Reverse-engineering the Self

L.A. Paul, Tomer D. Ullman, Julian De Freitas, and Joshua B. Tenenbaum

Presenter: Hanqi Zhou 2023.08.08

Background

- **Significant progress made in**
  - Machine learning models that can outperform human experts in a wide range of tasks
  - Cognitive models for planning and decision-making

- **A crucial gap exists in these models**
  - They lack an understanding of who or what is behind the planning and decision-making processes.

Motivation

- **The Need for a Self-aware Agent**
  - How to build an intelligent agent that behaves in a human-like manner?
  - The answer to this question is the agent itself, but the computational structure of such an agent remains undefined.
- **Beyond Data and Planning**
  - The missing component in current models cannot be addressed merely by adding more data or refining the planning framework.
  - The computational structure of the agent should support its ability to learn and think autonomously.

What is *think for itself*?

A truly intelligent agent that can think for itself

What is *think for itself*?

A truly intelligent agent that can think for itself

- **Definition**
  - An agent can recognize when its actions aren't leading to the desired outcome in unanticipated situations and can adjust its plans accordingly.
- **Features**
  - It can execute its plans in a flexible, general-purpose manner, learning autonomously without needing external intervention, guidance or reprogramming.

What is *think for itself*?

A truly intelligent agent that can think for itself (Cont’d)

- **Self-driving Car**
    - **Existing Systems**
        - Operate based on predefined rules and sensors' data.
        - React to specific situations based on pre-programmed responses.
    - **Truly Intelligent Agent**
        - Understand the context, anticipate potential future scenarios, and make decisions accordingly.
        - Learn from new traffic situations, adapt to different driving cultures, and even understand the intentions of pedestrians or other drivers, adjusting its behavior autonomously.

What is *think for itself*?

A truly intelligent agent that can think for itself (Cont’d)

- **A Truly Intelligent Agent**
  - Have a representation or model of its environment and its function within that environment.
  - Recognize and adapt to unanticipated problems.
  - Change its plans and identify new problems autonomously.
  - Execute plans flexibly and learn autonomously without external reprogramming.

What is *think for itself*?

Key components: centering & re-centering

- **Definition**
  - Centering: An intelligent agent represents itself as distinct from yet located within its environment.
  - Re-centering: a significant realignment or effective replacement of one model with another, as opposed to minor refinements.

What is *think for itself*?

Example

- **Scenario**
  - You wake up to pitch-black darkness in an unfamiliar room.
  - Feeling your way around, you touch a bedside table, then a lamp, and eventually find the light switch.
  - After turning on the light and leisurely glancing at the clock, you suddenly realize that you are late for a meeting where you are scheduled to present.
  - You quickly gather your belongings and rush out the door.

What is *think for itself*?

Example (Cont’d)

- **Initial Centering**: When you first awoke, you found yourself in a dark room, and your immediate problem was the need for light. This translated into the task of finding a light switch.

- **Re-centering**: The moment you glanced at the clock and realized you were late, the nature of your task changed - you had a more pressing problem of getting to your meeting on time. This realization required a shift in perspective.

- **Computational Feats**: To transition from the simple task of turning on the light to the more complex task of getting to your meeting on time, you had to perform several sophisticated computational tasks:

  - Recognize that your current actions (finding the light switch) were not aligned with your new goal (getting to the meeting on time).

  - Replace your old problem (navigating the dark room) with the new problem (reaching the meeting venue as quickly as possible).

  - Adjust your plans and actions based on this new centered perspective.

This work proposes …

A conceptual framework

- A truly intelligent agent can think, plan, and act autonomously, adapting to new situations without external intervention.
- The ability to center oneself and re-center oneself is a hallmark of intelligence.
  - This ability to "center" and "re-center" is crucial for an agent to adapt and respond to changes in its surroundings effectively.

This work proposes …

A conceptual framework with evidence

- Centering within space
  - For someone unfamiliar with maps, understanding that the dot represents their current location from a bird's eye view can be challenging.
  - This illustrates the cognitive leap required to understand one's position from a third-person perspective.

YOU
ARE
HERE
NOW

This work proposes …

A conceptual framework with evidence

- ○ Representing oneself within a virtual environment
  - When controlling an avatar in a video game or virtual world, the player must understand that the avatar represents them within that space.

Evidence of human

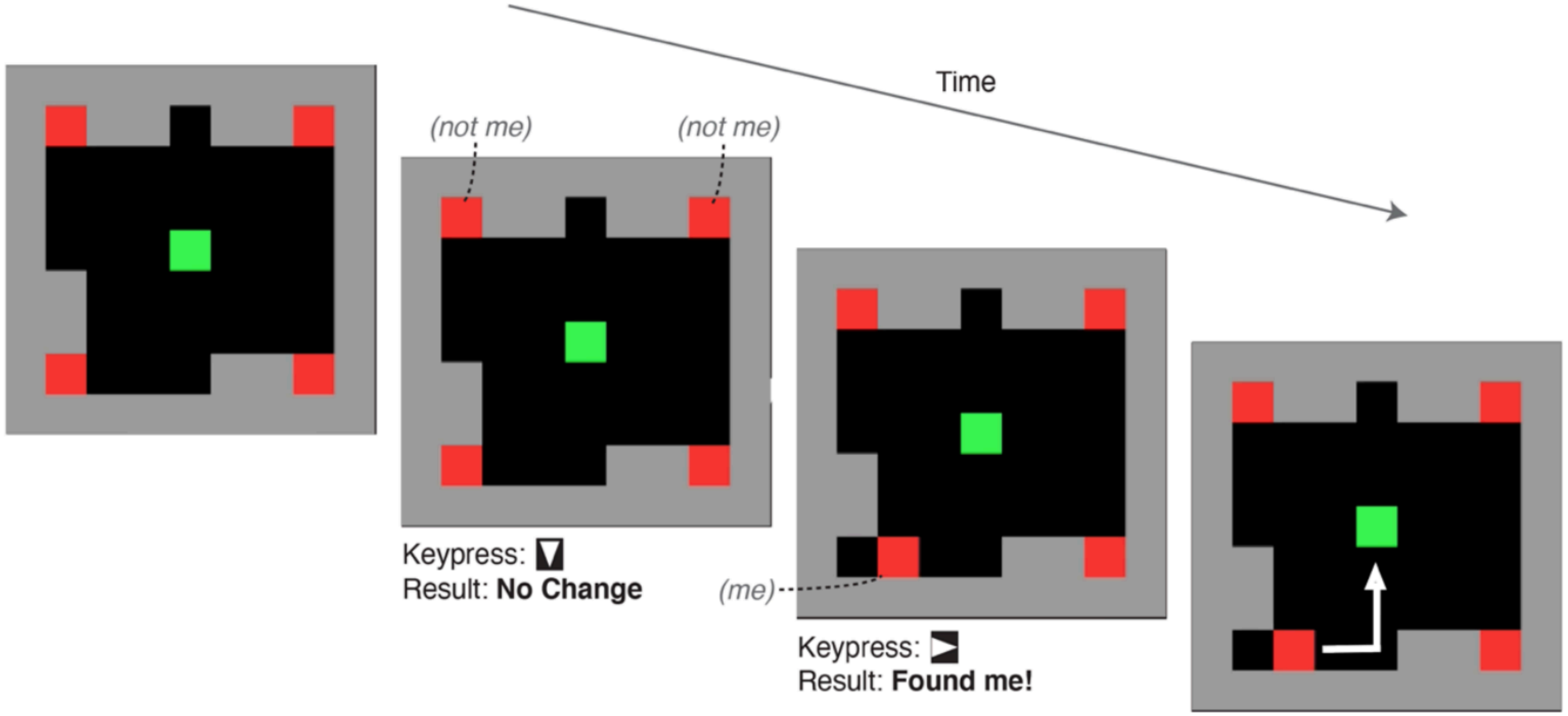
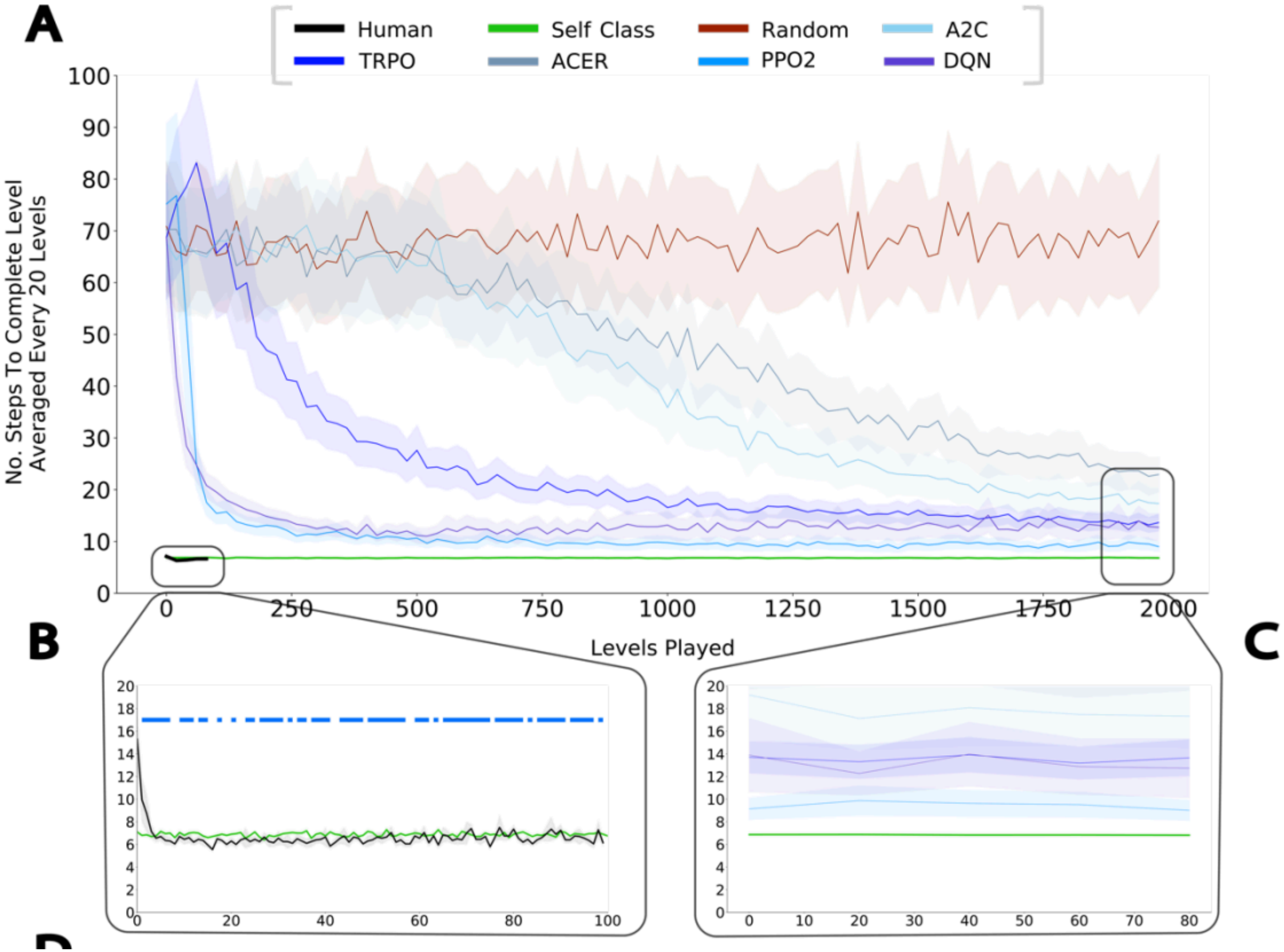


Figure 1. The Logic Game. There are 4 agents (red blocks), one of which is your avatar. The level ends when the avatar reaches the goal (green block). In this example, moving DOWN disambiguates the most possible selves (red)—the top two. If moving down produces no visible change, then you must be one of the bottom two agents. In order to disambiguate which of these bottom agents is your digital self, it is now equally informative to move RIGHT or UP. Moving RIGHT reveals that the digital self was in the bottom left corner. Knowing this, you navigate it to the reward (green).

https://www.hbs.edu/ris/Publication%20Files/23-017_e1f7f89b-20ed-4aaf-aaf4-36cdc08470cf.pdf

Evidence of human



https://www.hbs.edu/ris/Publication%20Files/23-017_e1f7f89b-20ed-4aaf-aaf4-36cdc08470cf.pdf

This work proposes …

A conceptual framework with evidence

- How humans navigate the world, both physically and socially
  - Humans can be seen as avatars in various contexts, representing themselves or others, and this perspective can offer insights into human cognition and behavior.

Interim summary

- An individual’s “having a self” is an individual manifesting the ability to

  (i) construct **a core self**,

  (ii) construct and hold **a model of the environment** that includes its core self in addition to independent temporal, spatial, agential, modal, and other relevant properties in its environment,

  (iii) use this information to build richly structured representations, including **a first-person representation of its embodied self, along with third-person representations of itself,** and align its core self with these representations to “center” itself on its body as an agent in its world, and

  (iv) **orient and reorient itself** by shifting its perceptual and cognitive center to new locations in new spaces in order to reframe and solve new problems.

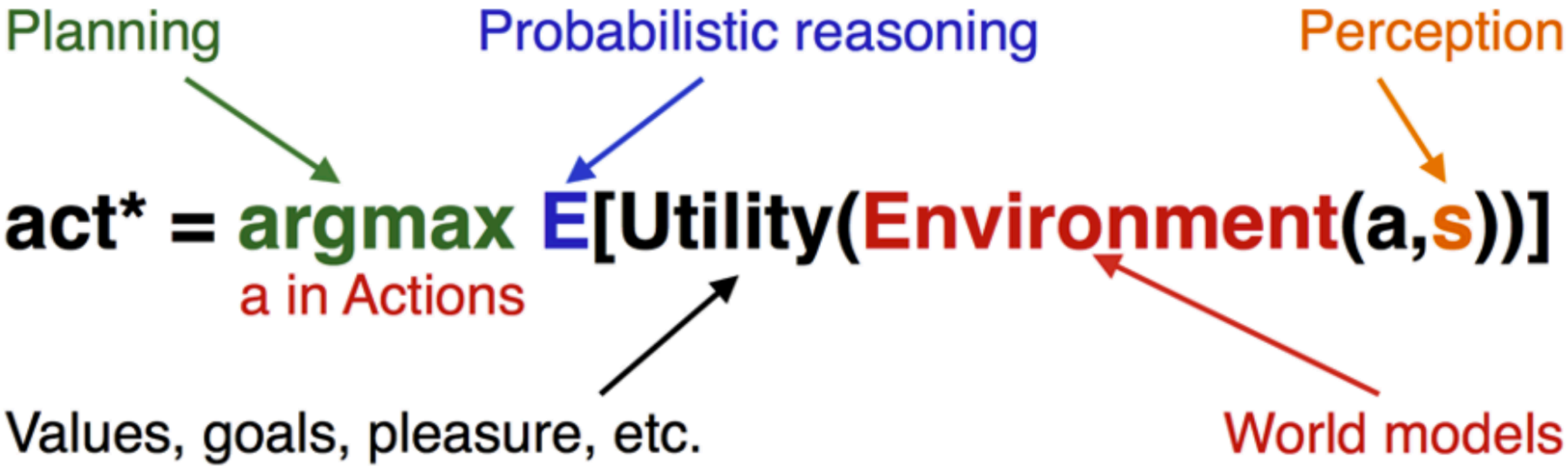
This work proposes …

A computational framework based on POMDPs

- Motivation
  - The need for machines to represent themselves in new possible situations, highlighting the adaptability and flexibility of human intelligence.

This work proposes …

A computational framework based on POMDPs



Planning

Probabilistic reasoning

Perception

act* = argmax E[Utility(Environment(a,s))]

a in Actions

Values, goals, pleasure, etc.

World models

What is POMDPs?

A computational framework based on POMDPs

○ MDP

- State space $S = \{s_1, s_2, \ldots, s_n\}$
- Action space $A = \{a_1, a_2, \ldots, a_n\}$
- Transition model $T(s, a) = P(s' \mid s, a)$
- Utility function $U(s, a) = R(s) - C(a)$

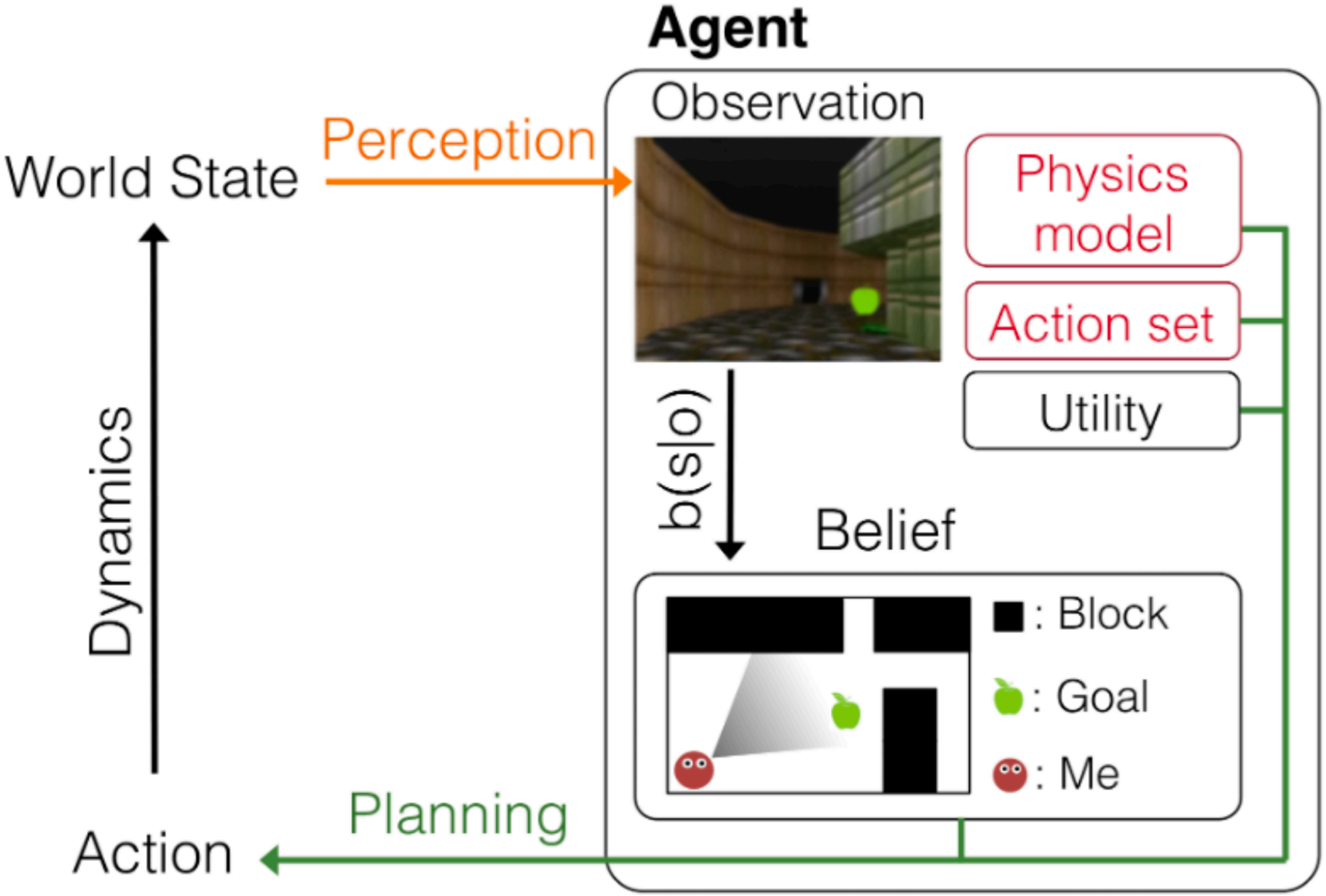
○ POMDPs

- Observation space Z
- Observation function $O(s, a) = P(z \mid s', a) = P(z \mid T(s, a), a)$
- Belief state $b(s)$

$$b(s_{t+1}) = P(s_{t+1} \mid s_t, z, o_{t+1}) \propto O(s_{t+1}, a) * \sum_{s} [T(s_t, a)b(s_t)]$$

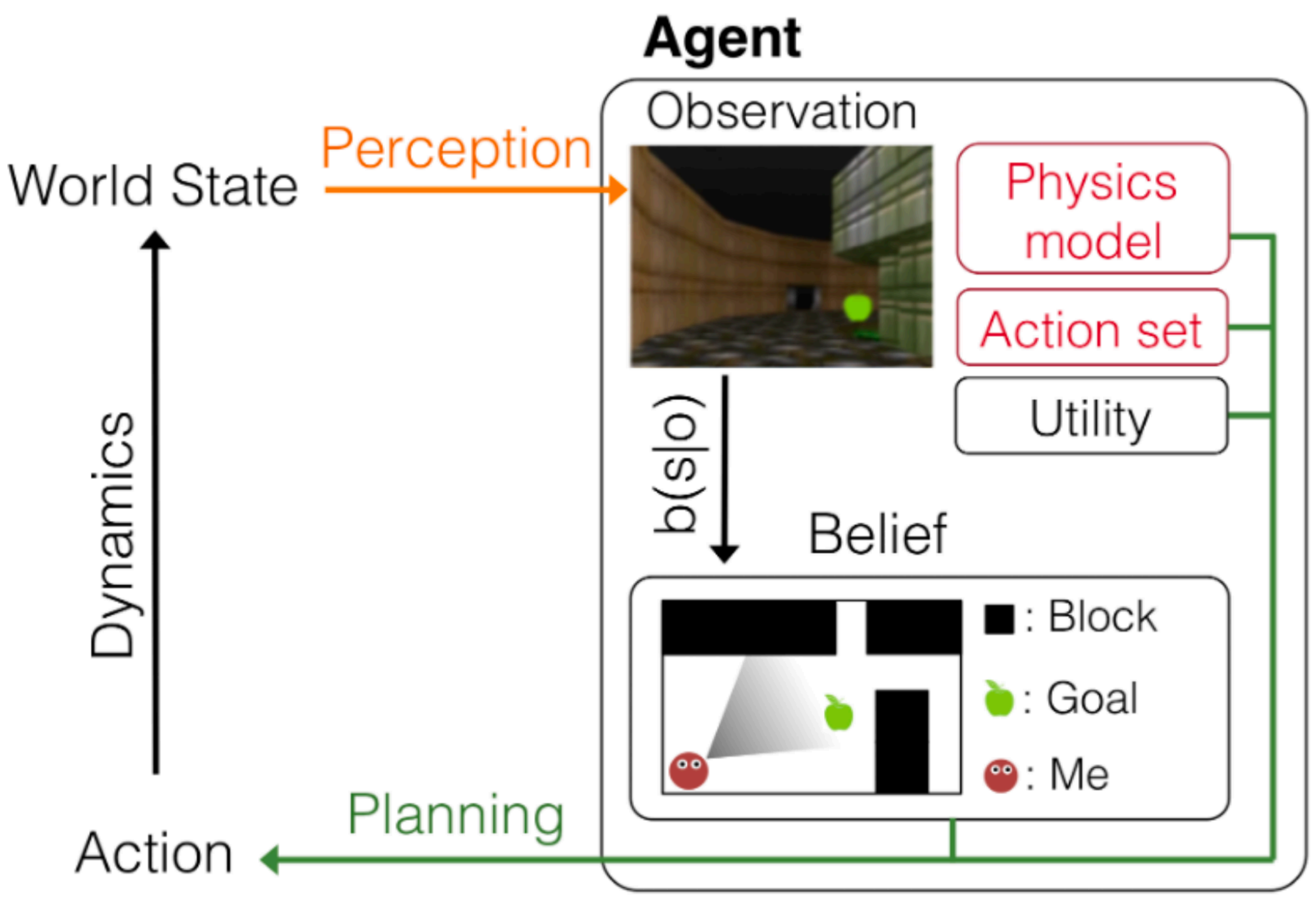
What is POMDPs?

A computational framework based on POMDPs (Cont'd)



This work proposes …

Enriched POMDPs (ePOMDPs)

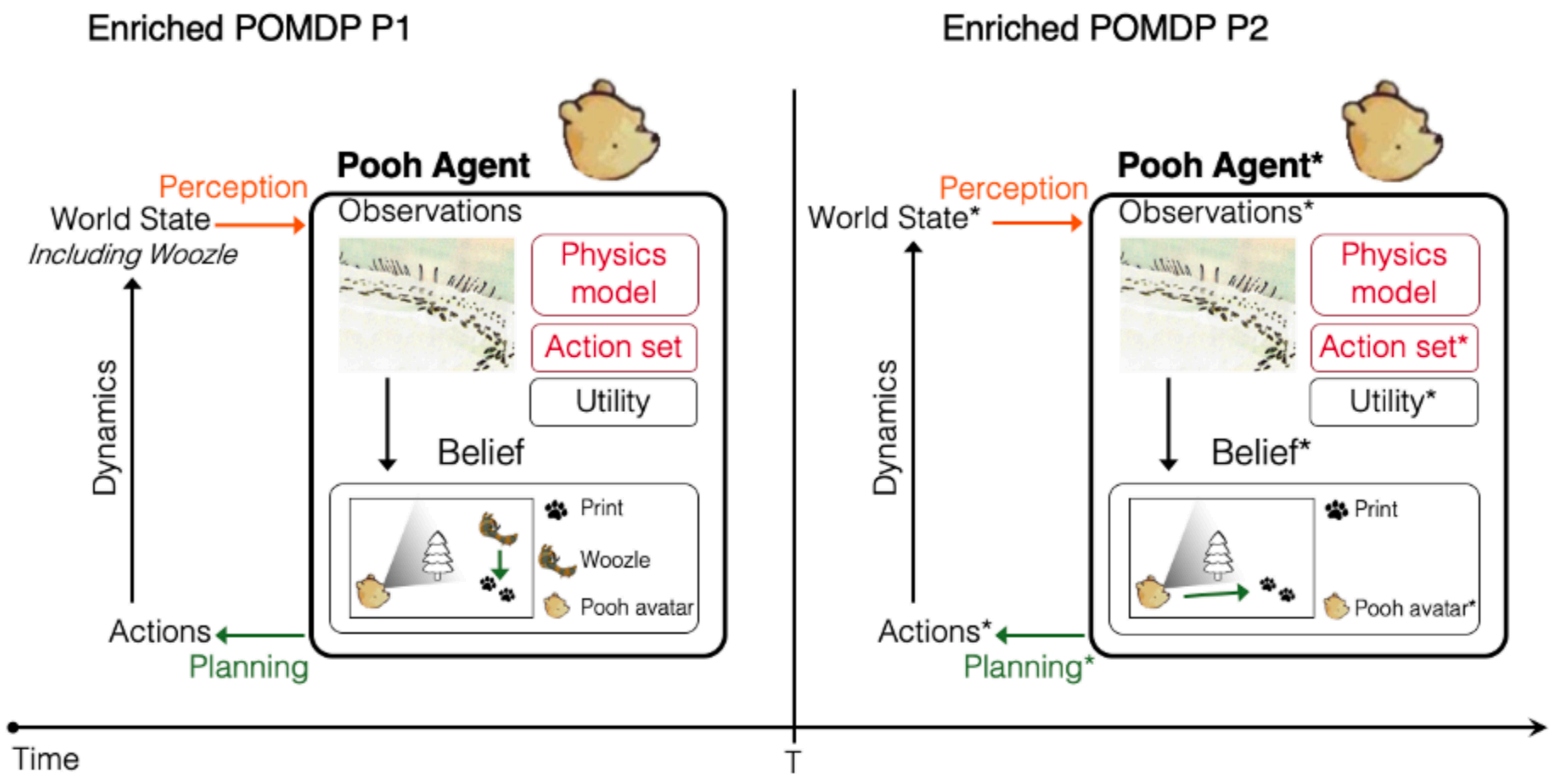


- **Separation between world (W) and agent (M) states:** Which objects are where & Where am I
- **Classes of world states:** An ePOMDP differentiates structure in the world state W by classes of entities that constrain and shape utilities, transitions, observations, and actions.
- **Enriched observation function O:** replace the arbitrary observation function with the process of sensory input (perception)
- **First- and third-person representations**
- **Enriched transition function T:** e.g. physical engine
- **Enriched action sets A:** organized hierarchically
- **Enriched utilities U:** e.g. harm, intention, and morality

This work proposes …

Meta ePOMDPs

- Self-correction and re-centering occur when an agent implicitly realizes the ePOMDP it has been employing no longer matches the world in a substantive way, generating a switch from one ePOMDP to another.



Summary

- The concept of "self" is crucial for creating truly intelligent agents.
- Truly intelligent agents should be able to think for themselves, adapt to new information, and re-center their beliefs.
- The introduced concepts, while promising, add complexities to the traditional POMDP framework, which might make them computationally challenging to implement.