AGENT: A BENCHMARK FOR CORE PSYCHOLOGICAL REASONING

Tianmin Shu
MIT

Abhishek Bhandwaldar
MIT-IBM Watson AI Lab

Chuang Gan
MIT-IBM Watson AI Lab

Kevin A. Smith
MIT

Shari Liu
MIT

Dan Gutfreund
MIT-IBM Watson AI Lab

Elizabeth Spelke
Harvard University

Joshua B. Tenenbaum
MIT

Tomer D. Ullman
Harvard University

ABSTRACT

For machine agents to successfully interact with humans in real-world settings, they will need to develop an understanding of human mental life. Intuitive psychology, the ability to reason about hidden mental variables that drive observable actions, comes naturally to people: even pre-verbal infants can tell agents from objects, expecting agents to act efficiently to achieve goals given constraints. Despite recent interest in machine agents that reason about other agents, it is not clear if such agents learn or hold the core psychology principles that drive human reasoning. Inspired by cognitive development studies on intuitive psychology, we present a benchmark consisting of a large dataset of procedurally generated 3D animations, AGENT (Action, Goal, Efficiency, coNstraint, uTility), structured around four scenarios (goal preferences, action efficiency, unobserved constraints, and cost-reward trade-offs) that probe key concepts of core intuitive psychology. We validate AGENT with human-ratings, propose an evaluation protocol emphasizing generalization, and compare two strong baselines built on Bayesian inverse planning and a Theory of Mind neural network. Our results suggest that to pass the designed tests of core intuitive psychology at human levels, a model must acquire or have built-in representations of how agents plan, combining utility computations and core knowledge of objects and physics.[1]

[1] The dataset is available at `https://www.tshu.io/AGENT`.

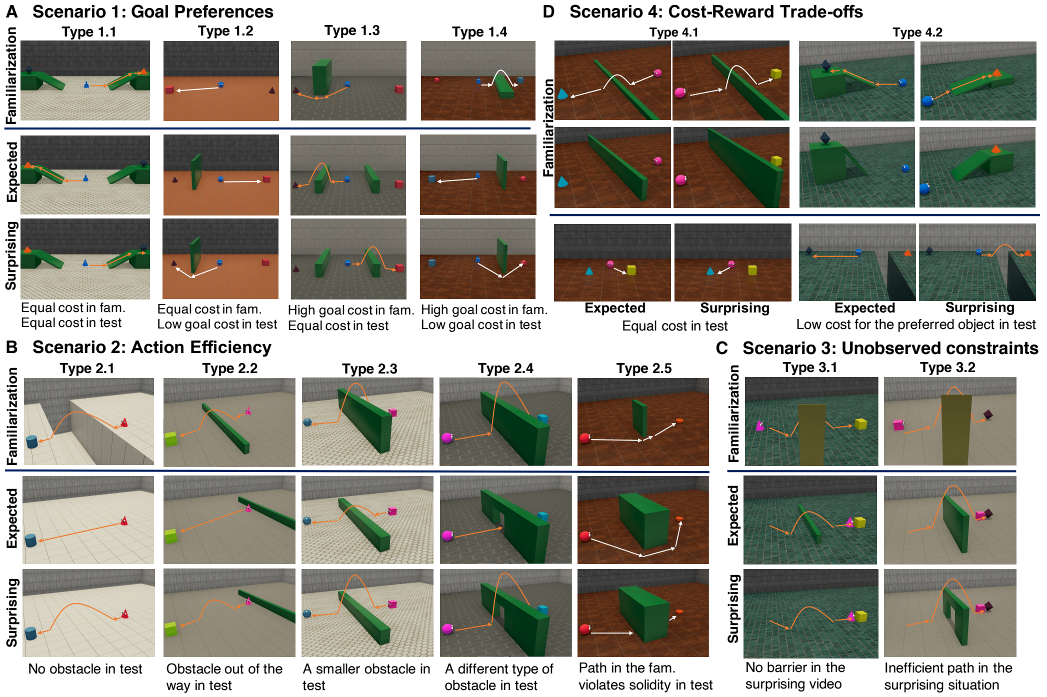
A Scenario 1: Goal Preferences
Type 1.1
Type 1.2
Type 1.3
Type 1.4
Familiarization
Expected
Surprising
Equal cost in fam. Equal cost in test
Equal cost in fam. Low goal cost in test
High goal cost in fam. Equal cost in test
High goal cost in fam. Low goal cost in test
D Scenario 4: Cost-Reward Trade-offs
Type 4.1
Type 4.2
Familiarization
Expected
Surprising
Equal cost in test
Expected
Surprising
Low cost for the preferred object in test
B Scenario 2: Action Efficiency
Type 2.1
Type 2.2
Type 2.3
Type 2.4
Type 2.5
Familiarization
Expected
Surprising
No obstacle in test
Obstacle out of the way in test
A smaller obstacle in test
A different type of obstacle in test
Path in the fam. violates solidity in test
C Scenario 3: Unobserved constraints
Type 3.1
Type 3.2
Familiarization
Expected
Surprising
No barrier in the surprising video
Inefficient path in the surprising situation

Figure 1: Overview of trial types of four scenarios in AGENT. Each scenario is inspired by infant cognition and meant to test a different facet of intuitive psychology. Each type controls for the possibility of learning simpler heuristics. Example trials are available at `https://www.tshu.io/AGENT`.