AGENT: A Benchmark for Core Psychological Reasoning

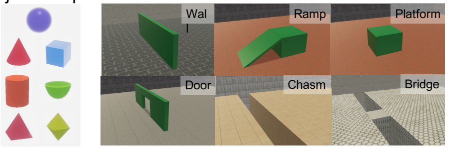
Tianmin Shu[1] Abhishek Bhandwaldar[2] Chuang Gan[2] Kevin A. Smith[1] Shari Liu[1]
Dan Gutfreund[2] Elizabeth Spelke[3] Joshua B. Tenenbaum[1] Tomer D. Ullman[3]
[1]MIT [2]MIT-IBM Watson AI Lab [3]Harvard Website: https://www.tshu.io/AGENT

Introduction

- Intuitive psychology, the ability to reason about hidden mental variables that drive observable actions, comes naturally to people.
- Despite recent interest in machine agents that reason about other agents, it is unclear if such agents learn or hold core psychological principles that drive human reasoning. states and are willing to be trained towards the goal.
- Inspired by cognitive development studies on intuitive psychology, we present a benchmark consisting of a large dataset of procedurally generated 3D animations, **AGENT** (**A**ction, **G**oal, **E**fficiency, co**N**straint, u**T**ility), structured around four scenarios (see the figure on the right).

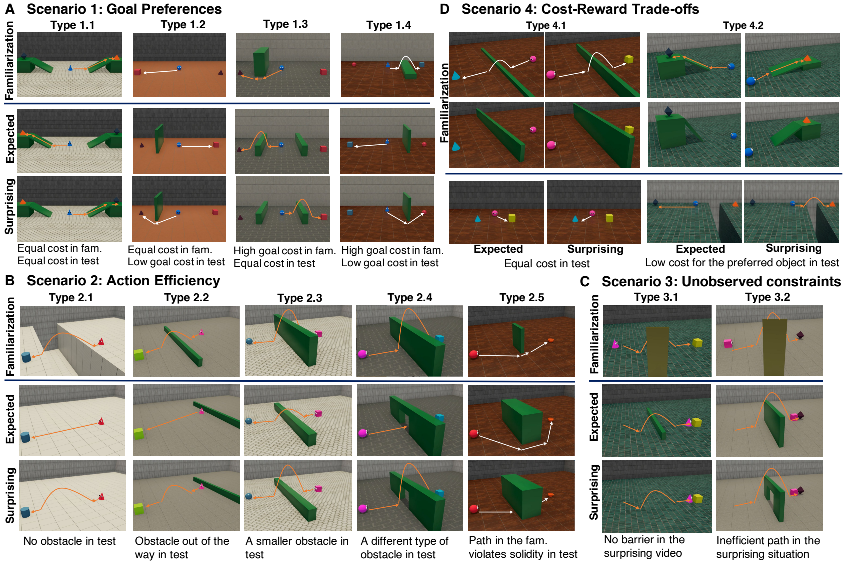
Dataset Structure and Evaluation

- 9240 videos synthesized in ThreeDWorld (TDW).
- 3360 trials in total, divided into 1920 training trials, 480 validation trials, and 960 testing trials. All training and validation trials only contain expected test videos.
- We provide RGB-D frames, instance segmentation, camera parameters, and ground-truth 3D states.
- 7 object shapes and 6 types of obstacles:



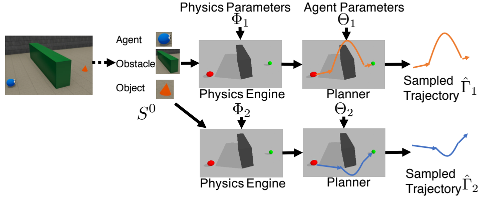
- Following Riochet et al. (2018), we define a metric based on relative surprise ratings. For a paired set of N_+ surprising test videos and N_- expected test videos (which share the same familiarization video(s)), we obtain two sets of surprise ratings, $\{r_i^+\}_{i=1}^{N_+}$ and $\{r_j^-\}_{j=1}^{N_-}$ respectively. Accuracy is then defined as the percentage of the correctly ordered pairs of ratings: $\frac{1}{N_+N_-}\sum_{i,j}\mathbf{1}(r_i^+ > r_j^+)$.

Overview of Trial Types of Four Scenarios in AGENT

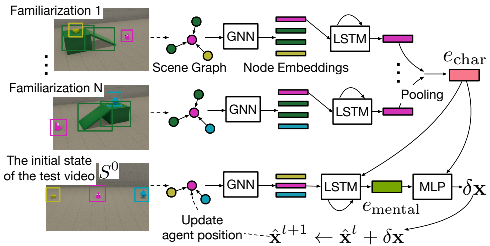


Baselines

BIPaCK: Bayesian Inverse Planning and Core Knowledge



ToMnet-G: Theory of Mind Neural Network with Graphs



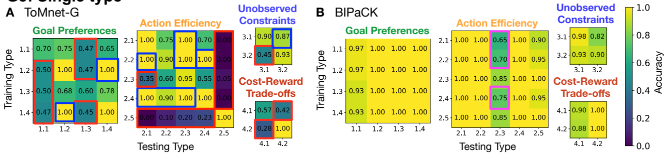
Experimental Results

All: Trained on all types and scenarios; G1: Leave one type out; G2: leave one scenario out

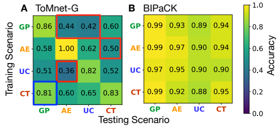
Condition	Method	Goal Preferences					Action Efficiency						Unobs.			Cost-Reward			All
		1.1	1.2	1.3	1.4	All	2.1	2.2	2.3	2.4	2.5	All	3.1	3.2	All	4.1	4.2	All	
	Human	.95	.95	.92	.97	.95	.87	.93	.86	.95	.94	.91	.88	.94	.92	.82	.91	.87	.91
All	ToMnet-G	.73	1.0	.53	1.0	.84	.95	1.0	.95	.88	1.0	.94	.95	.78	.85	.63	1.0	.82	.86
All	BIPaCK	.97	1.0	1.0	1.0	.99	1.0	1.0	.85	1.0	1.0	.97	.93	.88	.90	.90	1.0	.95	.96
G1	ToMnet-G	.63	.95	.53	1.0	.81	.95	.80	.45	.77	.05	.63	.45	.87	.70	.28	.42	.35	.63
G1	BIPaCK	.93	1.0	1.0	1.0	.98	1.0	1.0	.80	1.0	1.0	.97	.93	.82	.86	.88	1.0	.94	.94
G2	ToMnet-G	.50	.93	.50	.88	.73	.70	.60	.75	.75	1.0	.76	.60	.73	.68	.62	.98	.80	.74
G2	BIPaCK	.93	1.0	1.0	1.0	.98	1.0	1.0	.75	1.0	.95	.95	.88	.85	.87	.83	1.0	.92	.94

Red: poor generalization (no better than chance); **Blue**: good generalization; **Magenta**: Failures of BIPaCK

G3: Single type



G4: Single scenario



Dataset is available