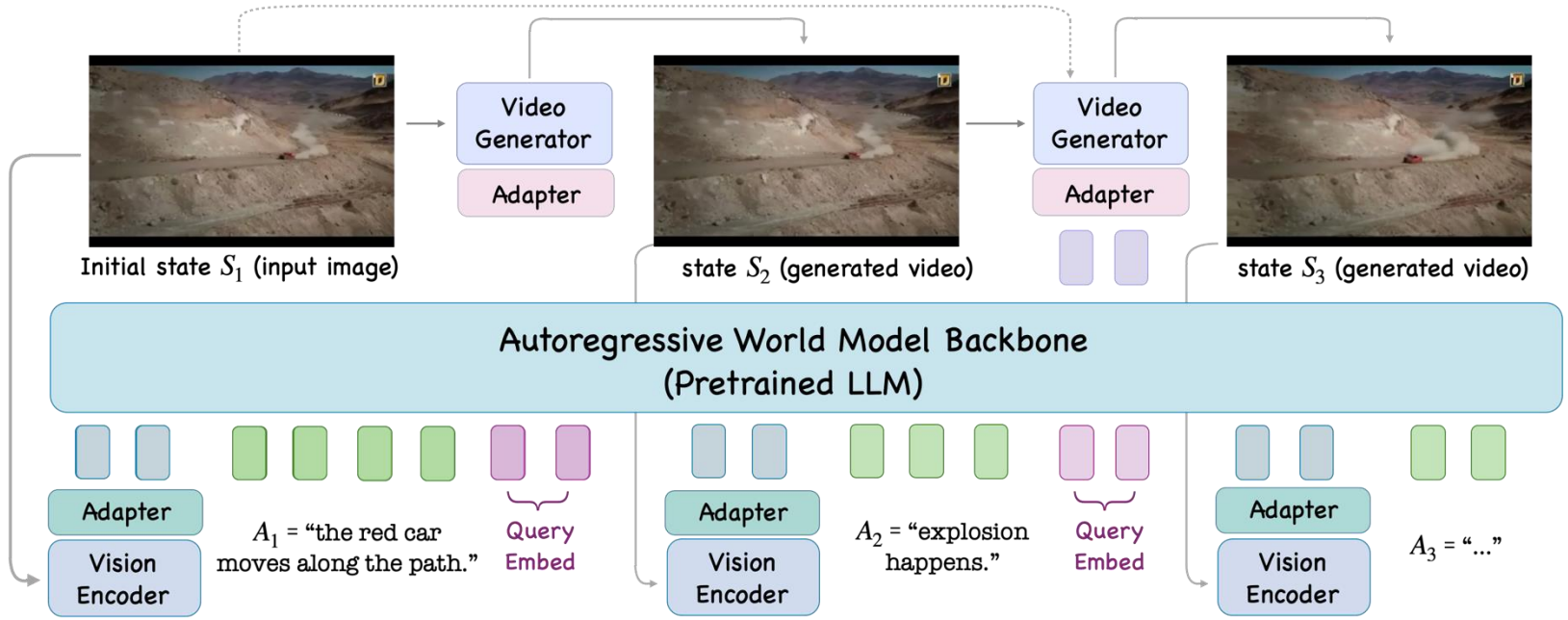
1) Integrating different spaces for simulation / reasoning: text, video, ...
3) Real-time control of the simulation through action inputs
   - Controllability allows to simulate many **counterfactual** worlds, and pick the best to actualize

Action planning for robots



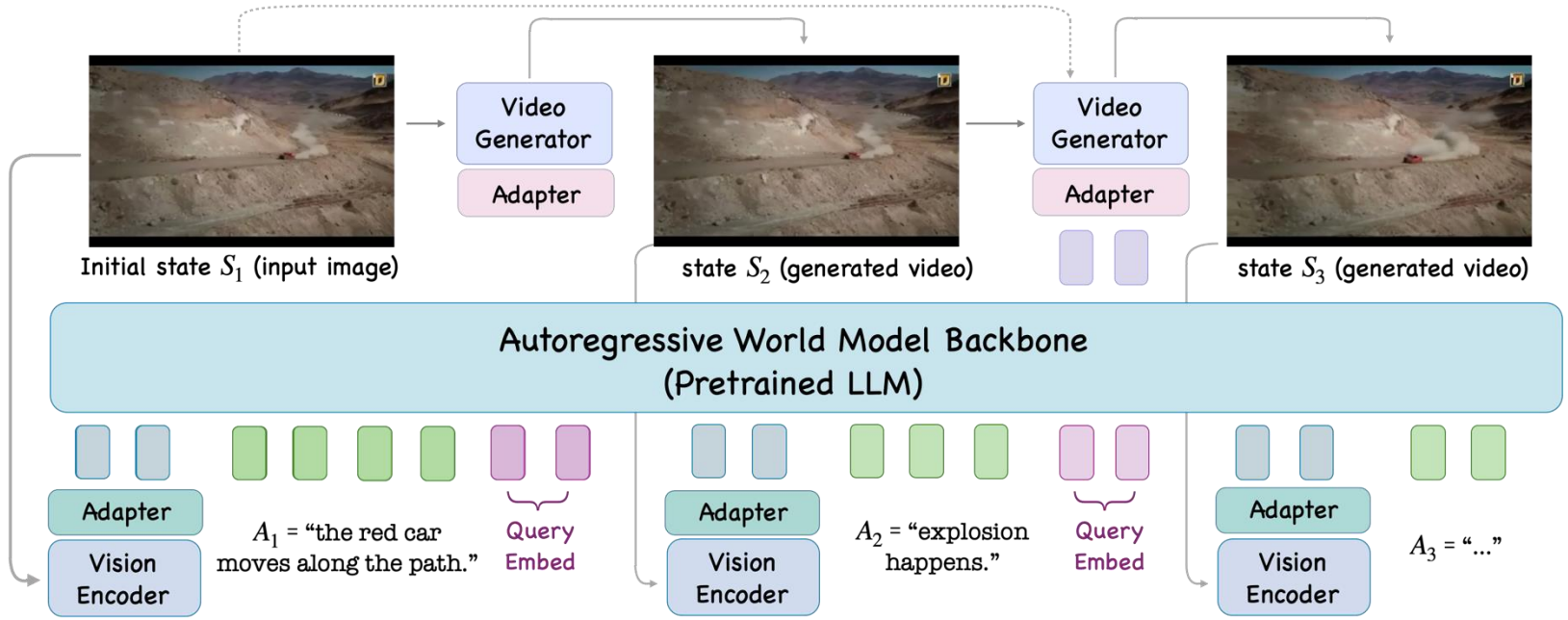
Pandora: stepping towards more general world models

2) Generalist language capability (like LLMs) + generalist vision capability (video pretraining)
   - Generalist pretrained **LLM** as the autoregressive backbone
   - Generalist pretrained **Video Diffusion Model** for visual simulation
     - Massive video pretraining enables consistent prediction of the physical world states
   - Instruction-tuning connects the backbone with video generator for real-time control
     - Using small high-quality real-time control text-video data



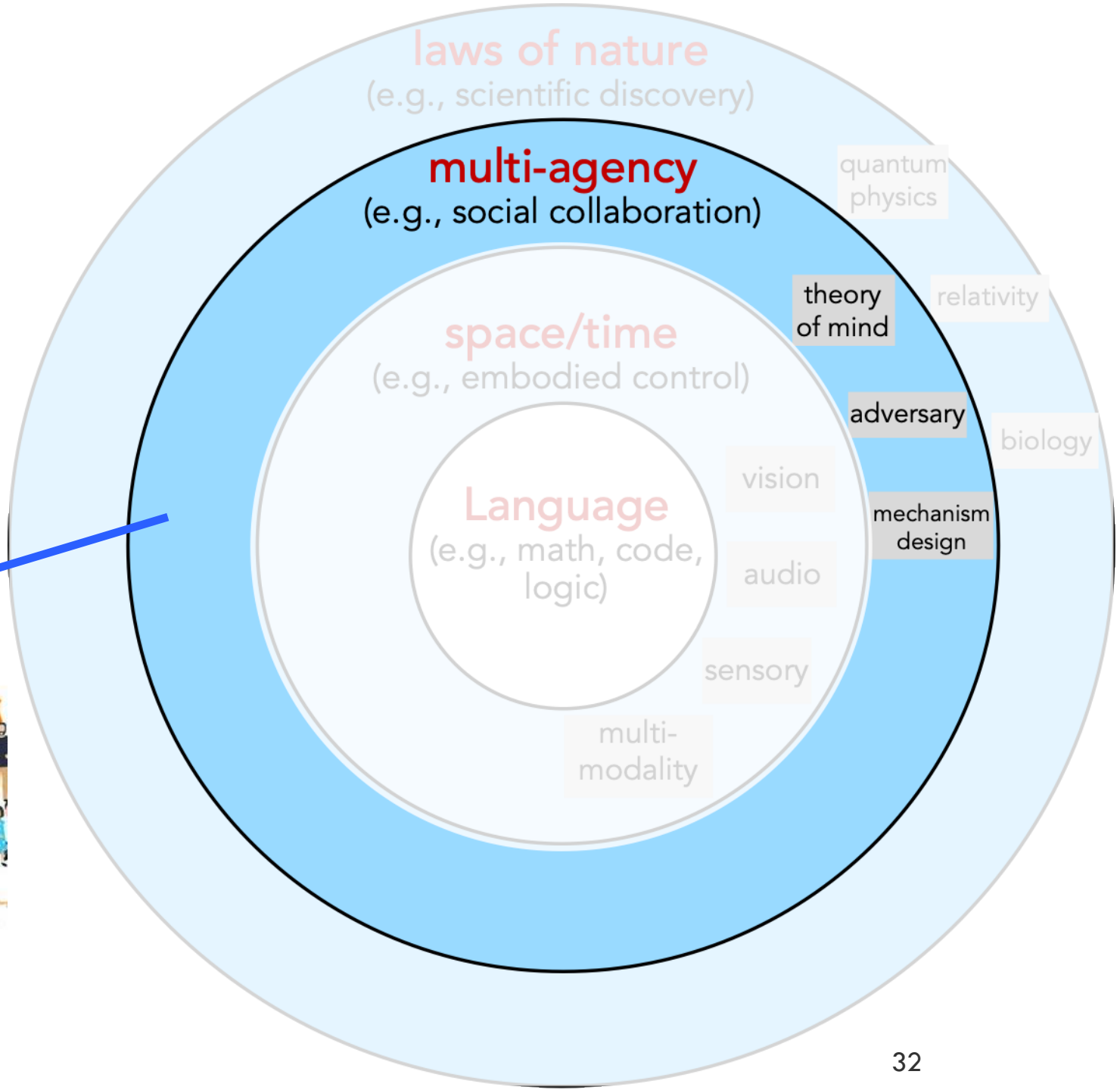
Pandora tepping towards more general world models

2) Generalist language capability (like LLMs) + generalist vision capability (video pretraining)
   - Generalist pretrained **LLM** as the autoregressive backbone
   - Generalist pretrained **Video Diffusion Model** for visual simulation
     - Massive video pretraining enables consistent prediction of the physical world states
   - Instruction-tuning connects the backbone with video generator for real-time control
     - Using small high-quality real-time control text-video data

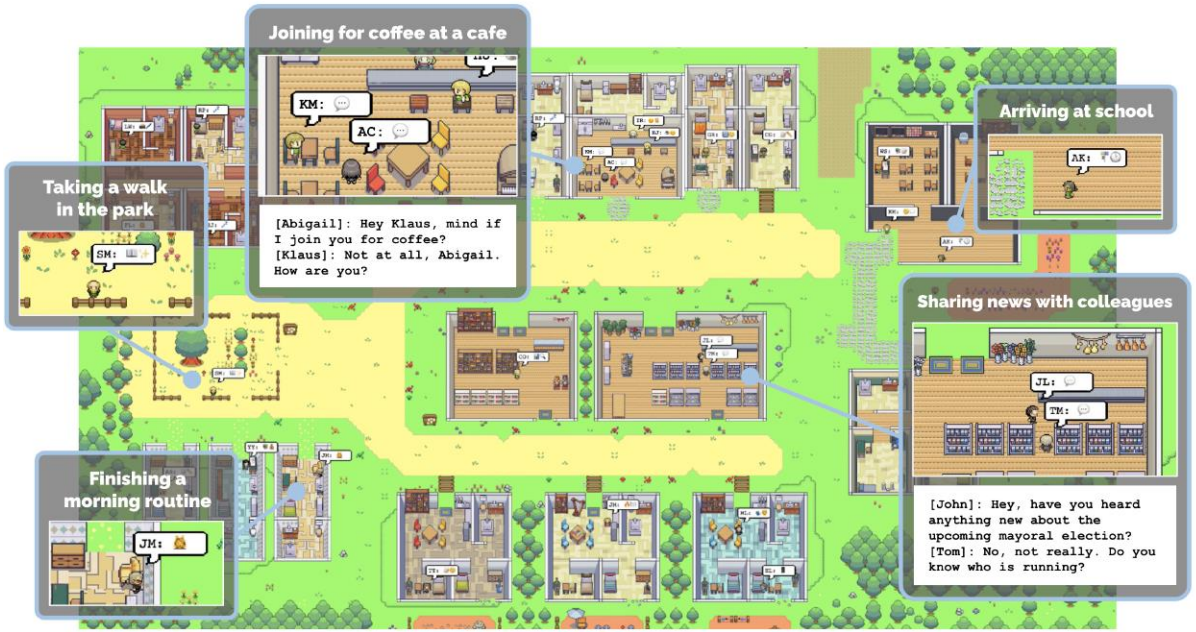


Simulative reasoning

Society of individual world models



Simulative reasoning

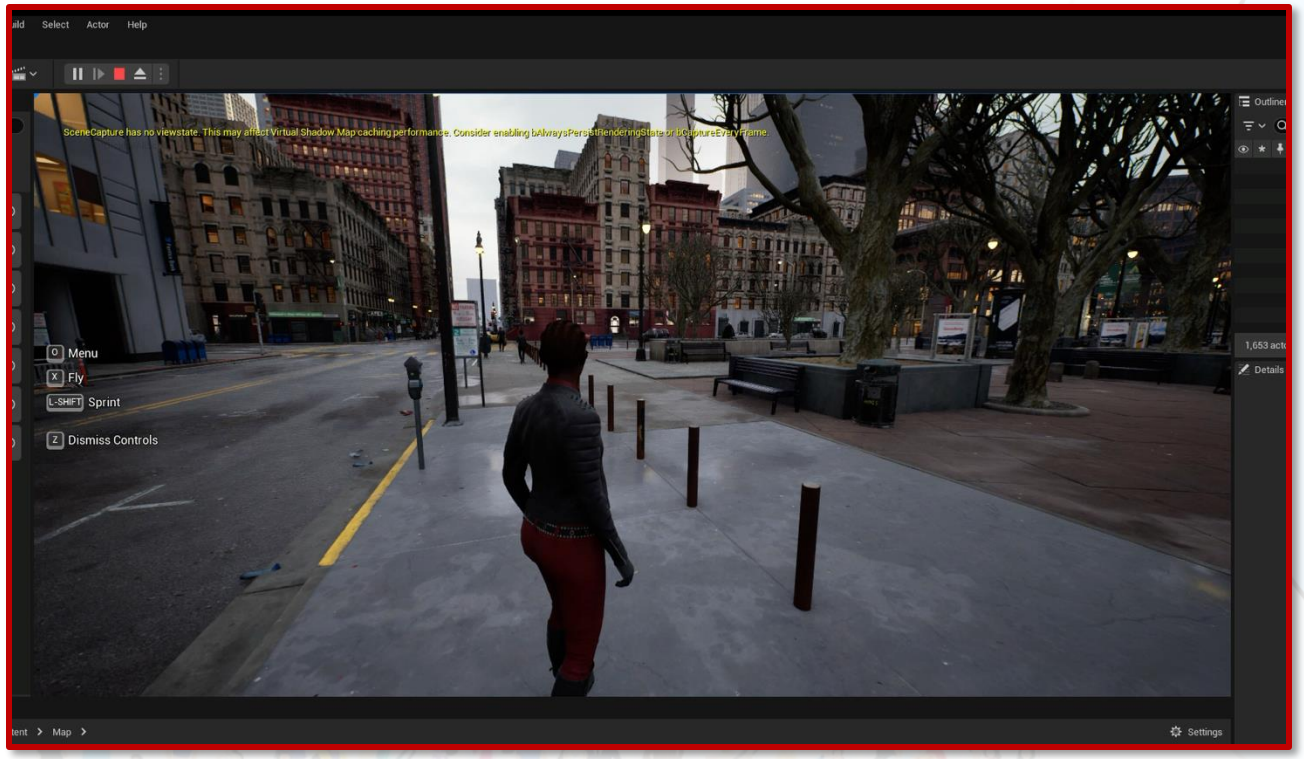


[Park et al., 2023]

25 agents, each controlled by individual LLM, converse with each other

- For studying emerging communication behaviors

Simulative reasoning

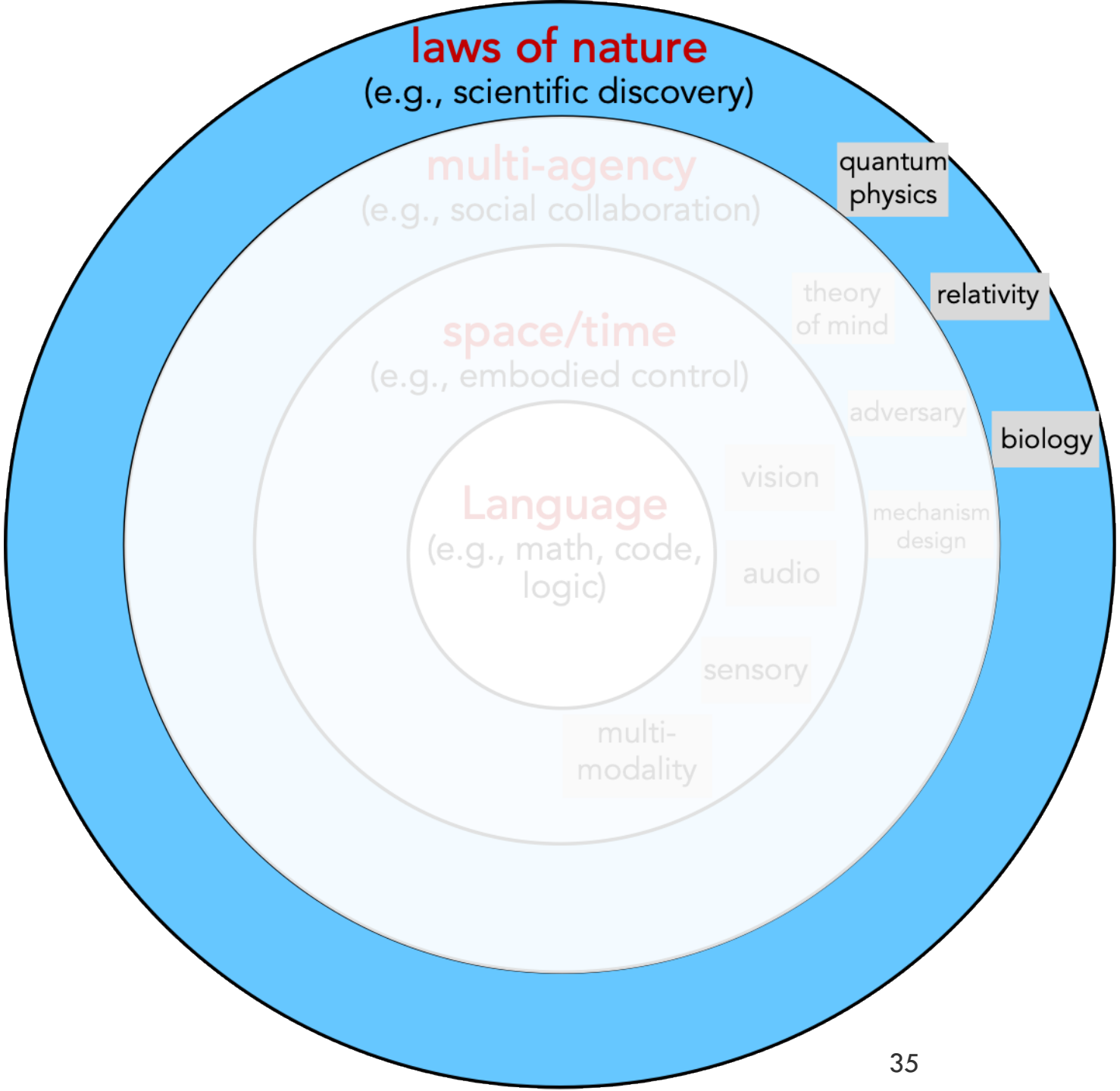


In progress

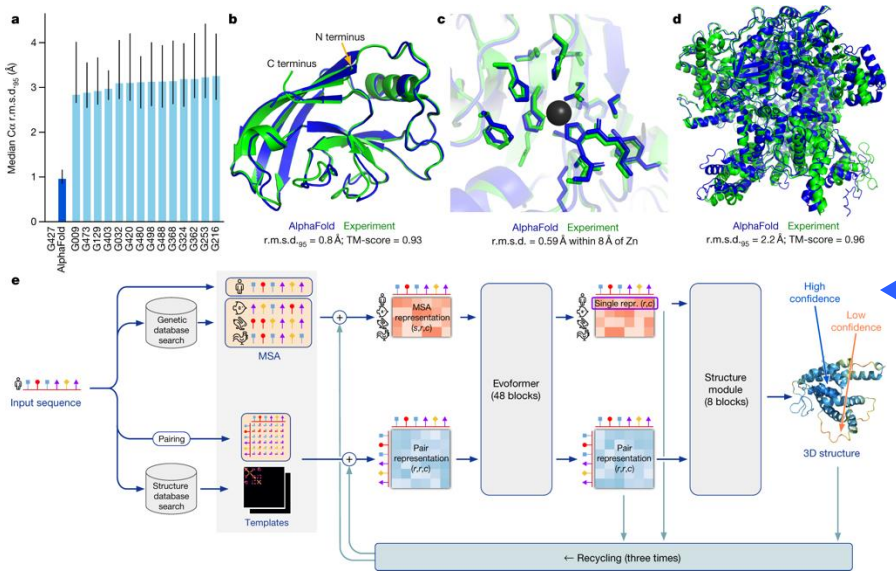
Richer and more realistic simulation of society

- Humans, vehicles, robots
- Simulating traffic, social, financial systems
- Could potentially be used for studying human-AI collaboration, education, social science, policy making, …

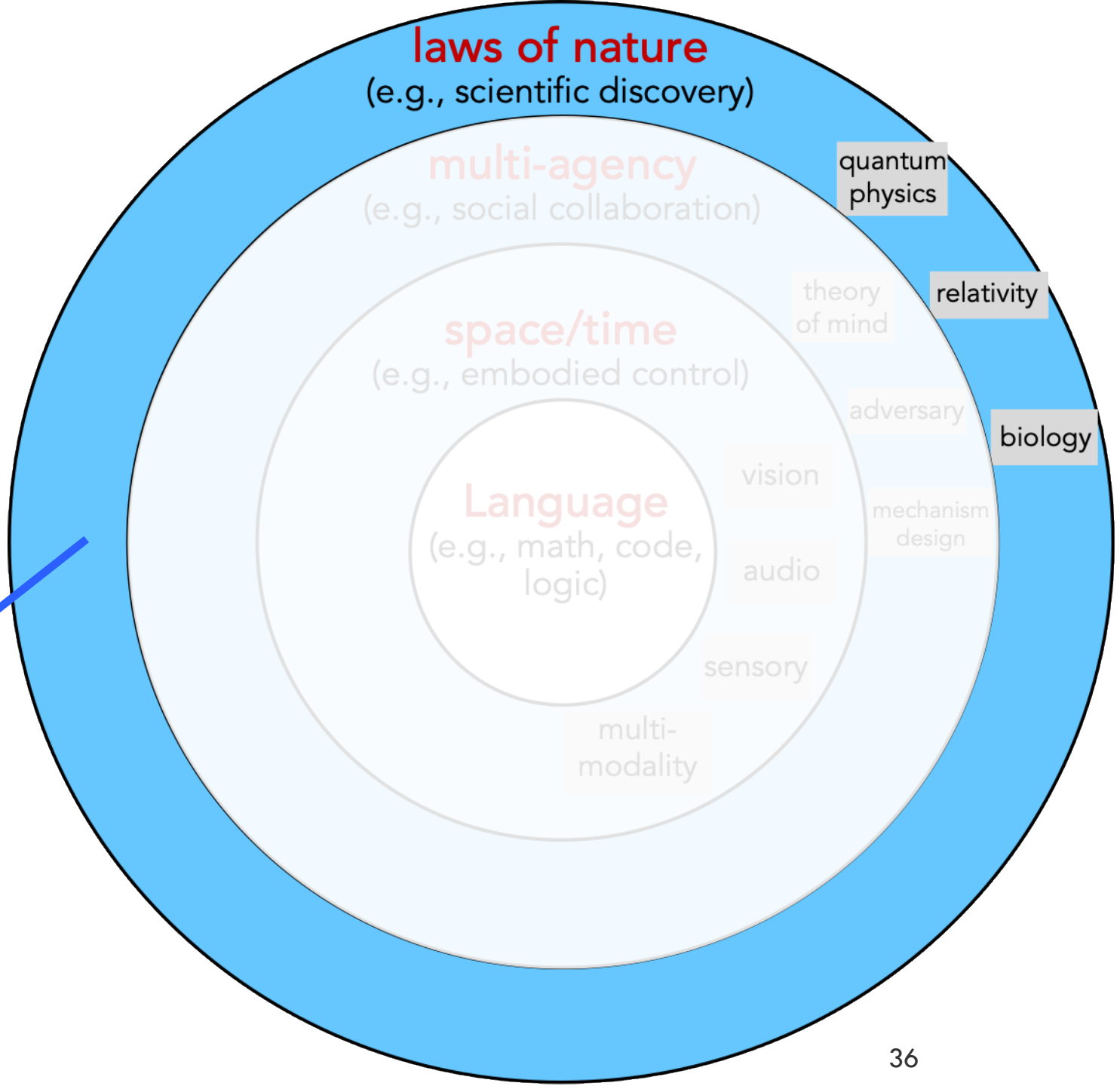
Simulative reasoning

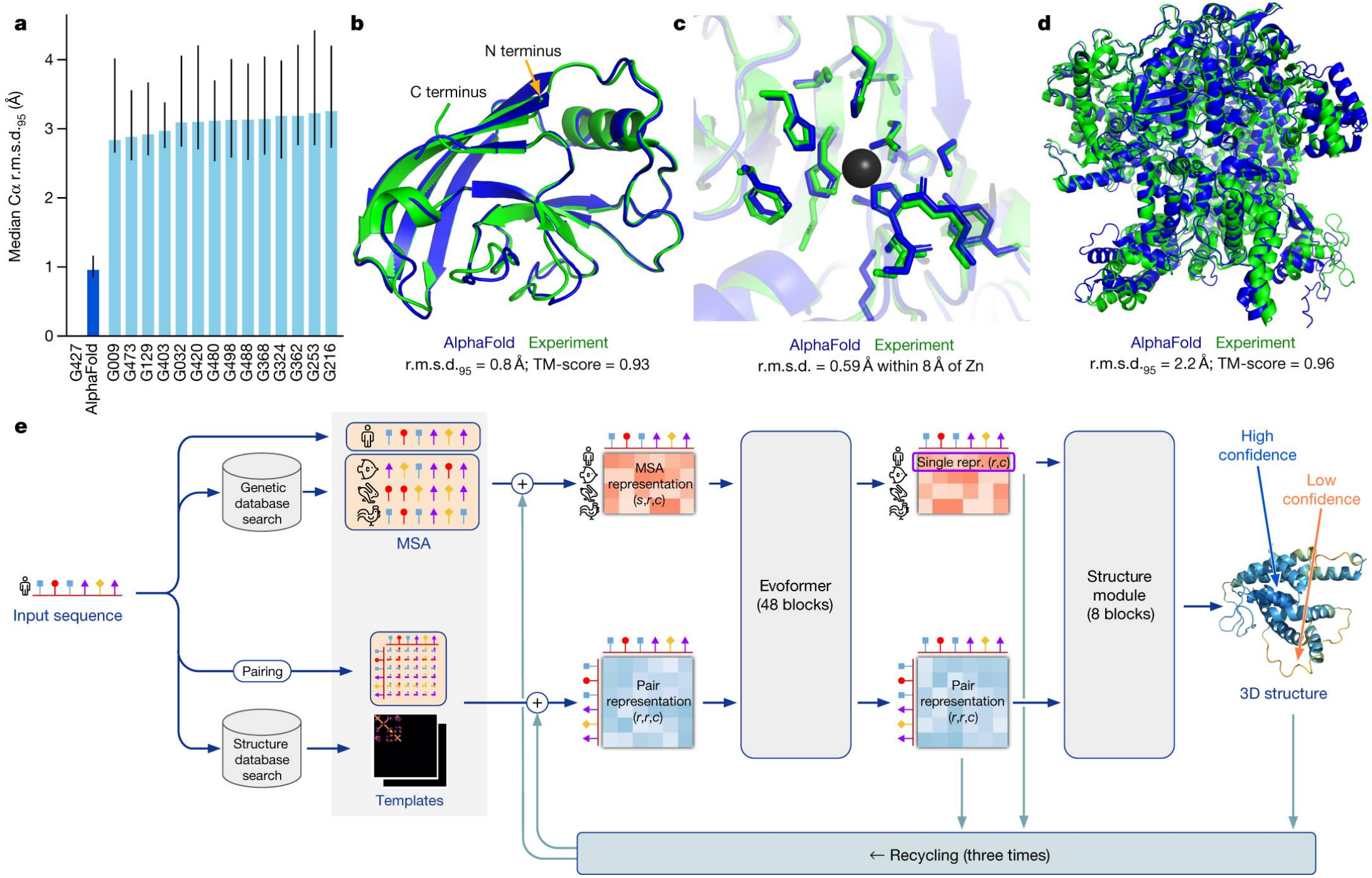


Simulative reasoning



AlphaFold
for protein structure prediction




a
Median Cα r.m.s.d.95 (Å)
0
1
2
3
4
G427
AlphaFold
G009
G473
G129
G403
G032
G420
G480
G498
G488
G368
G324
G362
G253
G216
b
N terminus
C terminus
AlphaFold Experiment
r.m.s.d.95 = 0.8 Å; TM-score = 0.93
c
AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn
d
AlphaFold Experiment
r.m.s.d.95 = 2.2 Å; TM-score = 0.96
e
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

MACE: A foundation model for atomistic materials chemistry

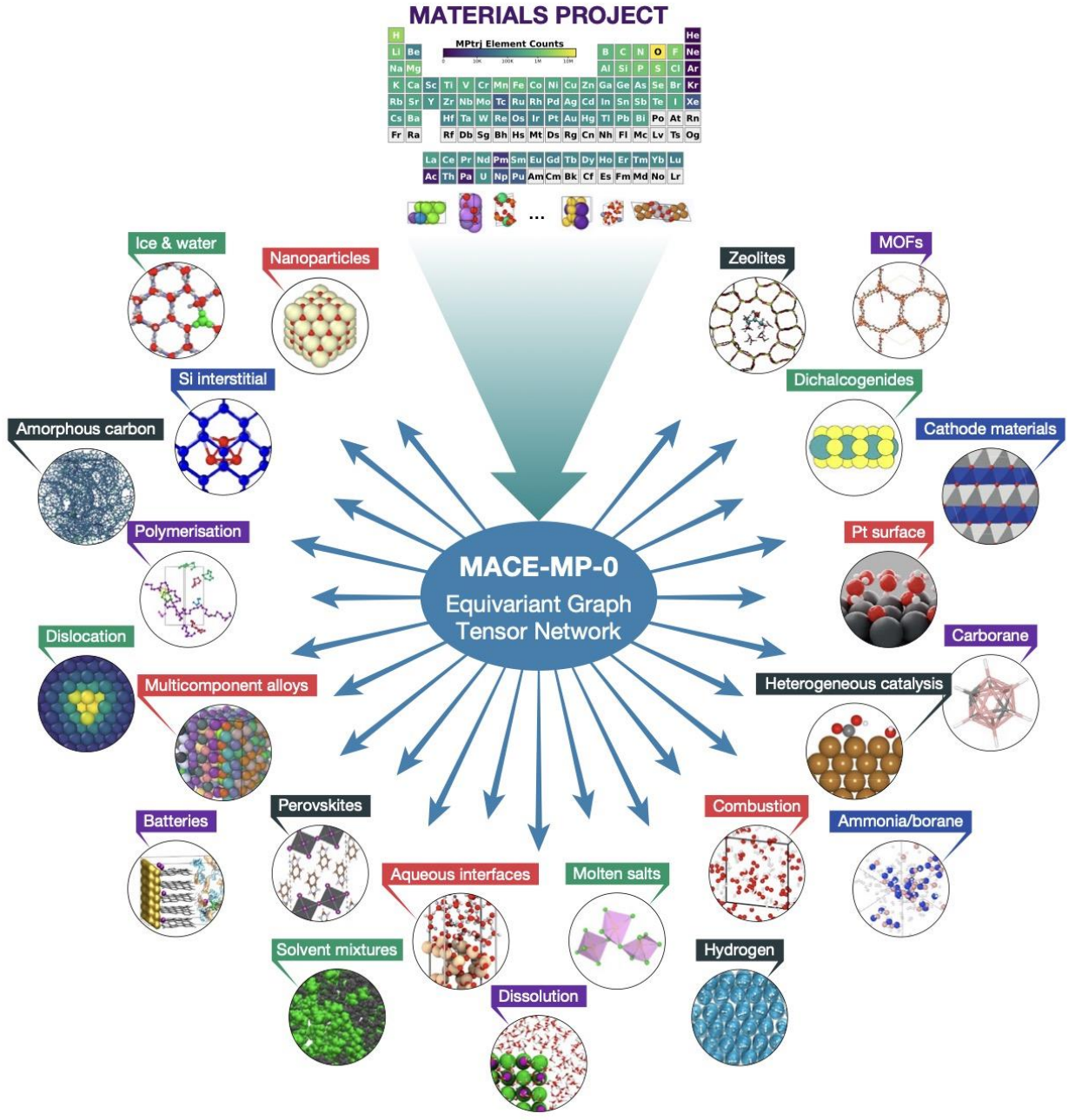
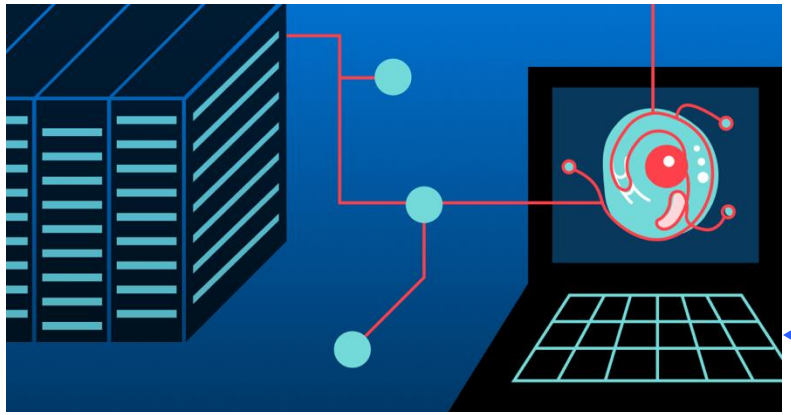
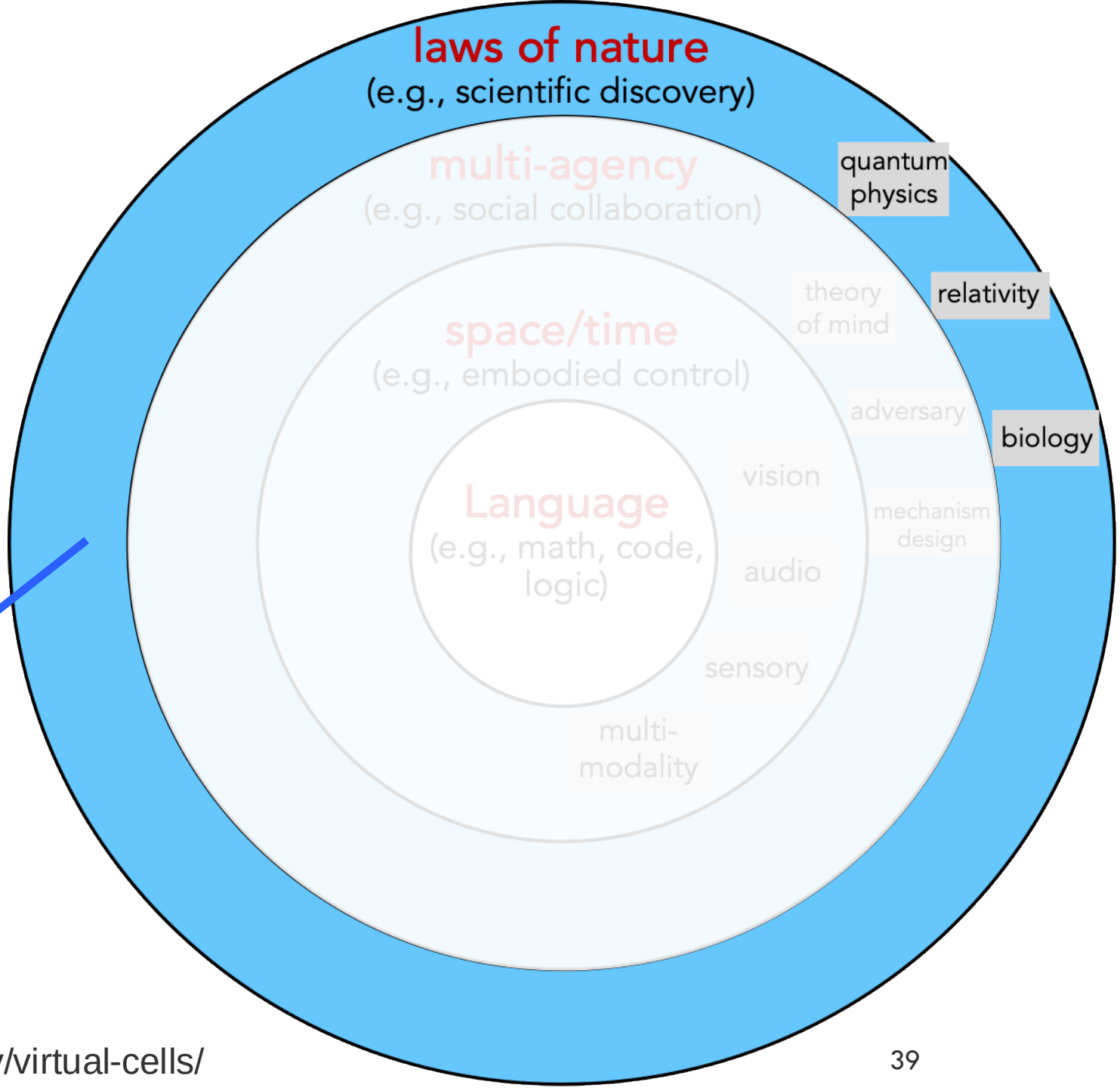


Figure 1: **A foundation model for materials modelling.** Trained only on Materials Project data (*19*) which consists primarily of inorganic crystals and is skewed heavily towards oxides, MACE-MP-0 is capable of molecular dynamics simulation across a wide variety of chemistries in the solid, liquid and gaseous phases.

Simulative reasoning

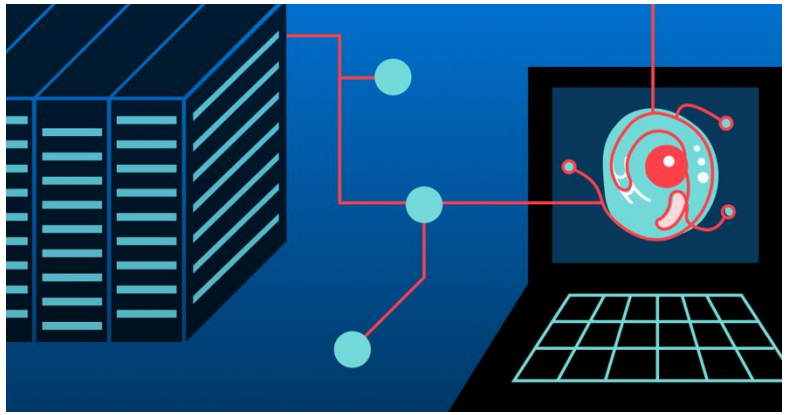


AI Virtual Cell
for drug discovery, etc.

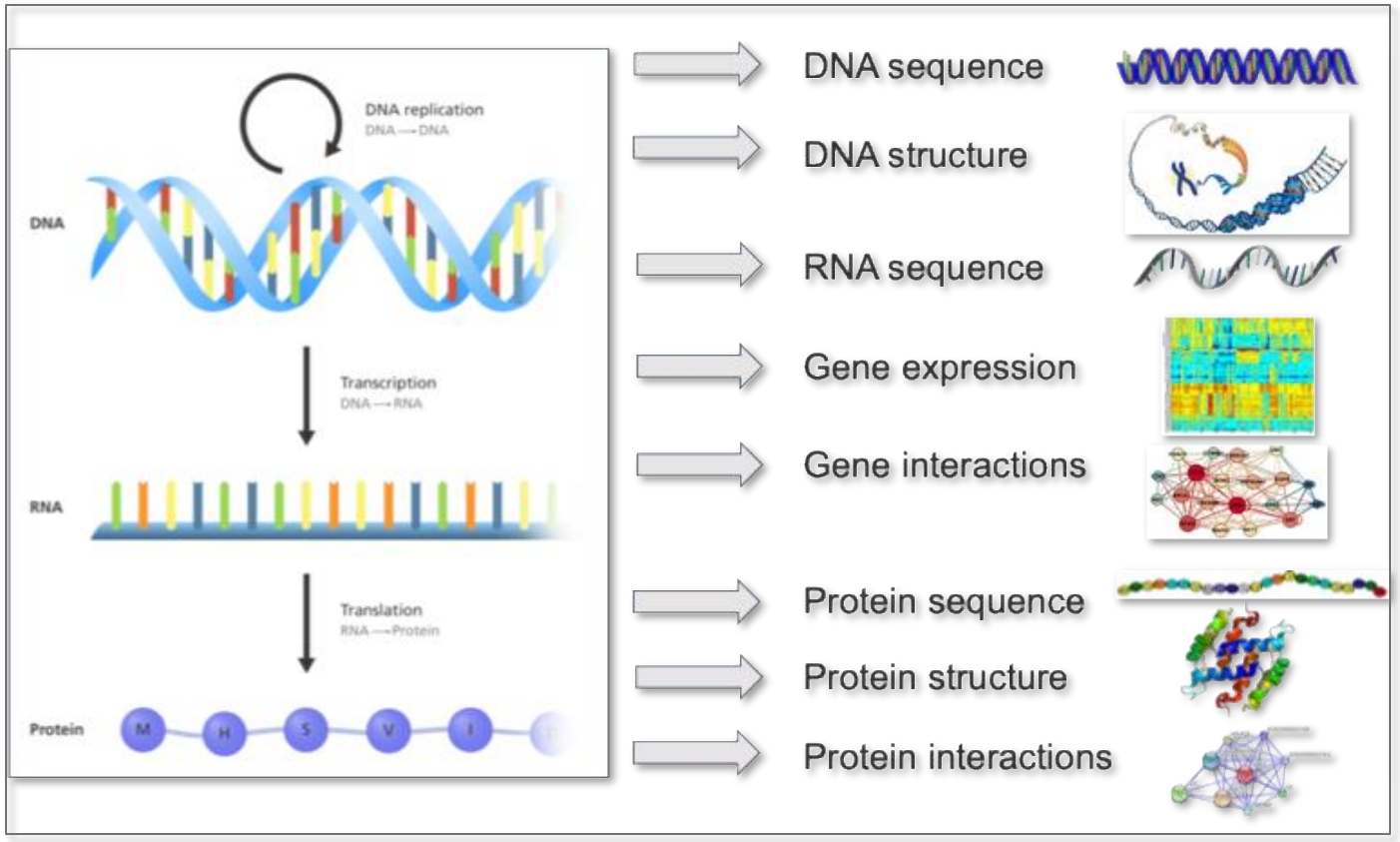


https://chanzuckerberg.com/science/technology/virtual-cells/

Simulative reasoning



AI Virtual Cell
for drug discovery, etc.



Extensive heterogenous biological data
(In progress)

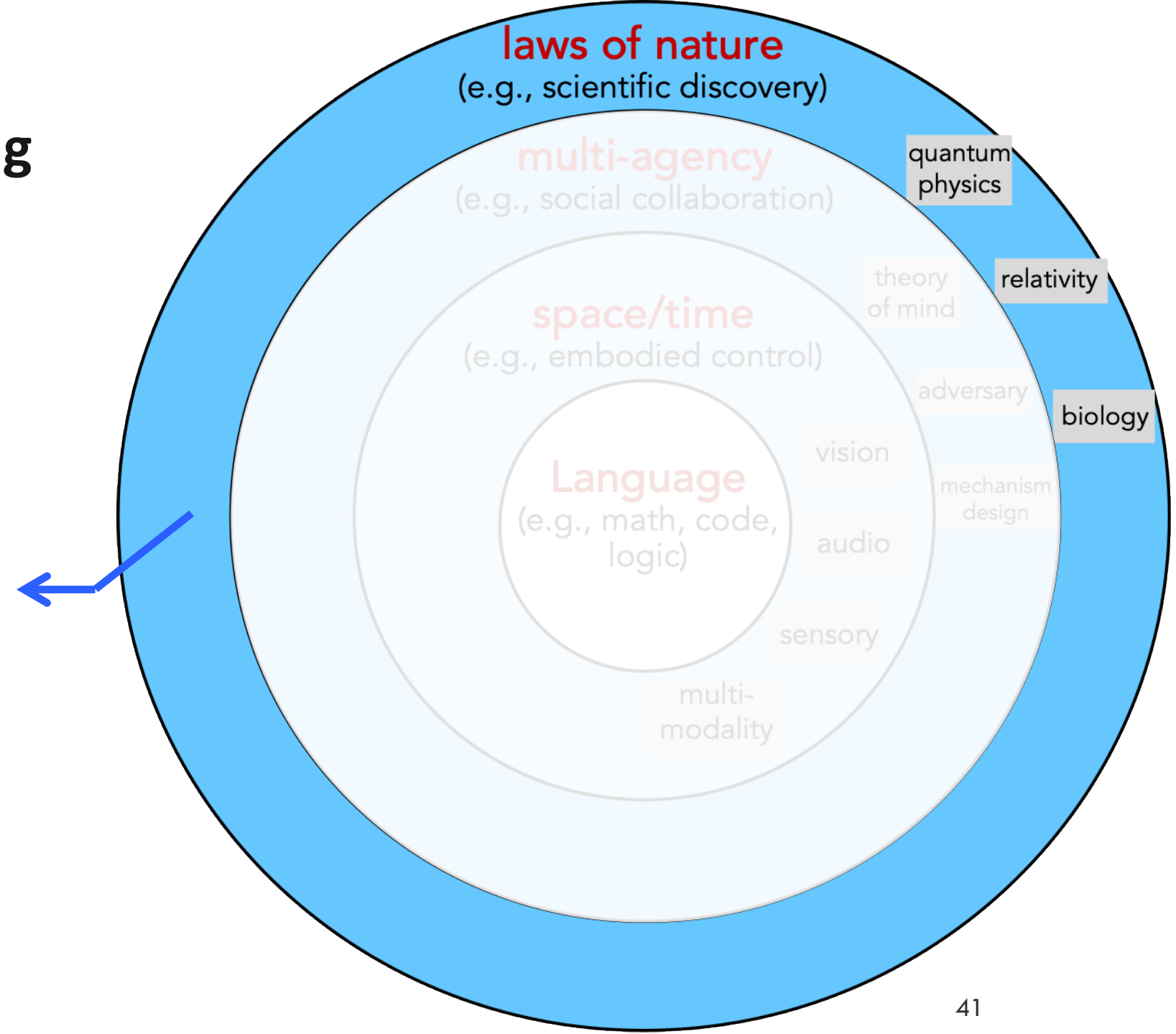
https://chanzuckerberg.com/science/technology/virtual-cells/

Simulative reasoning

Billions of years on Earth's evolutionary path

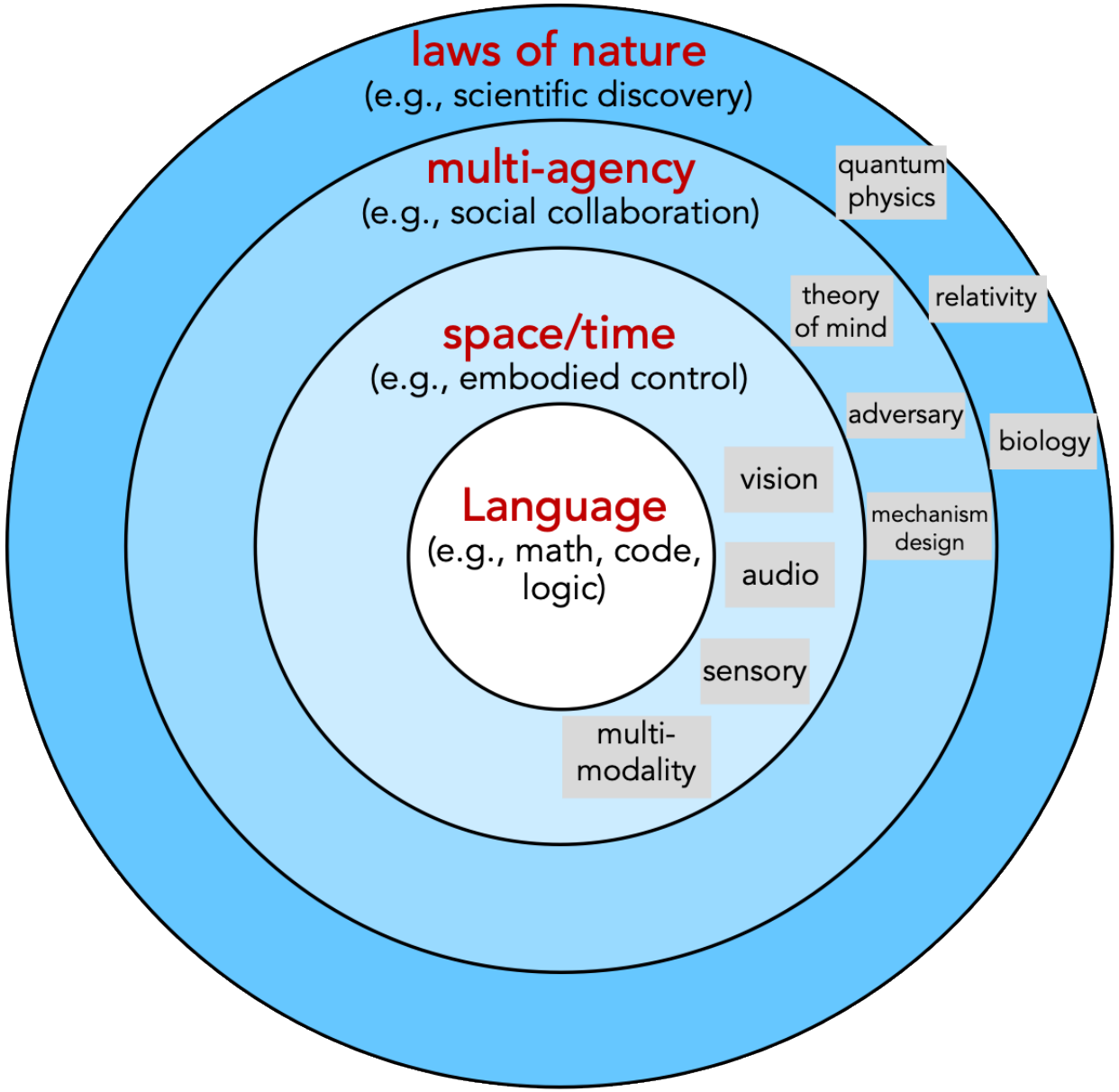
v.s.

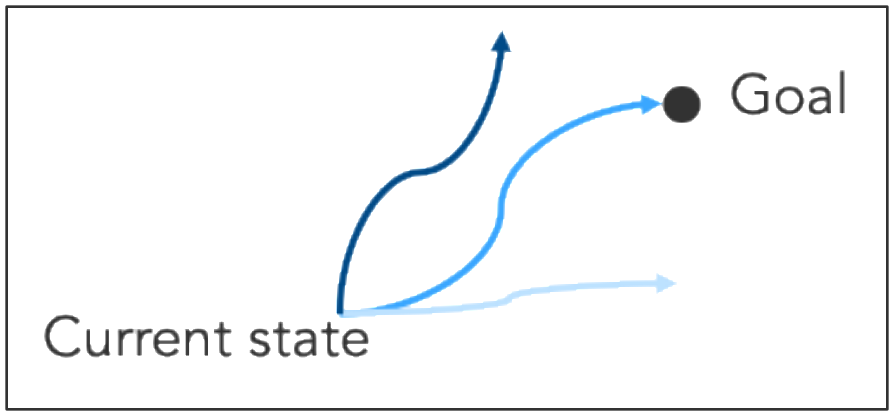
Moments of simulating Mars' civilization strategy

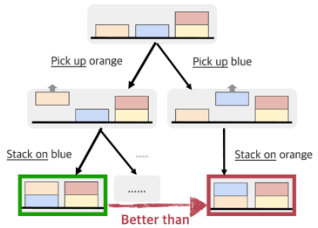


Summary

- Simulative reasoning based on world models
  - strategic planning via simulation
- "More simulation, more intelligence"
- Building general world models
  - Pandora

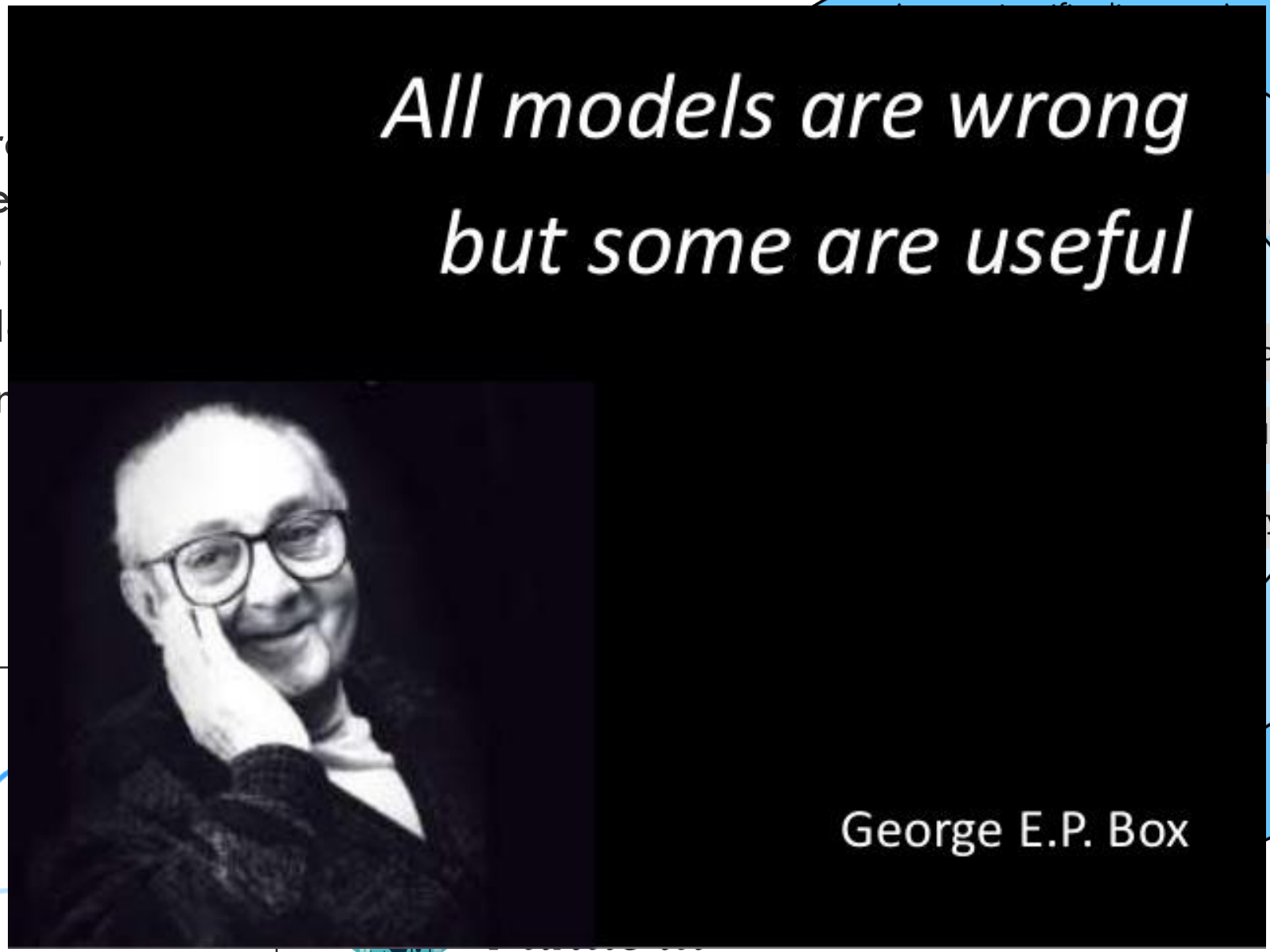


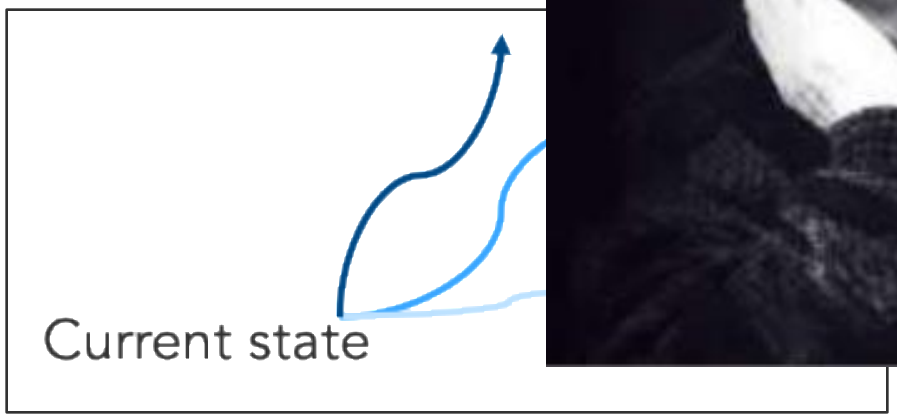




Summary

- Simulative r
  world mode
  - strategic p
- "More simul
- Building ger
  - Pandora





Questions?