Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others

Kanishk Gandhi[1], Gala Stojnic[2], Brenden M. Lake[1,2], Moira R. DIllon[2]

[1]Center for Data Science, New York University; [2]Department of Psychology, New York University

Introduction

- Human infants intuitively make rich inferences about the goals and preferences underlying others agents' actions [1-5].
- Machines, in contrast, are often trained to predict only agents' actions, not the intentions that underlie those actions.
- This impoverished "Machine Theory of Mind" may be a critical difference between human and machine intelligence [6].
- Recent computational work has aimed to address this difference, but this work has not been tested on a comprehensive benchmark that captures the rich and abstract nature of humans' intuitions about agents [7, 8].

Objectives

- Introduce a comprehensive benchmark for evaluating the common-sense understanding of the goals, preferences, and actions of other agents, appropriate for testing machines and humans alike
- Evaluating state-of-the-art computational models as baselines on this benchmark
- Piloting the benchmark with human infants as a first step towards its validation
- Comparing machine and human performance

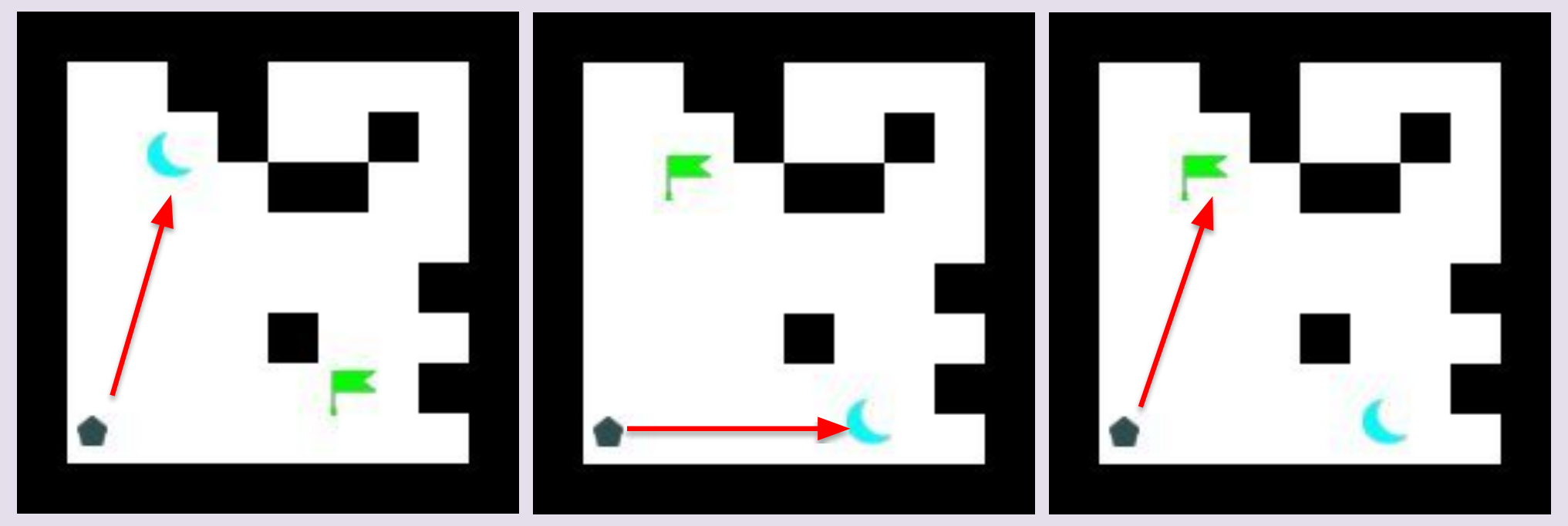
Baby Intuitions Benchmark (BIB)

- **BIB presents a suite of tasks evaluating core components of infants' rich intuitive reasoning about agents**:

1) **Preference Task:** Agents have preferences for specific objects
2) **Multiagent Task:** Agents have preferences that may not generalize to other agents
3) **Inaccessible Goal Task:** Obstacles restrict agents' actions and agents might move to a nonpreferred object when their preferred object is inaccessible
4) **Instrumental Task:** An agent's sequence of actions may be directed towards a higher-order goal
5) **Efficiency Task:** Rational agents act efficiently towards their goals.

- **BIB adopts the "Violation of Expectation" (VOE) paradigm:**
- A succession of 8 familiarization trials **(FAM phase)** introduces the main elements of the displays and allows the observer to form expectations
- To test trials (**TEST phase**) present an expected and unexpected outcome
- The *expected* outcome is typically perceptually dissimilar to the events in the familiarization while the *unexpected* outcome is typically more perceptually similar
- **Uses "Grid-World" environment:**
- Presented from an overhead perspective with simple shapes
- Fully observable to the agent and viewer
- Particularly suitable for testing AIs (e.g., easy procedural generation) [8]

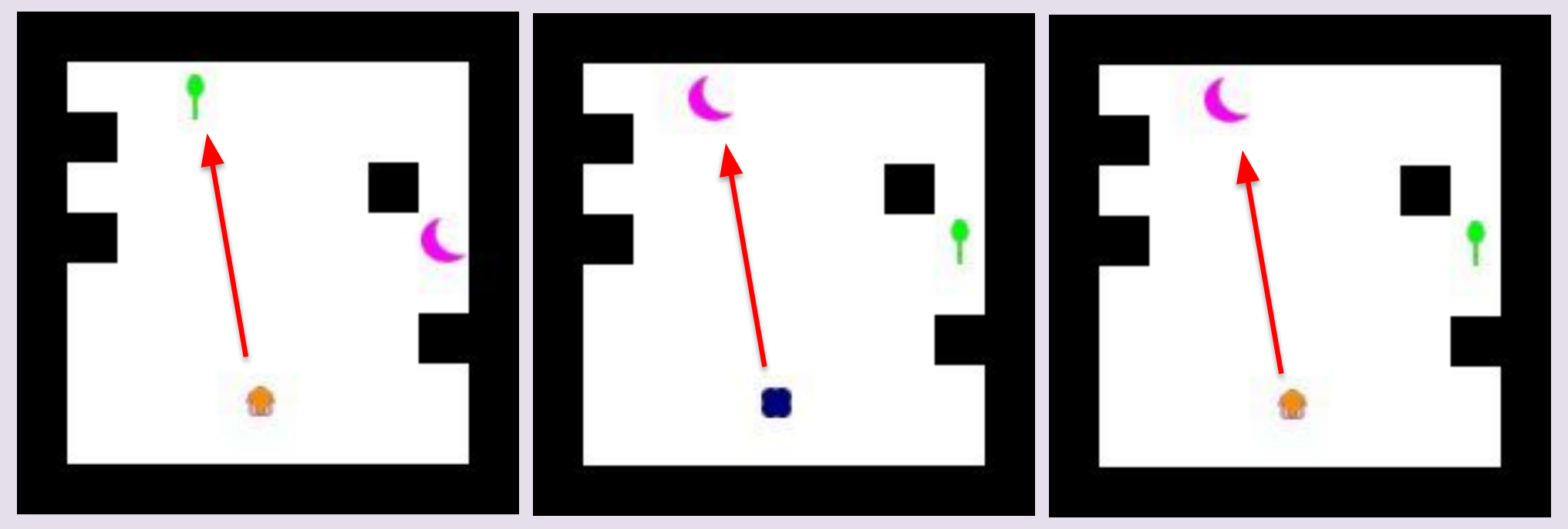
Preference



FAM (x8) Expected Unexpected

FAM: The agent repeatedly approaches the same object at (approximately) same location.
TEST: The agent approaches the preferred object at the new location (expected) or the nonpreferred object at the old location (unexpected).

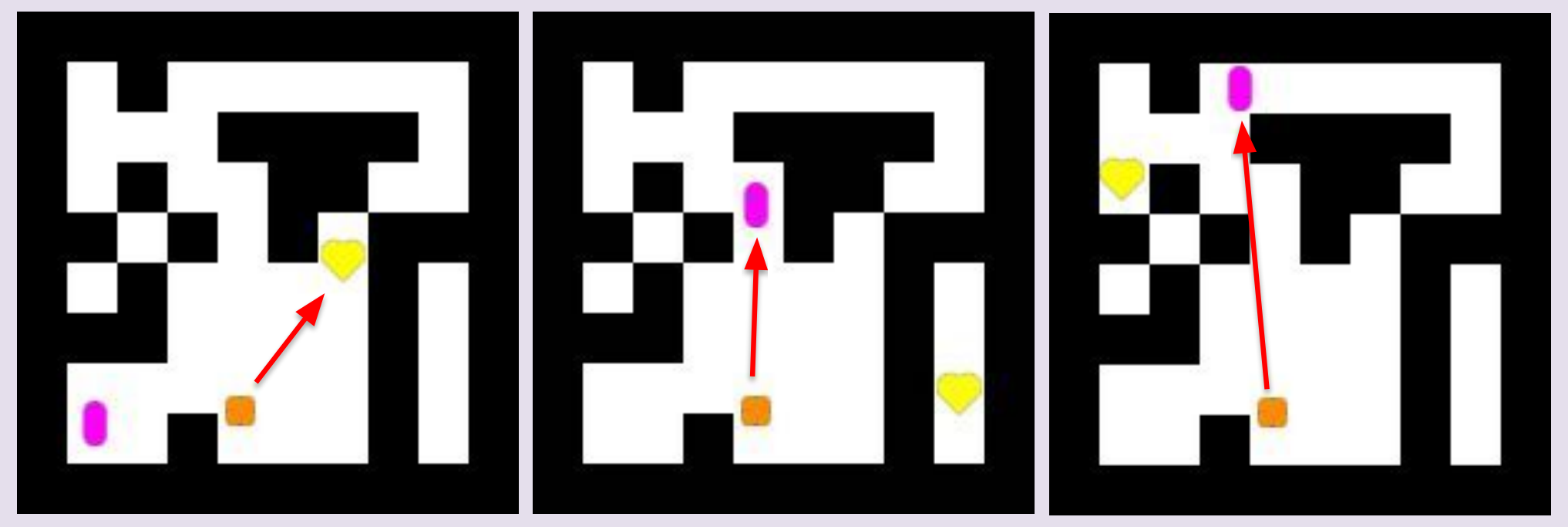
Multiagent Task



FAM (x8) No Expectation Unexpected

FAM: The agent repeatedly approaches the same object, which appears at varied locations.
TEST: A new agent approaches the nonpreferred object (no expectation) or the same agent approaches the nonpreferred object (unexpected).

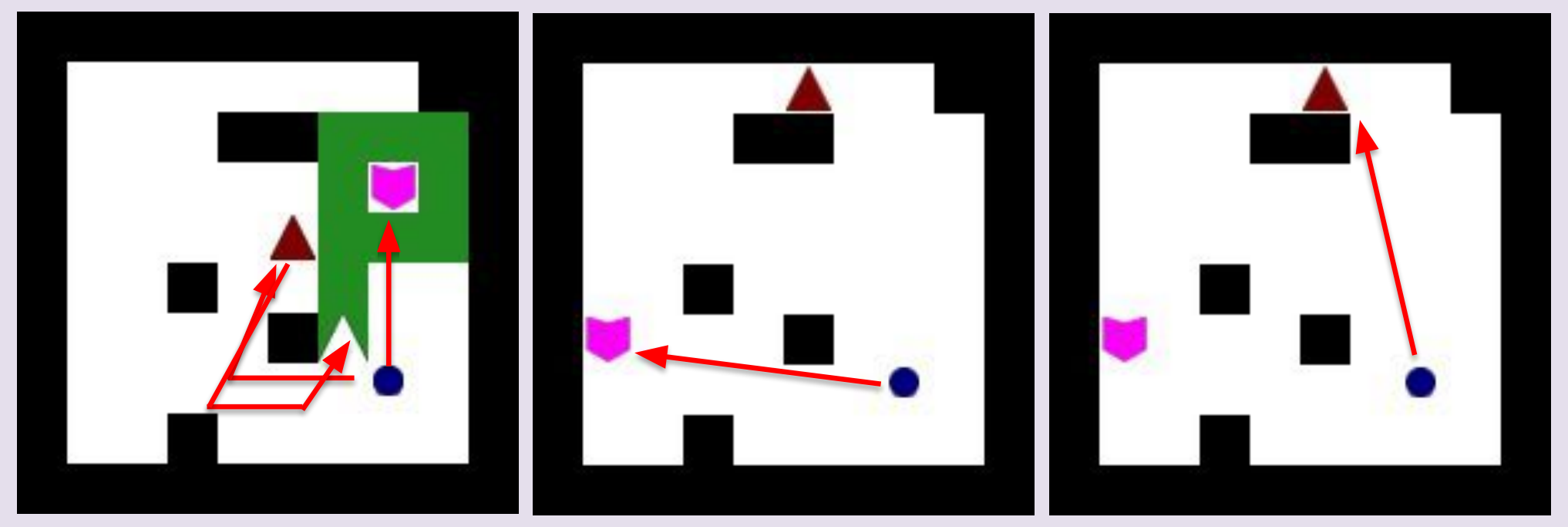
Inaccessible Goal Task



FAM (x8) Expected Unexpected

FAM: The agent repeatedly approaches the same object, which appears at varied locations.
TEST: The agent approaches the nonpreferred object when the preferred object is blocked (expected) or not blocked (unexpected).

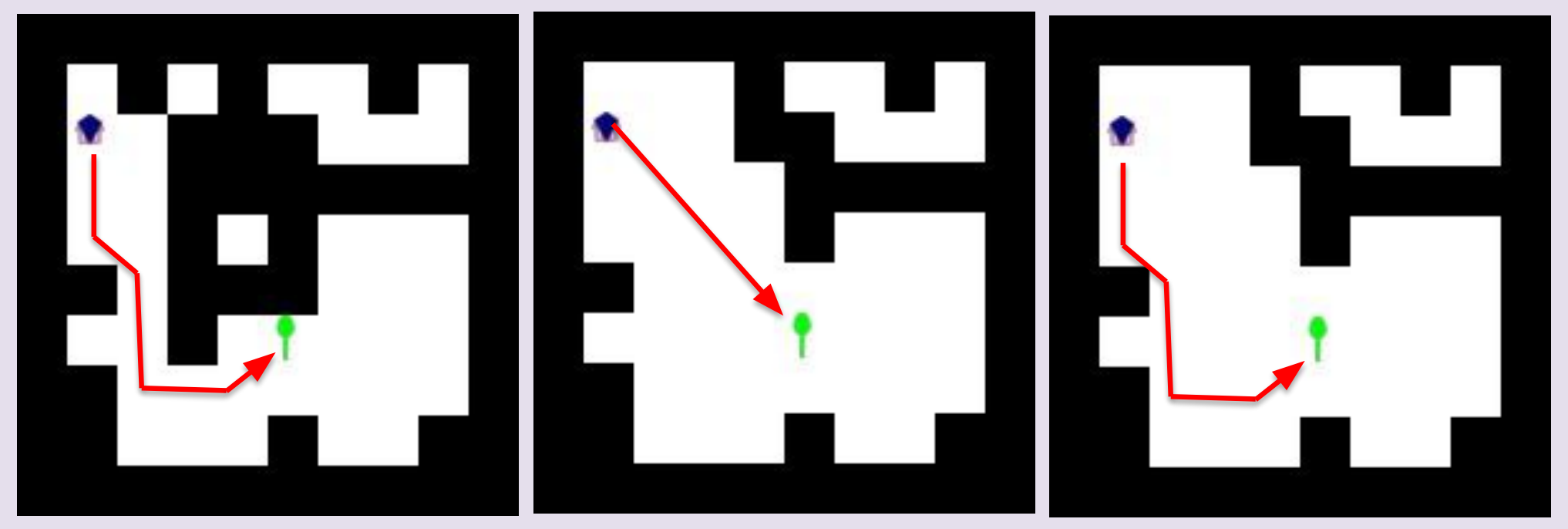
Instrumental Task



FAM (x8) Expected Unexpected

FAM: The agent retrieves the triangular key, inserts it into the green barrier, the barrier disappears, and the agent moves to the object.
TEST:
- NO BARRIER: There is no green barrier present. The agent moves to the object (expected) or to the key (unexpected).
- INCONSEQUENTIAL BARRIER: There is a barrier present, but it does not block the object. The agent moves to the object (expected) or to the key (unexpected).
- CONSEQUENTIAL BARRIER: The barrier blocks the object and the agent moves to the key (expected) or the barrier does not block the object and the agent moves to the key (unexpected).

Efficiency Task



FAM (x8) Expected Unexpected

FAM:
- RATIONAL: The agent moves around obstacles to its goal.
- IRRATIONAL: The agent moves along the same paths as the rational agent, but there are no obstacles.

TEST: There are no obstacles between the agent and its goal.
- PATH CONTROL: The distance between agent and goal is the matched. The agent takes the efficient, straight-line path (expected) or a curved path from familiarization (unexpected).
- TIME CONTROL: The agent takes the efficient, straight-line path (expected) or a curved path (unexpected), but the goal is closer to the agent and so the travel time is matched across trials.

Baseline Models' Performance

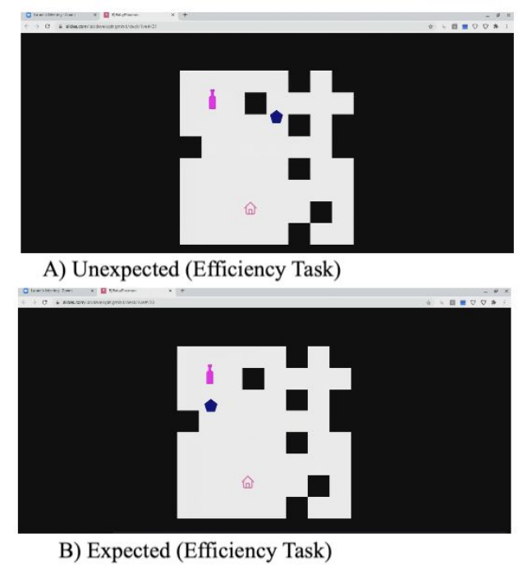
- Background training tasks are provided for the models to learn about the grid worlds, their elements, and the structure of the trials (thousands of episodes). These are systematically different from the evaluation tasks.
- Baseline models are variants of the Theory of Mind Network (ToMNet) [7], an approach to machine agency reasoning trained solely through observation:
  1) **Behavior Cloning (BC) model** operates on the actions taken by the agent and tries to model the demonstrated policy;
  2) **Video model** operates directly on the frames of a video and is trained to predict the video's next frame.
- We encode the familiarization trials as context either using a bidirectional RNN or an MLP.

BIB Agency Task	BC-MLP	BC-RNN	Video-RNN
Preference	26.3	48.3	47.6
Multi-Agent	48.7	48.2	50.3
Inaccessible Goal	53.1	46.6	66.0
Efficiency: Path control	96.0	95.8	99.8
Efficiency: Time control	99.1	99.1	99.9
Efficiency: Irrational agent	73.4	48.8	50.0
Efficient Action Average	85.5	73.1	74.9
Instrumental: No barrier	98.8	98.8	99.7
Instrumental: Inconsequential barrier	56.7	78.2	76.7
Instrumental: Blocking barrier	48.2	55.9	58.2
Instrumental Action Average	67.9	77.6	78.2

The scores quantify pairwise VOE judgements. The expectedness of a test trial is defined by its error on the most "unexpected" video frame (frame with the highest prediction error).

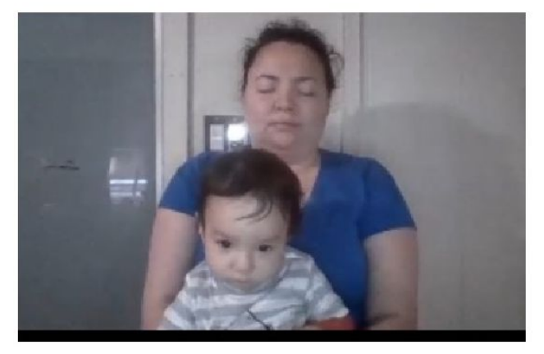
Pilot Validation with Infants: Preference Task and Efficiency Task (Path Control)

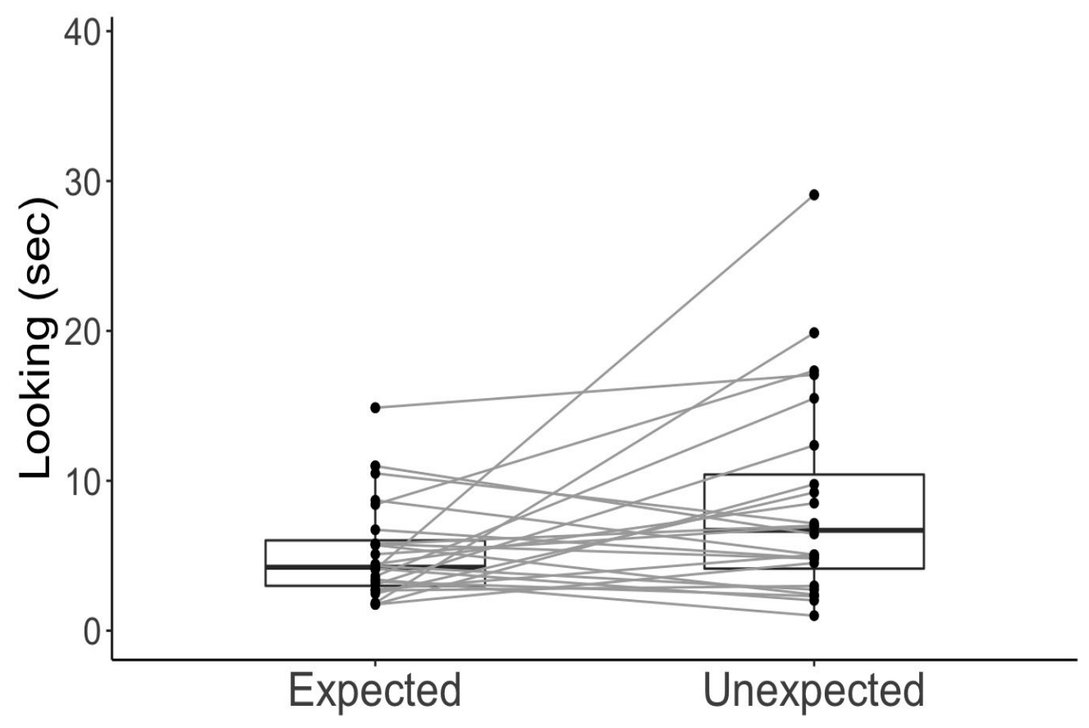
- **Participants**: 22 11-month-old infants were tested on both tasks (order counterbalanced). An additional four infants completed only one task, leading to the total of 24 infants per task (Total N = 26, 12 female, *M*age = 11.12 months, SD = 0.38).
- **Methods:** Infants were tested via Zoom. They saw one episode from the Preference Task and/or one from the Efficiency Task (Path Control). Infants' looking time was live coded using PyHab.

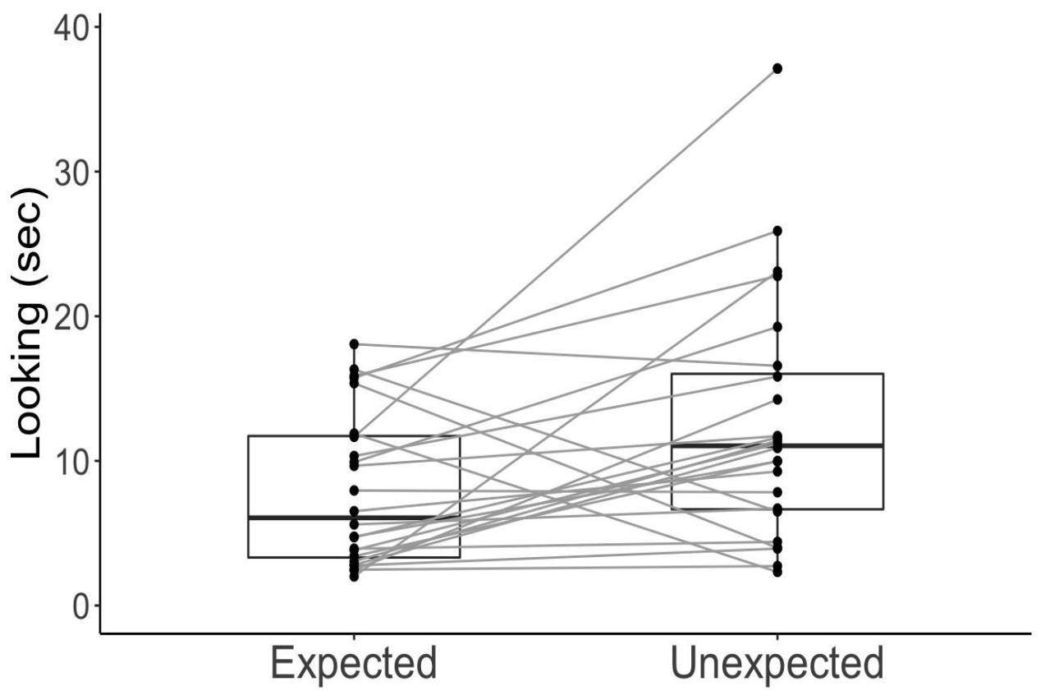


A) Unexpected (Efficiency Task)

B) Expected (Efficiency Task)







- **Results:** Infants looked longer to the unexpected compared to the expected outcomes on both the Preference Task (β=3.24, *p*=.040, *M*expected = 5.26, SE = 0.68; *M*unexpected = 8.50, SE = 1.40) and the Efficiency task (β=4.50, *p*=.016; *M*expected = 7.96, SE = 1.08, *M*unexpected = 12.47, SE = 1.70). .

Conclusions

- BIB's adoption of the content and methods of developmental psychology means its results can be interpreted in terms of human performance and makes it appropriate for direct validation with human infants.
- State-of-the art baseline models fail to demonstrate infant-like understanding of intentional agents. In particular, they fail to recognize that agents have object-based, as opposed to location-based goals, and they fail to modulate their predictions about an agent's efficient action based on whether that agent has previously acted rationally or irrationally.
- BIB's pilot validation with infants suggests successes on both the Preference Task and Efficiency Task.
- Human early agency-reasoning capacities are robust and persist even with simple overhead visual displays with minimal cues to agency.

References:

[1] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," Cognition, vol. 69, no. 1, pp. 1–34, 1998.
[2] F. Heider., and M. Simmel, "An experimental study of apparent behavior", The American journal of psychology, 57(2), 243-259.
[3] G. Gergely, Z. Nádasdy, G. Csibra, and S.Bíró, "Taking the intentional stance at 12 months of age," Cognition, vol. 56, no. 2, pp. 165–193, 1995.
[4] J. S. Buresh, and A. L. Woodward, "Infants track action goals within and across agents," Cognition, vol. 104, no. 2, pp. 287–314, 2007.
[5] A. L. Woodward, and J. A. Sommerville, "Twelve-month-old infants interpret action in context," Psychological Science, vol. 11, no. 1, pp. 73–77, 2000.
[6] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," Behavioral and brain sciences, vol. 40, 2017.
[6] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum, "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing," Nature Human Behaviour, vol. 1, no. 4, pp. 1–10, 2017.
[8] Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine theory of mind," in Proceedings of the 35th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4218–4227. [Online]. Available: http://proceedings.mlr.press/v80/rabinowitz18a.html

Acknowledgments:

This research was supported by the DARPA - Machine Common Sense (HR001119S0005 to B.M.L. and M.R.D.), as well as the National Science Foundation (DRL-1845924 to M.R.D). We thank Victoria Romero, Koleen McKrink, David Moore, Lisa Oakes, Clark Dorman, Thomas Schellenberg, Dean Wetherby, and Brian Pippin for their feedback and our participating families.