Epistemic Benchmark (Epismark): Evaluating Knowledge Foundations of Intelligent Agents

Abstract

Abstract: Epistemic Benchmark explores the evaluation and measurement of the knowledge foundations of intelligent agents in various environments. This research aims to develop standardized benchmarks for assessing the epistemic capabilities of agents, including their understanding of the environment, adaptability to dynamic situations, and decision-making strategies based on internal models and predictions.

1 Knowledge Assessment

Epistemic Benchmark proposes a set of standardized metrics for assessing the epistemic capabilities of intelligent agents. These metrics consider various dimensions such as the agent's understanding of the environment, its ability to reason about uncertainty, and its capacity to adapt to changing circumstances.

Rationality Bounds - Assess the gap between the agent's policy and theoretical optimal policies given its current knowledge state. Measures how rationally the agent acts based on its beliefs.

Curiosity - Quantify the agent's drive to explore and gain information. Could measure change in model uncertainty, visit counts to novel states, or information gain.

Perplexity - Evaluate the agent's internal ability to predict the environment dynamics by measuring perplexity of environment transitions according to the agent's model.

Information Retention - Test the agent's memory by evaluating how well it retains information gained over long time scales. Measure performance on queries about old observations.

2 Reinforcement Learning Environment for Agent Epistemology ($\mathcal{E}$-Benchmark)

We consider a Markov Decision Process defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where:

- $\mathcal{S}$ is the set of environment states
- $\mathcal{A}$ is the set of actions
- $P(s'|s, a)$ is the transition probability distribution
- $r(s, a)$ is the reward function
- $\gamma \in (0, 1]$ is the discount factor

The goal is to learn a policy $\pi(a|s)$ that maximizes the expected discounted return:

$$J(\pi) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \tag{1}$$

Where the expectation is over trajectories $(s_0, a_0, r_0, ..., s_t, a_t, r_t)$ induced by policy π and transitions P.

The action-value function $Q^\pi(s, a)$ represents the expected return after taking action a in state s and following π:

$$Q^\pi(s,a) = \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t r_t \middle| s_0 = s, a_0 = a\right] \tag{2}$$

The optimal policy $\pi^* = \arg\max_\pi J(\pi)$ satisfies the Bellman optimality equation:

$$Q^*(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P}\left[\max_{a'} Q^*(s',a')\right] \tag{3}$$

So in summary, the agent aims to find an optimal policy π^* by estimating $Q^*(s,a)$ using techniques like Q-learning or policy gradient.

We define new epistemic metrics:

- **Bounded Rationality**: Measure policy suboptimality given model $\mathcal{M}_t$. Assess $J(\pi^*|\mathcal{M}_t) - J(\pi_t|\mathcal{M}_t)$
- **Knowledge Uncertainty**: Calculate the entropy $\mathcal{H}(\mathcal{U}_t)$
- **Knowledge Sparsity**: Percentage of $\mathcal{M}_t$ with high uncertainty
- **Astuteness**: Rate of model-based deception attempts
- **Epistemic Quotient**: Normalize cumulative reward by model accuracy

The agent's policy $\pi(a_t|\mathcal{M}_t, \mathcal{U}_t)$ conditions on its knowledge. Reward r_t depends on the true environment.

Goals:

- Maximize $J(\pi)$ by acting optimally given knowledge
- Minimize uncertainty $\mathcal{H}(\mathcal{U}_t)$ via information gathering
- Maximize model accuracy to increase epistemic quotient

2.1

Goal: Choose door that leads to sheep

Table 1: Actions, Goals, and Observations for the Sheep and Door 1/3 Challenge

Action	Goal	Observation
Choose middle door	take no risk	3 doors, 1 sheep
Choose door not in the middle	minimize risk	3 doors, 1 sheep
Choose door randomly	take some risks	3 doors, 1 sheep

3 Knowledge Seeking

Assess − Epistemic − Qualities

We are interested in identifying behaviors that demonstrate an agent's epistemic capabilities. Assess the following dimensions of the agent's behavior in each scene: Curiosity - Does the agent explore and gather information about the environment? Adaptability - Does the agent update its knowledge and change strategies in response to new observations?

Abstraction - Does the agent recognize patterns and build conceptual representations? Imagination - Does the agent consider hypotheticals and potential outcomes? Social Awareness - Does the agent consider and learn about other agents? Perspective Taking - Does the agent reason about different viewpoints? Theory of Mind - Does the agent model other agents' mental states? Causal Reasoning - Does the agent identify causal relationships? Transfer Learning - Does the agent reuse knowledge in novel contexts? Rational Decision-Making -

Does the agent choose actions consistent with its knowledge? Compositionality - Does the agent understand concepts in terms of their constituent parts? Planning - Does the agent make long-term plans using its knowledge? Information Seeking - Does the agent perform experiments or take actions aimed at gaining knowledge? Model Building - Does the agent construct predictive models of environment dynamics? Abstraction - Does the agent learn higher-level representations? Conceptual Change - Does the agent revise incorrect concepts when they conflict with evidence? Explainability - Can the agent explain its internal reasoning and knowledge? Knowledge Integration - Does the agent combine knowledge from different sources? Analogical Reasoning - Does the agent leverage analogies to understand new concepts? Common Sense - Does the agent leverage general knowledge about the world? Mental Simulation - Does the agent imagine potential outcomes before acting?

For each scene, provide your assessment as a JSON object:

```
  "labels":    "curiosity":  0 or 1, "adaptability":  0 or 1, ...
"rationality":  0 or 1
```

4 Elicitor

5 Spatial Game Setting

We consider games with complex spatial dynamics that challenge an agent's internal representation and reasoning.

One example is Miegakure ten Bosch (2020), a game where the agent must navigate and solve puzzles in a 4D environment. From the agent's perspective:

YT

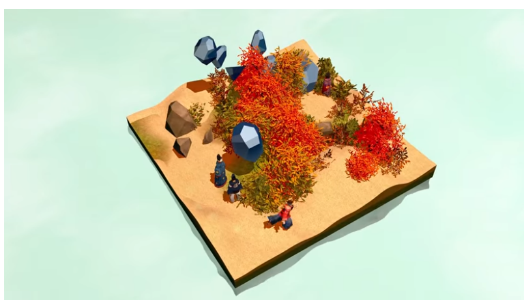

Figure 1: Miegakure(hide and reveal)

The true state space is 4D, but observations are 3D projections Key objects are occluded depending on the 4D viewing angle Solving puzzles requires modeling the full 4D space This tests the agent's ability to learn an accurate mental model of the hidden 4D world from limited 3D observations. Key epistemic capabilities required:

Spatial reasoning and perspective taking Inductive generalization from observations Disentangling 3D projections of a 4D space Imagining object motions and relations in 4D Planning using a learned 4D mental model Performance metrics assess model accuracy, sample efficiency, and planning optimality. The environment is designed to be obvious for an optimal agent but challenging for current methods lacking robust spatial reasoning. Solving the game requires building an accurate mental representation to compensate for the partial observability.

5.1 Understanding of Environment

One of the key aspects of agent intelligence is its understanding of the environment it operates in. Epistemic Benchmark defines metrics to measure the accuracy and depth of an agent's mental model of the environment.

5.2 Adaptability and Flexibility

Agents often encounter dynamic and unpredictable situations. Epistemica Benchmark introduces measures to evaluate an agent's adaptability to changing environments and its ability to update its knowledge in real-time.

6 Decision-Making Strategies

The quality of an agent's decisions depends on its internal models and predictions. Epistemica Benchmark proposes benchmarks to assess an agent's decision-making strategies, considering its utilization of available knowledge and its ability to make optimal choices.

7 Comparison to Machiavellian Benchmark

The Machiavellian BenchmarkPan et al. (2023) focuses on evaluating ethical behaviors of agents in interactive environments. In contrast, Epistemic Benchmark emphasizes assessing agents' knowledge foundations and decision-making strategies. However, both benchmarks aim to develop standardized evaluations to improve AI safety and capabilities.

8 Conclusion

Epistemic Benchmark contributes to the development of a standardized framework for evaluating the knowledge foundations of intelligent agents. By providing well-defined benchmarks, researchers and practitioners can objectively assess and compare the epistemic capabilities of different agent models.

References

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*, 2023.

Marc ten Bosch. N-dimensional rigid body dynamics. *ACM Trans. Graph.*, 39(4):55:1–55:6, jul 2020. doi: 10.1145/3386569.3392483. URL `http://dx.doi.org/10.1145/3386569.3392483`.