Modeling Surprise in a Physically Grounded Joint Inference of Preference, Knowledge, and Perceptual Access

Harry Chen (harryc@mit.edu)
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology

Karen Chung (seoyeon@mit.edu)
Department of EECS, Massachusetts Institute of Technology

Abhishek Bhandwaldar (Abhi.B@ibm.com)
MIT-IBM Watson AI Lab

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology

Tomer D. Ullman (tullman@fas.harvard.edu)
Department of Psychology, Harvard University

Tianmin Shu (tianmin.shu@jhu.edu)
Department of Computer Science & Department of Cognitive Science, Johns Hopkins University

Abstract

Even infants understand other agents can have partial observability of the world, and show varying degrees of uncertainty about the knowledge and preferences of others. This work models people's inference of another agent's preference and knowledge given limited perceptual access, as measured by their surprise response. We propose POPKI (Physically-grounded Observation, Preference, and Knowledge Inference), a Bayesian inverse-planning method that models graded surprise in the inference of preference, knowledge, and perceptual access in rich 3D environments. To test our model, we extended the AGENT dataset to trials that probe preference, knowledge, and perceptual access. Experimental results show POPKI replicates humans' varying degrees of surprise when judging the behavior of agents under different visibility conditions. These results suggest that reasoning about how agents plan in imagined physical states according to their knowledge under limited observability is key to reverse-engineering human-like uncertainty judgments in psychological reasoning tasks.

Keywords: intuitive psychology; perceptual access; knowledge; Bayesian inverse planning; violation of expectation

Introduction

A deep understanding of human mental states is essential for effectively integrating machine agents into human-centric environments. Previous works have investigated what type of machine models can reverse-engineer human-level Theory of Mind (ToM). Most notably, recent ToM benchmarks such as the Baby Intuitions Benchmark (Gandhi et al., 2022) and AGENT (Shu et al., 2021) offer a wide set of tasks that probe how much a model grasps fundamental ToM principles, based on the influential Violation of Expectation paradigm used in developmental studies (e.g., Woodward, 1998; Csibra et al., 2003; Liu et al., 2017). Prior research on modeling surprise in these developmentally inspired benchmarks has focused on situations where agents have complete visibility of their surroundings. However, when agents lack full observability, human surprise judgments reflect graded uncertainty (Luo & Baillargeon, 2008). It remains unclear how we can model such graded surprise judgments.

We propose POPKI (Physically-grounded Observation, Preference, and Knowledge Inference), a Bayesian inverse-planning approach that models graded surprise in rich 3D environments by jointly inferring preferences, knowledge, and perceptual access. Our experimental results show POPKI replicates the varying degrees of human surprise when observing the behavior of agents under different conditions of perceptual access, unlike a strong neural network baseline.

AGENT+ with Perceptual Access Trials

The original AGENT benchmark (Shu et al., 2021) only depicted agents with full perceptual access. Following Luo & Baillargeon (2008), we expanded AGENT to create AGENT+ with three new trial types that test limited perceptual access (Figure 1). Each trial begins with a familiarization video that displays an agent's preference for a particular object, followed by a test video in which the agent either chooses the same object as before or a different one. We refer to the former trials as 'expected' and the latter as 'surprising.'

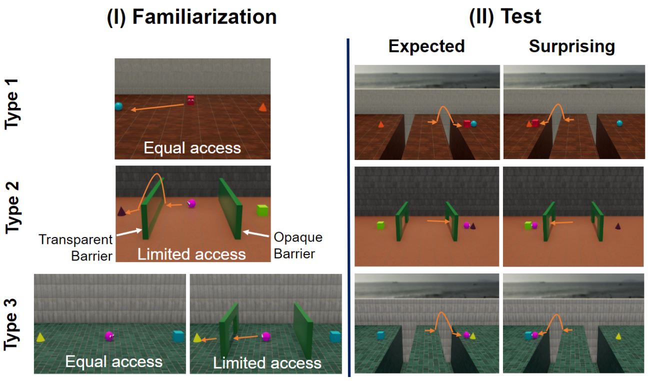

Figure 1: Overview of the perceptual access trials. One of the barriers in the Type 2 familiarization video is transparent. We show two frames for the Type 3 familiarization video because the agent first looks at both objects (equal access) before the barriers drop (limited access).

In Type 1 stimuli, the agent sees both objects during familiarization, giving it full knowledge of their locations. In Type 2, the agent only sees one object, limiting its knowledge. In such a case, the agent may be locally selecting a globally dispreferred object, since it does not know another one exists. In Type 3, the agent initially sees both objects, but one becomes obscured by barriers, altering its perceptual access but not its knowledge. In all test videos across stimuli types, the agent can see both objects.

Computational Model

POPKI jointly infers preference, knowledge, and perceptual access as shown in Figure 2. The POPKI model is built in the probabilistic programming language Gen (Cusumano-Towner et al., 2019). It leverages PyBullet as its physics engine, and RRT* (Karaman & Frazzoli, 2011) as its planner. Unlike the Bayesian inverse-planning model used to solve the tasks in the original AGENT benchmark, POPKI introduces partial observation $o^t \sim O(o|s^t)$ given a state s^t at time t, and knowledge inference $k^t(g) \in 0,1$ of a goal object $g \in \mathcal{G}$ ($\mathcal{G}$ is the set of goal objects). $k^t(g) = 1$ means the agent knows g exists.

Given the familiarization video, represented by a trajectory of states s^t at each step ($\Gamma_{fam} = s^{1:T}$), we conduct a simultaneous inference of knowledge k^t, perceptual access o^t, and agent parameters $\Theta = (R, C)$ that include the reward R and the cost C. We also infer the physics parameters Φ that define the physical characteristics of the physics engine. The

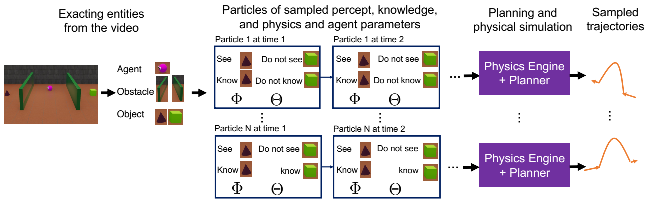

Figure 2: Overview of the POPKI model. We approximate the physical scene in a physics engine using extracted object entities (with ground-truth annotations). For each particle hypothesis, we sample an agent trajectory conditioned on that hypothesis.

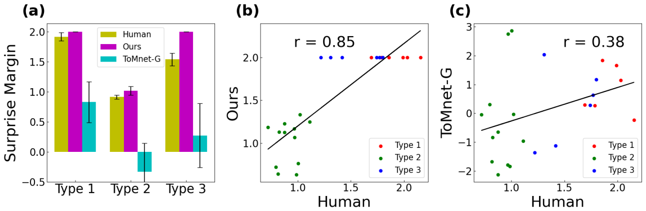

Figure 3: Surprise margin comparisons between humans and models. (**a**) Averaged surprise margin for each type (error bars show standard errors). (**b**) Correlation between humans and ours. (**c**) Correlation between humans and ToMnet-G.

inference is formulated as:

$$P(\Phi,\Theta|\Gamma_{\text{fam}}) \propto \sum_{k^{0:T},o^{1:T}} P(s^{1:T}|k^{0:T},R,C,\Phi)P(k^{1:T}|o^{1:T},k^0) \cdot P(o^{1:T}|s^{1:T})P(k^0)P(R)P(C). \quad (1)$$

In order to compute the likelihood of a trajectory given a specific hypothesis, we use a two-step process. First, we simulate the trajectory based on the imagined environment, constructed from the current assumed knowledge. Then, at each step of this trajectory, we assess the likelihood using a Gaussian distribution. In this distribution, the mean corresponds to the trajectory's coordinate at that specific step, while the standard deviation—a constant manually set to match the data—accounts for minor deviations from the projected path.

We use Sequential Monte Carlo to update the knowledge, perceptual access, and preference inference at each step. We maintain a set of particles that each contain a hypothesis of knowledge, perceptual access, and the physics and agent parameters. We use uniform priors for R and C, and a Bernoulli prior for k^0 with the P(knowing an unobserved goal) at 0.5.

For each test video, we measure surprise using the expected log-likelihood of observing the agent trajectory in the test environment with respect to the posterior distribution of Φ and Θ inferred from the familiarization video: $E_{\Phi,\Theta}[\log P(\Gamma_{\text{test}}|\Phi,\Theta)]$, where

$$P(\Gamma_{\text{test}}|\Phi,\Theta) = \sum_{k^{0:T},o^{1:T}} P(s^{1:T}|k^{0:T},R,C,\Phi) \cdot P(k^{1:T}|o^{1:T},k^0)P(o^{1:T}|s^{1:T}). \quad (2)$$

Results

Human Experiment We recruited 84 participants (mean age = 39.8; 65 female) on Prolific to judge 48 trials (24 pairs of surprising and expected trials). People rated how 'surprising' an agent's behavior was in a test video on a scale from 0 to 100. Each trial received ratings from 10 participants. The study was approved by an institutional review board.

Baseline We adopt ToMnet (Rabinowitz et al., 2018) as a baseline, specifically the extended model ToMnet-G (Shu et al., 2021), which employs a graph-NN to encode states of entities that appear in the videos. ToMnet-G is trained on a dataset of 360 trials from the new perceptual access scenario.

Our Model Prior work (Gandhi et al., 2022; Shu et al., 2021; Zhi-Xuan et al., 2022) only evaluated whether the surprising test video receives a higher surprise rating than its expected counterpart. Here, we further examine the surprise margin as an indicator of uncertainty. Formally, we define r^+ and r^- to be the surprise ratings for the paired surprising and expected test videos, respectively. The surprise margin is calculated as $r^+ - r^-$. Note that surprise ratings were standardized first, following prior works (e.g., Smith et al., 2019; Shu et al., 2021). We standardize our model scores between pairs of surprising and expected trials, ensuring that the surprise margin reflects the surprise rating within trials without adjusting for other surprise margins from other trials of the same type. This calibration method explains the consistent surprise margin of approximately 2.0 for trial types 1 and 3 observed in our model in Figure 3**(b)**. This approach does not compromise the robustness of our model's outcomes.

Figure 3**(a)** shows that the average surprise margin, based on human ratings, is significantly lower for Type 2 compared to Type 1 and Type 3, consistent with findings in Luo & Baillargeon (2008). Our POPKI model's surprise margins show similar trends. By contrast, ToMnet-G's surprise margins for the three types do not exhibit a weaker surprise in Type 2 trials. In Figure 3**(bc)**, we show the surprise margin for each paired trial. The correlation between our model's surprise judgments and human ratings is 0.85 ($p < 0.001$), markedly higher than ToMnet-G's correlation with human ratings (0.38; $p = 0.07$). Notably, while ToMnet-G succeeded on the original AGENT benchmark, it rated the expected test video as more surprising in 41.7% of all paired trials, contrary to humans and POPKI , which judged it less surprising on average.

Conclusion

In this work, we expanded the AGENT benchmark by introducing new trials that evaluate a model's ability to reason about goal-directed behavior under limited perceptual access. To handle these trials, we develop a new model (POPKI) that sequentially models an agent's knowledge and perceptual state, combined with a physically grounded generative model of the agent's behavior. Our experimental results demonstrate that POPKI can successfully account for the varying degrees of surprise that humans exhibit when judging an agent's actions under different conditions of perceptual access.

References

Csibra, G., Bíró, Z., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cogn. Sci.*, *27*(1), 111–133.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., & Mansinghka, V. K. (2019). Gen: a general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th acm sigplan conference on programming language design and implementation* (pp. 221–236).

Gandhi, K., Stojnic, G., Lake, B. M., & Dillon, M. R. (2022). *Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others.*

Karaman, S., & Frazzoli, E. (2011). Incremental sampling-based algorithms for optimal motion planning.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017, November). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Luo, Y., & Baillargeon, R. (2008). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*, 489–512. doi: 10.1016/j.cognition.2006.10.007

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227).

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., ... Ullman, T. D. (2021). Agent: A benchmark for core psychological reasoning. In *Proceedings of the 38th international conference on machine learning.*

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, *32*.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022). Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. *arXiv preprint arXiv:2208.02914*.