# An Infant-Cognition Inspired Machine Benchmark for Identifying Agency, Affiliation, Belief, and Intention

**Li Wenjie** [1] **Shannon C. Yasuda** [2] **Moira R. Dillon** [2] **Brenden M. Lake** [1,2]
{wenjieli, shannon.yasuda, moira.dillon, brenden}@nyu.edu
[1] Center for Data Science, New York University
[2] Department of Psychology, New York University

## Abstract

Human infants have remarkable abilities to reason about the underlying and invisible causes that drive others' actions. These abilities are at the core of human social cognition throughout life. Artificial Intelligence (AI) systems continue to fall short in achieving this same commonsense social knowledge. Recent benchmarks focusing on social cognition and theory of mind have begun to address the gap between human and machine social intelligence, but they do not fully consider the social reasoning required to understand scenarios with multiple interacting agents. Building on such benchmarks, we present eight new tasks focusing on different early social competencies, as informed by behavioral experiments with infants. We use a self-supervised Transformer model as a baseline test of our new tasks, and in addition, we evaluate this model on a previous social-cognitive benchmark. While our model shows improved performance on the previous benchmark compared with other data-driven models, it performs sub-optimally on our new tasks, revealing the challenge of learning complex social interactions through visual data alone.

**Keywords:** Social Cognition; Theory of Mind; Deep Learning; Artificial Intelligence; Cognitive Development

## Introduction

Human communication, collaboration, and learning are deeply rooted in an ability to understand and interpret the social world, including identifying other agents and their affiliations, beliefs, and intentions (Astington & Pelletier, 1998; Krych-Appelbaum et al., 2007; Resches & Pereira, 2007). While human infants display remarkable proficiency in such understanding (J. K. Hamlin, Wynn, & Bloom, 2007; Powell & Spelke, 2013; Sommerville & Crane, 2009; Woodward, 1998), deep learning systems often struggle with even basic social-cognitive tasks (Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus & Davis, 2019). Modern deep learning architectures and training paradigms, particularly those focused on supervised learning, tend to reduce behavioral data to labels of classification problems, neglecting the nuances and complexities that factor into human reasoning (Carreira & Zisserman, 2017). Moreover, although recent large language models succeed in many language-based tasks (OpenAI, 2023), their successes are not robust, for example, to variations of classic social-cognitive tasks (Ullman, 2023). By starting from infants' foundational knowledge of the social world, we can begin to identify the building blocks and inductive biases essential for the development of flexible social reasoning, highlighting key elements missing from current AI systems aiming to capture human intelligence.

Initial steps have been taken to build human-like AI systems with core cognitive abilities (Gandhi, Stojnic, Lake, & Dillon, 2021; Rabinowitz et al., 2018; Shu et al., 2021), but the focus of this prior work has typically been on simpler social scenarios involving just one agent, omitting the more complex dynamics of scenarios with multiple interacting agents. For example, the Baby Intuitions Benchmark (BIB) directly compared the performance of machines and infants on six tasks assessing an observer's inferences about individual agents' goal-directed actions on objects (Gandhi et al., 2021; Stojnić, Gandhi, Yasuda, Lake, & Dillon, 2023). BIB offered both an important starting point for such a research program as well as a framework in which to create new tasks probing foundational social cognition not covered in its initial set of tasks.

Building on this framework established in BIB (Gandhi et al., 2021), we introduce eight new tasks focusing on different early social competencies, including infants' reasoning about other agents' goals, affiliations, beliefs, and intentions. These tasks are structurally complex, for example, challenging AI systems to track multiple agents' mental states and to differentiate among various passive, goal-directed, and socially driven behaviors. Due to this complexity, the new tasks are expected to pose a significant challenge for current AI systems. Along with our evaluation tasks, we also designed twelve background-training tasks to give machines an opportunity to learn the environment and related, but not overlapping, cognitive representations to those tested in the evaluation tasks. As a baseline, we evaluated our new tasks using a state-of-the-art Transformer model (Arnab et al., 2021; Vaswani et al., 2017) trained on next-frame prediction, employing a self-supervised paradigm without relying on synthetic labels or negative examples. Although we found that our model performed better than previously tested baselines on BIB's tasks, it made sub-optimal predictions on our new tasks, exposing weaknesses in its abilities to understand the complex causal relations captured by our tasks.

## Tasks

Eight new evaluation tasks and twelve new background training tasks, each with thousands of episodes, challenge AI systems to identify agency, affiliations, and the beliefs and intentions of interacting agents. A task is a video containing a 2D grid world with simple shapes, whose actions represent the social interactions among animate agents (Heider & Simmel, 1944). This design eliminates the vision challenges of naturalistic scenes, probing a machine's ability to learn higher-level cognitive representations from lower-level visual information (Gordon, 2016; Springer, Meier, & Berry, 1996). Moreover, an optional JSON file documents each frame's en-



In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.* 

vironment and object properties (e.g., coordinates), allowing researchers to vary the amount of oracle information.

Like BIB, our new tasks rely on a violation-of-expectation (VOE) paradigm, commonly used to test infants. For infants, the VOE paradigm uses their looking times to different outcomes to measure their implicit predictions: Infants tend to look longer at outcomes they find surprising (Spelke, 1985). To adopt this paradigm for the current tasks, we made pairs of videos, which start with the same eight familiarization events drawn from the same statistical distribution but end in either an expected or unexpected test event, given the context set up by the familiarization. A model is tasked to determine which of the two videos contains the more expected test event.

We include two tasks focused on social affiliations between agents (Approach), two tasks focused on the attribution of goals to agents, not objects (Object Goal), two tasks focused on the attribution of beliefs to agents (False Belief & True Belief), and two tasks focused on agents' helping and hindering behaviors (Helping & Hindering). Each task consists of 1000 episodes. Below we provide further detail about each task, explaining their structure and criterion for success.

## Approach: Social & Instrumental

Do models predict that an agent will imitate the actions of another agent it had affiliated with?

**Developmental Background.** Infants predict that members of the same social group will exhibit similar actions and that individual agents will approach other agents whose actions they imitate (Powell & Spelke, 2013, 2018). Unpublished research outlined in Spelke (2022), moreover, suggests that infants 7.5- to 13.5-months-old are surprised by group-inconsistent actions when those actions are non-causal, but they have no expectation when the actions include contacting an object and changing its color.

**Familiarization Events.** An agent approaches one of two target agents to establish a social affiliation (Figure 1a & b).

**Test Events.** As shown in Figure 1a & b, the test environment contains a new goal object. In Approach: Social, the agents never interact with the goal object. In the beginning of the event, the two target agents each sequentially move in unique patterns. The main agent then moves the same way either as the target agent it had previously approached (expected) or as the target agent it had previously not approached (unexpected).

To deter machines from solving the task with trivial heuristics such as pattern matching, we also compare Approach: Social with Approach: Instrumental, where agents are strategically placed so that the movement of the main agent could also be interpreted as an efficient and goal-directed action towards the new object, making it ambiguous whether its action is imitative and socially motivated. It is therefore less expected in Approach: Instrumental versus Approach: Social that the main agent's movement should be similar to the movement of the target agent it had approached. To keep the task formats consistent, we nevertheless treat the less expected outcome as an unexpected outcome.

## Object Goal: Agent & Object

Do models recognize and attribute goals only to an agent that displays self-propelled, efficient motion, but not to an object that is moved by an external force?

**Developmental Background.** Infants recognize that agents, but not objects, exhibit self-propelled motion (Cicchino & Rakison, 2008), have object-based goals (Woodward, 1998), and move rationally and efficiently toward their goals (Csibra, Gergely, Bıró, Koos, & Brockbank, 1999). For example, Woodward (1998) found that 5-month-old infants expected a hand, but not a mechanical claw, to reach consistently for a goal object, not to a goal location.

**Familiarization Events.** Each video contains a constantly rotating spinner, an ambiguous element that acts as either a goal-directed agent or passive object, and two static target objects (Figure 1c & d). In Object Goal: Agent (Figure 1c), the shape, positioned a short distance away from the spinner, initiates its own movement. In Object Goal: Object (Figure 1d), the shape begins moving only after contact with the spinner. In both scenarios, the shape moves until it contacts one of the target objects. The shape collides with the same target object on the same side of the grid world across trials. A gray square under the spinner ensures visual consistency between familiarization and test events.

**Test Events.** As shown in Figure 1c & d, the objects' locations are switched relative to their locations during familiarization. A gray square occludes the shape's starting position and potential contact with the spinner to make the cause of its movement ambiguous. At the start of each event, the shape emerges from behind the square and goes directly to one of the two objects. In the agent condition, the shape either approaches the previously contacted object at a new location (expected), or a new object at the previously visited location (unexpected). Similar to the Approach task, the evaluation compares the model's performance across the two conditions, Object Goal: Agent versus Object Goal: Object. Since objects cannot have preferences, here it is expected in Object Goal: Agent and (relatively) unexpected in Object Goal: Object that the shape approaches the previously contacted object at a new location. Conversely, it is unexpected in Object Goal: Agent and (relatively) expected in Object Goal: Object that the shape approaches the new object at the previously visited location.

## False Belief & True Belief

Do models predict that an agent will act based on what it has been present to observe?

**Developmental Background.** Older infants and younger toddlers predict that agents will act on objects based on where they last saw those objects (Onishi & Baillargeon, 2005; Scott & Baillargeon, 2017), and they prefer agents who intend to help based on what they last saw (Woo, Liu, Gweon, & Spelke, 2021; Kiley Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013).

**Familiarization Events.** In the first kind of event, an agent

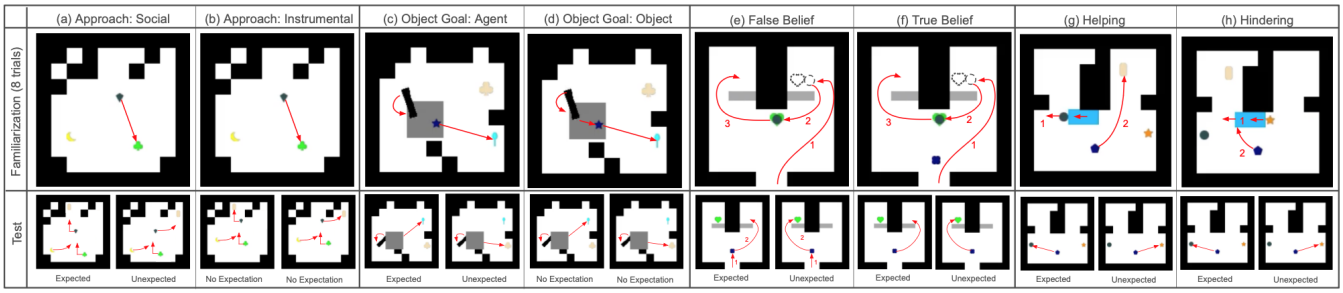


Figure 1: **Schematic Overview of the Evaluation Tasks.** Eight familiarization events are presented first to set up an expectation about the underlying agency, affiliations, beliefs, or intentions driving the behavior of each moving shape. At test, either of two test events are presented, one consistent and one inconsistent with the expectation set up by the familiarization events. Here, red arrows indicate the shapes' movements, and numbers indicate the order of these movements. For clarity, this figure only partially represents the familiarization events for (e) through (h).

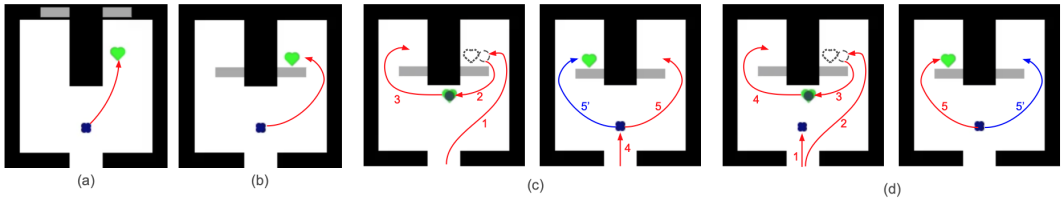


Figure 2: **Schematic of the False Belief and True Belief tasks.** Arrows indicate the direction of movements, and numbers indicate the order of these movements. In the first trial (a), a clover-shaped agent approaches an observable heart-shaped goal object. In subsequent familiarization events (b), the agent searches for its goal in the same room, even when grey occluders obstruct its view. At test, (c) in the False Belief task, when the clover-shaped agent is absent, a circular agent moves the goal to the other room (2,3) before leaving. The clover-shaped agent enters (4) and either goes to the original room, failing to find the goal object (red arrow, 5: expected), or to the room where the object had been moved (blue arrow, 5': unexpected). (d) In the True Belief task, the clover-shaped agent enters (1) before the circular agent enters (2), witnessing the object change location (3,4). Here, the clover-shaped agent would be expected to search in the new room for its goal (red arrow, 5).

moves toward a visible goal object located in one of two rooms (Figure 2). In the following events, the goal object is always placed in the same room, but depending on the placement of the two grey occluders, the agent may or may not be able to see the object's location (Figure 2a & b). The agent always moves to the same room to find the object, establishing that the agent looks for the object where it had last seen it.

**Test Events.** The goal object is initially located in the same room, but a second agent switches its location to the other room. In the True Belief task, the first agent sees the object change location. In the False Belief task, the first agent is not present during the switch, and so does not see the object change location. The first agent later searches in the original location (expected for False Belief, unexpected for True Belief) or the new location (unexpected for False Belief, expected for True Belief).

## Helping & Hindering

Do models infer that a goal-directed agent prefers an agent who helps it reach its goal, and not one that hinders it?

**Developmental Background.** Infants prefer agents who help others achieve their goals (Fawcett & Liszkowski, 2012; J. Hamlin, 2015; J. K. Hamlin & Wynn, 2011; Premack & Premack, 1997). For example, J. K. Hamlin (2013) found that 10-month-old infants were more likely to reach for a puppet who removed a door blocking another puppet's preferred object instead of reaching for a puppet who also removed a door, but one blocking the puppet's non-preferred object.

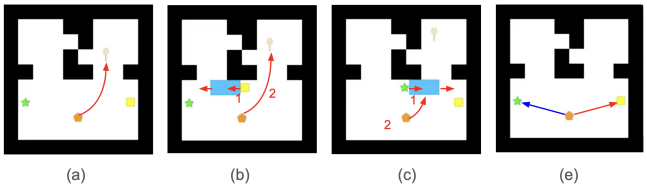


Figure 3: **Schematic of the Helping and Hindering tasks.** Arrows indicate the direction of movements, and numbers indicate the order of these movements. In the depicted scenario, a pentagonal agent moves toward a spoon-shaped goal (a). In the Helping task (b), a clover-shaped agent removes the obstacle to help the pentagonal agent reach its goal object. In the Hindering task (c), a star-shaped agent places an obstacle between the pentagonal agent and the object to hinder the agent. In both tasks (e), the pentagonal agent is expected to approach the clover-shaped agent (red arrow), and is unexpected to approach the star-shaped agent (blue arrow).

**Familiarization Events.** In the first kind of event (Figure 3a), two agents observe another agent approach a goal

object. In the second kind of event, a barrier is placed in the environment. In the Helping task (Figure 3b), a helping agent removes the barrier blocking the goal-directed agent from reaching its goal while the other agent does nothing. In the Hindering task (Figure 3c), a hindering agent moves the barrier to prevent the goal-directed agent from reaching its goal while the other agent does nothing.

**Test Events.** At test, the goal and barriers are removed. The three agents are in the same room, placed apart from each other. In the Helping task, it is expected that the goal-directed agent approaches the helper and unexpected that it approaches the stationary agent. In the Hindering task, it is expected that the goal-directed agent approaches the stationary agent and unexpected that it approaches the hinderer.

## Background Training Tasks

Infants bring knowledge to experimental studies in the lab, unlike deep learning systems that start with limited inductive biases. We provide twelve background training tasks for data-driven learning algorithms to obtain background knowledge about social cognition, simple physics, and the format of our tasks. For example, machines could learn that agents have object-based preferences (Figure 4b), that objects move upon contact with a rotating spinner and stop after colliding with a wall, and the causal relations between the familiarization events and the test event. Importantly, the background tasks are designed to differ in low-level ways from the evaluation tasks (e.g., an agent never navigates to one of two possible goal objects in the background training tasks like they do in the evaluation tasks), requiring models to form abstract and integrated representations from the background training to succeed at the evaluation. Detailed descriptions and illustrations of the complete background training set are on the project website.

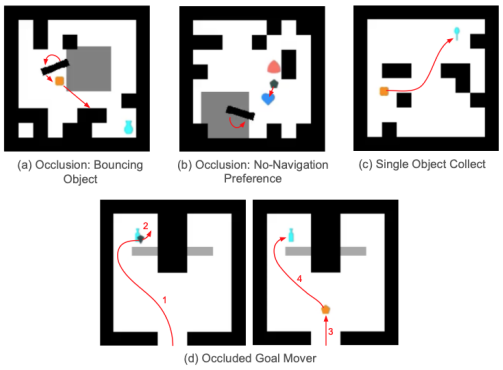


Figure 4: **Selected Background Training Tasks.** Models are trained on twelve background tasks that introduce the environment's dynamics as well as individual components of agents' and objects' behaviors that can be combined to solve the evaluation tasks.

Notably, and similar to infants' everyday experiences, the training set consists of expected examples only. While it is challenging to match the complexity of infants' real-world experiences, we believe this training set offers a reasonable foundation for meaningful comparison. We encourage the use of additional data to enhance machines' performance.

## Baseline Model

We propose a baseline model (Figure 5) that aims at predicting the subsequent frame in a test trial, given the preceding frames and one randomly selected familiarization trial as context. We choose a Transformer architecture for its strength in processing sequential data (Vaswani et al., 2017; Dosovitskiy et al., 2020; Arnab et al., 2021). The core challenge for the model is to learn to represent temporal-spatial continuities, causal relationships, and key concepts in social reasoning from the background training set.

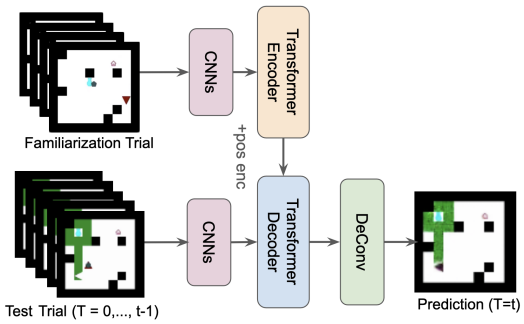


Figure 5: **Model Architecture and Training Procedure.** The Transformer model predicts the test frame at time $t$, given the frames from a familiarization event and test frames from time 0 to $t-1$.

The baseline model is trained with all the background training tasks from both the current work and BIB. Each task video is segmented into nine events–eight familiarization events followed by one test event. Frames are sampled from each event with a stride of 25 frames, with an upper limit of 20 frames for a single event [1]. The training objective of the model is to minimize the mean square error (MSE) loss between the predicted frames and target frames on a pixel-wise basis. A three-layer 2D CNN transforms each video frame into a series of patch embeddings, which are augmented with sinusoidal temporal-spatial positional encodings to preserve their sequential relations within the video. The embeddings of a familiarization event are encoded with self-attention to establish the task context. Prediction always starts from the second frame of a test event. To predict the $t^{th}$ frame, the Transformer decoder first applies self-attention on all preceding test frames up to the $(t-1)^{th}$ frame, and then incorporates the encoded context into the decoder using cross-attention. Finally, a two-layer 2D deconvolutional network transforms the decoder output into RGB format.

After 100 epochs of training, the model yields a loss of $5.5 \times 10^{-4}$ on the validation set, underscoring the model's predictive accuracy. For comparison, the baseline MSE between two successive frames was $2.6 \times 10^{-3}$ on average. Vi-

[1] We evenly sample 20 frames for rare events with more than 500 frames.

Figure 6: **Example Prediction on a Held-out task.** A clover-shaped agent navigates to the bottle-shaped goal object. Left: previous frame, Middle: current frame (target), Right: model prediction.

sualizations of model predictions on a held-out task (Figure 6) reveal that the predictions preserve the low-level shapes and colors of the elements in the environment (although with some blurring), as well as some higher-level properties such as efficient goal-directed action.

## Alternative Models

We also compared our model with the other models previously tested on BIB. These models employ different architectures and training approaches and, in some cases, include oracle information and hand-selected priors (Table 1). BC-MLP and BC-RNN are behavioral cloning models trained to predict future coordinates of the main agent (Gandhi et al., 2021). Video-RNN, like our model, processes and constructs pixel images of task frames. The VT model uses Transformer-like attention mechanisms to predict both frames and agent coordinates (Hein & Diepold, 2022). Finally, HBToM predicts the coordinates of the agent by leveraging a hierarchical Bayesian approach to construct relevant cognitive functions in a probabilistic model (Zhi-Xuan et al., 2022). The prediction accuracy of these models on BIB is shown in Table 2. Future work might evaluate these models on our new tasks since these previous models either make predictions based on the positions of a single, rather than multiple, agents (BC-MLP, HBToM, and VT), or require task-specific knowledge (HBToM).

## Evaluation

Models' predictions are successful when they identify the one video from the pair that conforms to social commonsense targeted in each task. For our baseline model, this means its prediction error on the expected video is lower than that of the unexpected video. The performance of the baseline model on our new tasks is presented in Table 3 along with its performance on BIB in Table 2. We discuss some of the predictions of the baseline model in detail below to illustrate its capabilities and limitations.

**Goal Preference (BIB).** VT and the baseline model substantially outperform MLP- and RNN-based models in BIB's preference and inaccessible goal tasks. As shown in Table 2, VT scores 82.1% and 89.8%, and our Transformer model scores 73.7% and 78.8%. These results highlight the strength of the attention mechanism in VT and our Transformer model in capturing element-wise relations.

**Instrumental Action (BIB).** The current baseline model makes accurate predictions (97.9%) in the No Barrier task, slightly above chance predictions (57.3%) in the Inconsequential Barrier task, and below chance predictions (21.4%) in the Blocking Barrier task. The model often fails to capture the sequential and causal relations in the events. For example, in the Blocking Barrier task (Figure 7), the model predicts that the green wall will fade before the agent even reaches the lock. This is likely because the model associates fading with the number of frames that have elapsed in the video instead of learning the causal relation between the key and the wall. This heuristic fails during the evaluation because the agent has to travel farther (taking more frames) to the key.

**Helping & Hindering.** The current baseline model achieves

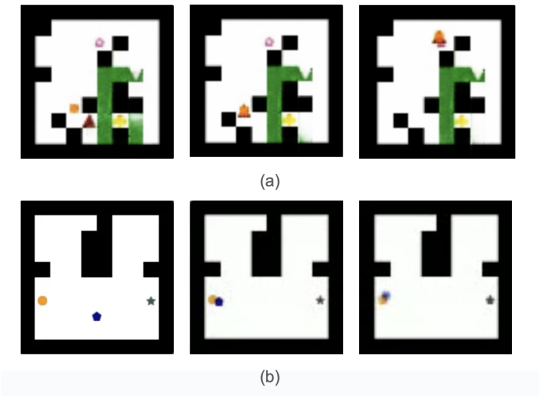


Figure 7: **Example Predictions on Evaluation Tasks.** (a) In an Instrumental Action: Blocking Barrier task, the baseline model predicts that the orange agent will approach the red key (left). Later, the model predicts that the green wall will fade when the agent picks up the key (middle) and when it approaches the lock with the key (right). (b) In a Hindering task, the model correctly predicts that the agent will approach the stationary agent (right), but fails to generate the agent at the correct location near the target agent (center).

an accuracy of 60.7% and 58.3%. A closer examination of the predicted frames suggests that the model might have been capable of identifying the target agent but fails to predict the approach trajectory. In the Hindering task (Figure 7b), a blue pentagonal agent is seen below a stationary agent (orange circle) and a hindering agent. The model correctly predicts that the pentagonal agent will approach the orange circular agent, but the two agents overlap at the orange agent's upper-right corner (right), instead of the lower-right corner, which would be on the path to the orange agent (center). This discrepancy can likely be traced back to the model's over-reliance on the statistics of the background training, where the main agent always starts from the top part of the grid in tasks showing hindering behaviors.

**True & False Belief.** The current baseline model achieves an accuracy of 97.5% and 2.4%, indicating that it expects the agent to approach the goal behind the wall, regardless of its belief about the relocation. Such predictions could be attributed to either genuine ignorance or false belief or to a failure to understand the role of the occluder in creating an environment with only partial visibility.

Table 1: **Oracle information used by different models.**

| | Deep Learning Models | | | | | Bayesian Principled Model |
|---|---|---|---|---|---|---|
| Privileged Information | BC-MLP | BC-RNN | Video-RNN | VT | Transformer (Ours) | HBToM |
| Environment meta-data (element type, coordinates, etc.) | x | x | | x | | x |
| Built-in inductive biases | | | | x | | x |

Table 2: **Comparisons of model prediction accuracy (%) on BIB.**

| | Deep Learning Models | | | | | Bayesian Principled Model |
|---|---|---|---|---|---|---|
| Task Name | BC-MLP | BC-RNN | Video-RNN | VT | Transformer (Ours) | HBToM |
| Preference | 26.3 | 48.3 | 47.6 | 82.1 | 73.7 | 99.7 |
| Multi-Agent | 48.7 | 48.3 | 50.3 | 49.1 | 50.2 | 99.2 |
| Inaccessible Goal | 73.3 | 80.7 | 71.8 | 78.9 | 78.9 | 99.7 |
| Efficiency: Path Control | 94.0 | 92.8 | 99.2 | 96.0 | 96.0 | 94.9 |
| Efficiency: Time Control | 99.1 | 99.1 | 99.9 | 99.0 | 99.9 | 97.2 |
| Efficiency: Irrational Agent | 73.3 | 55.7 | 50.1 | 29.5 | 80.4 | 96.6 |
| Instrumental: No Barrier | 98.8 | 98.8 | 99.7 | 98.7 | 97.9 | 98.8 |
| Instrumental: Inconseq Barrier | 55.2 | 78.2 | 77.0 | 96.9 | 57.3 | 97.0 |
| Instrumental: Blocking Barrier | 47.2 | 56.6 | 62.5 | 82.1 | 21.4 | 99.7 |

Table 3: **Baseline prediction accuracy (%) on the new social cognition tasks.**

| Task Name | Baseline |
|---|---|
| Approach: Social | 38.6 |
| Approach: Instrumental | 50.2 |
| Goal Attribution: Agent | 40.8 |
| Goal Attribution: Object | 53.5 |
| True Belief | 97.5 |
| False Belief | 2.4 |
| Helping | 60.7 |
| Hindering | 58.3 |

## Discussion

Machine benchmarks focusing on human social cognition create valuable opportunities to develop AI systems with human-like and human-compatible competencies. By comparing the predictions of AI and infants, we can align the goals of AI systems with the foundational cognitive building blocks of human cognition. Our present work builds on previous work, in particular on the Baby Intuitions Benchmark (Gandhi et al., 2021; Stojnić et al., 2023), by introducing eight new evaluation tasks that explore various social-cognitive abilities present from human infancy, including reasoning about agency, affiliations, belief, and intention. We also generated twelve background training tasks to provide machines an opportunity to learn the environmental dynamics and individual components of agents' and objects' behaviors that could be combined to solve the evaluation tasks.

We updated the lower-bound for data-driven machine social reasoning with a Transformer baseline, which is trained with a self-supervised learning paradigm. Its non-task-specific architecture and training procedure provide an adaptable pipeline for future datasets structured around the VOE paradigm. Despite showing promise on BIB's tasks without oracle guidance (Table 1), its performance on our new tasks highlights the necessity for more powerful AI systems capable of reasoning about complex environments and the social relationships among agents.

What can we learn from existing computational models of social cognition, and how do we create AI systems that can identify agency, affiliation, belief, and intention like infants do? Our results point to two complementary strategies: creating more realistic training data and integrating cognitive inductive biases into model architectures. The disparity between the artificial and passive learning environments we provided cannot fully capture the data distribution in the rich, multi-modal, and interactive experiences that support infant learning. Efforts to bridge this gap have included capturing infants' sensory experiences through head-mounted cameras (Vong, Wang, Orhan, & Lake, 2024; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021), eye-tracking (Mendez, Yu, & Smith, n.d.; Candy et al., 2023) and simulating interaction with the environment via embodied agents (Wykowska, Chaminade, & Cheng, 2016). Our evaluation tasks are poised to serve as a critical testing ground for models trained on these datasets.

On the other hand, structured Bayesian models aim to capture the innate knowledge and biases that infants possess to facilitate efficient generalization. Existing models, such as BIPaCK (Shu et al., 2021) and HBToM (Zhi-Xuan et al., 2022), perform well on synthetic benchmarks built with similar priors as the models. However, they may generalize poorly to real scenarios whose distributions are not fully captured by the priors selected by the modelers. With the long-term goal of developing machines that have infant-like social reasoning, we hope that our current work stimulates future investigations that generate new modeling strategies as well as new approaches that combine the strengths of existing strategies.

## Acknowledgement

The authors would like to thank Emin Orhan, Kevin Smith, Wentao Wang, Alexa Tartaglini, Yanli Zhou, as well as the anonymous reviewers for their suggestions and feedback. The project website is https://github.com/wliwenjieli/InfantSocialBenchmark.

## References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 6836–6846).

Astington, J. W., & Pelletier, J. (1998). The language of mind: Its role in teaching and learning. *The handbook of education and human development: New models of learning, teaching and schooling*, 569–593.

Candy, T. R., Dalessandro, A., Tellez, V., Biehn, S., Mestre, C., Haaff, T., ... Smith, L. (2023). The distribution of gaze positions of human infants in natural behavior. *Journal of Vision*, *23*(9), 4999–4999.

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6299–6308).

Cicchino, J. B., & Rakison, D. H. (2008). Producing and processing self-propelled motion in infancy. *Developmental Psychology*, *44*(5), 1232.

Csibra, G., Gergely, G., Bıró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason'in infancy. *Cognition*, *72*(3), 237–267.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fawcett, C., & Liszkowski, U. (2012). Observation and initiation of joint action in infants. *Child Development*, *83*(2), 434–441.

Gandhi, K., Stojnic, G., Lake, B. M., & Dillon, M. R. (2021). Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, *34*, 9963–9976.

Gordon, A. (2016). Commonsense interpretation of triangle behavior. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 30).

Hamlin, J. (2015). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for helpers over hinderers in the hill paradigm. *Frontiers in psychology*, *5*, 1563.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive development*, *26*(1), 30–39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Hein, A., & Diepold, K. (2022). Comparing intuitions about agents' goals, preferences and actions in human infants and video transformers. In *Svrhm 2022 workshop @ neurips.*

Kiley Hamlin, J., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209-226.

Krych-Appelbaum, M., Law, J. B., Jones, D., Barnacz, A., Johnson, A., & Keenan, J. P. (2007). "i think i know what you mean": The role of theory of mind in collaborative communication. *Interaction Studies*, *8*(2), 267–280.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Marcus, G., & Davis, E. (2019). *Rebooting ai: Building artificial intelligence we can trust*. Vintage.

Mendez, A. H., Yu, C., & Smith, L. B. (n.d.). Controlling the input: How one-year-old infants sustain visual attention. *Developmental Science*, e13445.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *science*, *308*(5719), 255–258.

OpenAI. (2023). *ChatGPT*. (https://chat.openai.com/chat)

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, *110*(41), E3965–E3972.

Powell, L. J., & Spelke, E. S. (2018). Third-party preferences for imitators in preverbal infants. *Open Mind*, *2*(2), 61–71.

Premack, D., & Premack, A. J. (1997). Infants attribute value±to the goal-directed actions of self-propelled objects. *Journal of cognitive neuroscience*, *9*(6), 848–856.

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).

Resches, M., & Pereira, M. P. (2007). Referential communication abilities and theory of mind development in preschool children. *Journal of Child Language*, *34*(1), 21–52.

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249.

Shu, T., Bhandwaldar, A., Gan, C., Smith, K., Liu, S., Gutfreund, D., ... Ullman, T. (2021). Agent: A benchmark for core psychological reasoning. In *International conference on machine learning* (pp. 9614–9625).

Sommerville, J. A., & Crane, C. C. (2009). Ten-month-old infants use prior information to identify an actor's goal. *Developmental science*, *12*(2), 314–325.

Spelke, E. S. (1985). Preferential-looking methods as tools for the study of cognition in infancy.

Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1* (Vol. 1). Oxford University Press.

Springer, K., Meier, J. A., & Berry, D. S. (1996). Nonverbal bases of social perception: Developmental change in sensitivity to patterns of motion that reveal interpersonal events. *Journal of Nonverbal Behavior*, *20*, 199–211.

Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, *235*, 105406.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, *5*, 20–29.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504-511. doi: 10.1126/science.adi1374

Woo, B. M., Liu, S., Gweon, H., & Spelke, E. (2021). Who needs more help? sixteen-month-old infants prefer to look at and reach for helpers who help with harder tasks. In *Proceedings of the annual meeting of the cognitive science society.*

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

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375.

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*.