An Infant-Inspired Benchmark for Machine Social Cognition

Wenjie Li
Center for Data Science
New York University
wenjieli@nyu.edu

Shannon Yasuda
Department of Psychology
New York University
shannon.yasuda@nyu.edu

Moira Dillon
Department of Psychology
New York University
moira.dillon@nyu.edu

Brenden Lake
Center for Data Science
New York University
brenden@nyu.edu

Abstract

Preverbal infants have remarkable abilities to understand others' intentions, goals, and social affiliations, which lay the groundwork for complex cognitive and language development, crucial for navigating the intricacies of human social dynamics throughout life. In contrast, recent development in Artificial Intelligence systems are often designed to emulate human-like behaviors by extracting common sense knowledge from language or sensory data. Despite their impressive performance in many aspects, they fail to recover some foundational theory of mind capacities found in early infancy. This discrepancy highlights critical gap to create AI that understands humans and think like humans. To address this, we introduce a suite of theory of mind tasks drawn from studies of infants social cognition. This work is an extension of our prior work Baby Intuition Benchmark (BIB), Expanding on our prior work the Baby Intuition Benchmark (BIB), a suite of theory of mind tasks drawn from infant studies, to challenge AI systems to understand others' goals and social affiliations, enriching Baby Intuition Benchmark, our prior benchmark focusing on reasoning about other goal-driven agents. We evaluate both benchmarks using a Transformer model trained with a self-supervised learning paradigm. The model shows improved performance over existing baselines, elevating the upper-bound of deep learning models' capacities in causal and object-oriented reasoning. However, it still demonstrates limitations of AI to represent others' mental states, underscoring the challenges in achieving human-like theory of mind reasoning in AI.

1 Introduction

Human communication, collaboration, and learning are deeply rooted in our ability to understand and interpret the minds of others. This fundamental capacity underpins every aspect of our social interactions and intellectual growth, forming the cornerstone of our collective and individual experiences. (Astington and Pelletier, 1996; Krych-Appelbaum et al., 2007; Resches and Pereira, 2007). Preverbal infants, despite their limited experience in the world, demonstrate remarkable proficiency in inferring the mental states and social affiliations of other individuals (Woodward, Sommerville, Powell, years?). In contrast, deep learning systems frequently struggle with basic social reasoning tasks, stemming from a lack of inductive biases essential for understanding other agents (Gandhi et al.,2021; Stojnic et al., 2023; Lake et al.,2017; Marcus and Davis, 2019). Often, these systems

Preprint. Under review.

resort to pattern matching in others' actions, neglecting to infer the underlying goals, preferences, and social affiliations of other individuals. Predominantly, modern deep learning architectures and training paradigms, particularly those focused on supervised learning, tend to treat behavioral data as mere classification problems, thus overlooking the more nuanced signals present in latent mental states (image classification, Kinetics, Something-Something, autonomous vehicle). The recent advent of large language models has demonstrated major strides in achieving human-level intelligence by extracting semantic information from vast language data (citation, gpt4 report?). However, the representations and predictions of these models are not robust to classic theory of mind tasks (Ullman, 2023; other theory of mind llm paper) demonstrating a failure of these models to truly reason about others like humans do. In contrast, infants, with exposure to much less data, display flexible, sensitive, and robust theory of mind capabilities. By starting from infants' small knowledge repertoire, we can begin to identify the foundational building blocks and inductive biases essential for the development of versatile social reasoning skills, highlighting key elements missing in current AI development.

Initial steps have been taken to bridge the gap between Artificial Intelligence (AI) systems and infant social cognition (Ghandi et al., 2021; Shu et al., 2021). The Baby Intuition Benchmark (BIB) introduces a suite of six tasks to assess understanding and inferences of others' goals, preferences, and mental states in infants and machines. These tasks evaluate whether infants or machines understand agents are driven by goal objects instead of goal locations (Woodward); they navigate and use efficient actions to reach goals; different agents may have different goals; and sometimes intermediate goals and instrumental actions are taken to reach the final goals. Published concurrently, AGENT offers a similar benchmark with a set of complementary tasks about comprehension of goal-directed agents. However, these existing benchmarks omitted a broad range of cognitive abilities, which we sort to expand in this work. The tasks we introduce are more complex and cognitively demanding, often requiring tracking of multiple agents' mental states or identifying different types of visually similar passive, goal-directed, or socially affiliated entities. These capacities emerge in later developmental stages and and thus we expect the new benchmark to be more challenging and demanding for both infants and machines.

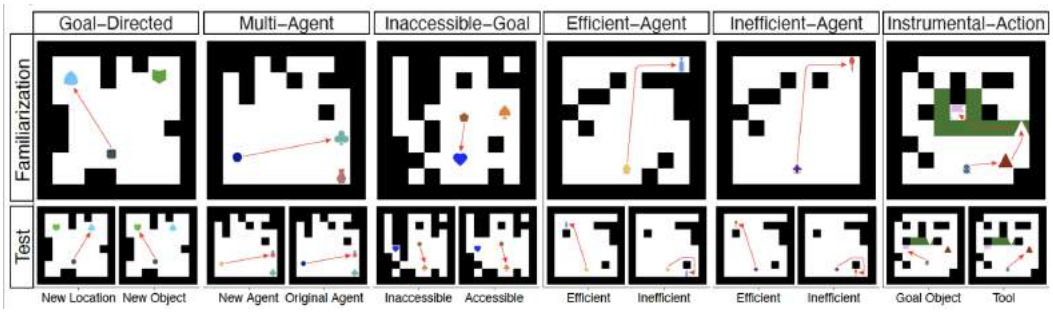

Figure 1: Existing tasks in Baby Intuition Benchmark. Figure credit to Stojni et al., 2023

To initiate the first step towards aligning machine theory of mind with that of infants, we employ a state-of-the-art Transformer model (Vaswani et al., 2017) as a baseline for the benchmarks. The model is trained in a self-supervised paradigm to perform a next-frame prediction task. Using no artificial labels, we hope to emulate the data availability in humans' day-to-day world.

2 An Infant-Inspired Benchmark for Machine Social Cognition

As an expansion of BIB (Kanishk et al., 2022), our new tasks continue the same format, using videos of geometric shapes moving in a grid-world environment (Heider and Simmel, 1944). It is shown that infants and adults can attribute animacy to simple geometric shapes based on behavioral cues alone. We employ this approach for its efficiency in task generation and its abstraction from real-life agents. This choice streamlines the engineering process for creating thousands of videos to train and test models, and eliminates the vision challenges of machine understanding of naturalistic scenes, thereby focusing on the machines' ability to learn higher-level cognitive functions. It also requires the machines to learn a flexible mental representation of the events in the scene to succeed at each task. instead of relying on low-level perceptual cues.

To facilitate a direct comparison between machine learning models and infants in laboratory settings, like BIB, the social cognition tasks are designed within the Violation-of-Expectation (VOE) paradigm, a classic method in developmental psychology experiments. Each task is structured into nine trials within the same environment. The initial eight trials act as a familiarization phase, designed to set up an expectation, consistently sampling from the same statistical distribution, followed by either an expected test trial or an unexpected test trial. Despite being perceptually similar, these conditions represent distinct concepts that align or conflict with the familiarization trials, ensuring that task completion hinges on a deep conceptual understanding rather than mere pattern recognition. In developmental studies, infants' looking times to either test trial are evaluated. If they've understood the familiarization trials, infants will look longer at unexpected events. We can similarly test machines' "expectations" by having them make predictions at test, and comparing these predictions to both the expected and unexpected test trials.

We present five new tasks focusing on 1) imitation, 2) social approach, 3) helper/hinderer, 4) true/false belief, and 5) goal attribution, each of which consists of 4000 videos. The first four tasks evaluate an observer's ability to deduce social affiliations by identifying actions of imitation, help, or hindrance among social partners. The Goal Attribution task examines whether infants or machines can discern goals attributed to agents as opposed to inanimate objects, offering key insights into whether simple geometric shapes are effectively perceived as distinct entities based on their movement patterns. The following sections will provide detailed insights into each task, delineating their structure, execution process, and the criteria that define successful completion

2.1 Inferring Social Affiliation From Social Imitation

Can an AI system infer that an agent will approach and, therefore affiliate with, another agent whose behavior it socially imitates, but not when its imitative actions are required to reach an instrumental goal?

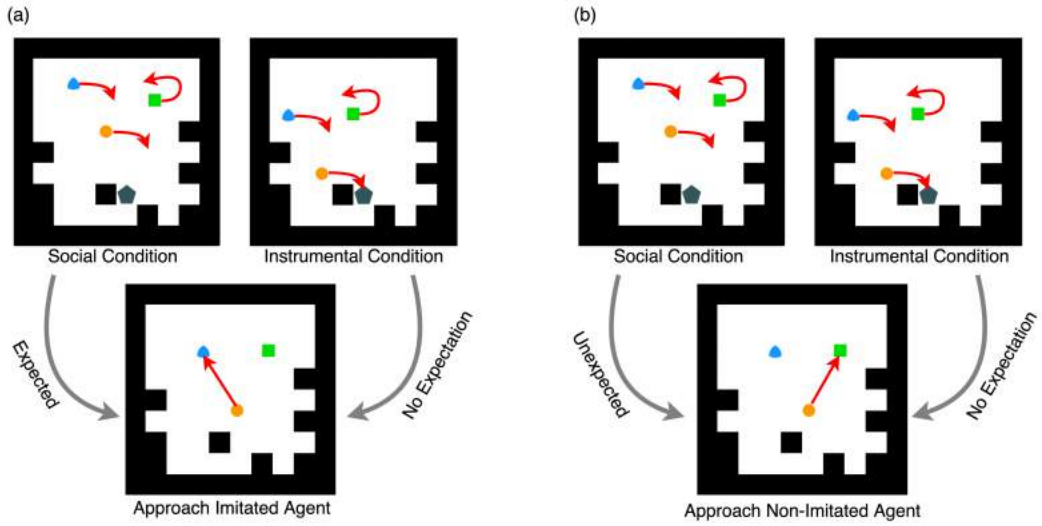

Figure 2: Imitation

Developmental Background. Infants expect an agent to approach and affiliate with others it imitates. Research by Powell & Spelke (2018) demonstrated that infants as young as 4 months expect imitators to approach, and therefore affiliate with, those whose sounds they imitate. Pilot research by Powell & Spelke et al. (2014) has shown that infants as young as 7.5 month olds were surprised by group-inconsistent actions only when those actions were non-causal. When the actions caused members of a social group to make contact with a goal object, which changed color upon contact, infants did not consider the action to be group-inconsistent. Other research has demonstrated that older infants, when imitating others, can distinguish between others' social and instrumental actions (Gergely et al., 2002; Carpenter et al., 2005; Powell, 2021) suggesting that pure social imitation,

as opposed to goal-driven behavior, does not directly serve the agent's own goals, thus eliciting a stronger social preference. Can AI systems distinguish between purely social and instrumental actions, associating only the former with social affiliation?

Familiarization Trials. In this task, we contrast conditions of social and instrumental imitation. Each condition involves two target agents, and a goal object. The target agents each exhibit a unique movement pattern (e.g. going down and to the right, or going up and to the left), with one demonstrating this pattern at the start of each trial. The imitating agent then replicates the movement pattern (going down and to the right) of one of the target agents. In the social condition, the imitating agent has sufficient space to mimic either target without otherwise engaging with objects of obstacles in the environment. In the instrumental condition, a goal object is placed at the end of the imitating agent's path, with obstacles arranged such that the chosen movement pattern is the only efficient way to reach the goal. The goal object changes color upon contact, signifying the causal effect of the approaching action (Liu, 2019). To maintain visual consistency between conditions, the same goal object is positioned in the same location, but the positions of the agents are shifted in the social condition such that the goal object is off of the trajectory of the imitating agent.

Test Trials. In the test trials, the goal object is removed. The imitating agent either approaches the target agent it imitates during familiarization (expected in the social condition, no specific expectation in the instrumental condition) or the one it does not imitate (unexpected in the social condition, no specific expectation in the instrumental condition). The observer is deemed successful if it assigns a higher probability a) for the imitator to approach the agent it imitated in the social condition compared to the instrumental condition, or b) for the instrumentally imitative agent rather than the socially imitative agent to approach the agent that was not imitated.

2.2 Predicting Imitation From Social Affiliation

Does an AI system expect an agent to socially imitates the actions of those it socially affiliates with?

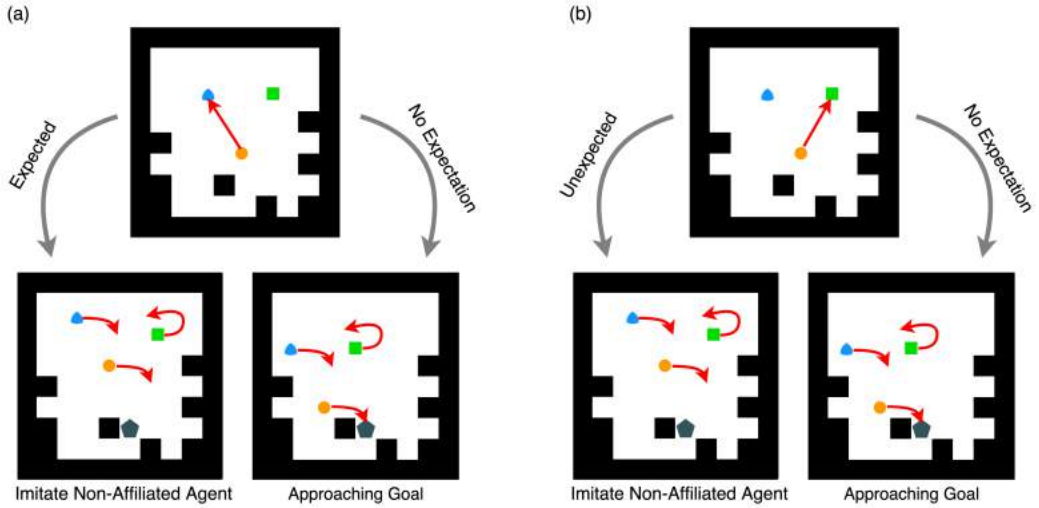

Figure 3: Social Approach

Developmental Background. Infants anticipate that members of the same social group will exhibit similar behaviors. In research by Powell and Spelke (2013), 8-month-old infants observed two groups of visually similar geometric figures maintaining proximity and performing circular "dance" movements within their respective groups. Notably, infants showed surprise when an individual aligned its movements with a member of the opposite group rather than conforming to the dance of its own group. Further experiments revealed that infants held the same expectations for agents from the same social group even if they looked visually distinct.

Familiarization Trials. An agent consistently approaches one of two target agents (blue triangle in (a) or green square in (b))across trials to establish social affiliation.

Test Trials. In the test trial, the two target agents each display a unique movement pattern. After sequential demonstrations of these patterns at the trial's outset, the imitating agent adopts the movement pattern of one of the target agents. There are two distinct conditions: In the instrumental condition, a target object and obstacles are strategically placed near the imitated agent, making its movement pattern the only efficient path to the target object. Here, observers should have no particular expectation regarding which target agent the imitator mimics, as any similarity in actions could be coincidental, stemming from the agent's goal-driven behavior (both test trials have no expectation). Conversely, in the social condition, the agent has the freedom to imitate either target agent and it is therefore expected when it mimics the agent it previously affiliated with and unexpected when it mimics the agent it did not previously affiliate with.

2.3 Inferring social preference to helpers over hinderers

Can an AI system infer that a goal-driven agent has a social preference for another agent who helps it achieve its goal over another agent who hinders it?

Developmental Background. In a 2007 empirical report, Hamlin, Wynn, and Bloom provided the first evidence that preverbal infants at 6 and at 10 months of age evaluate others on the basis of their helpful and unhelpful actions toward unknown third parties. In their "hill paradigm," a Climber puppet tried but failed to climb a steep hill, and was alternately bumped up the hill by the Helper and bumped down the hill by the Hinderer. After being habituated to these events, both 10- and 6-month-olds selectively reached for the Helper over the Hinderer. Research by Premack and Premack (1997) demonstrated that 12 month old infants can distinguish between socially negative and positive actions carried about by agents represented as simple 2D shapes. Further research as demonstrated that infants as young as 3 months old preferentially reach for a helper, who helps a climber reach its goal of climbing up a hill compared to a hinderer, who pushes the climber down the hill (Hamlin et al., 2010; Hamlin et al., 2007; Hamlin, 2015). These results have been replicated in multiple scenarios (Hamlin and Wynn, 2011; Hamlin et al., 2013). Infants also expect the agent being helped to approach the helper over the hinderer (Fawcett & Liskowski, 2012; Lee & Song, 2014; Kuhlmeier et al., 2003).

Familiarization Trials. The setting involves three agents positioned at the lower half of the environment, with the upper side divided by a wall into two separate rooms. A goal object appears alternately in one of these rooms in each trial. During the first four trials, two agents observe a third agent moving towards the goal object, which changes color upon contact. In the final four trials, a red barrier emerges, obstructing the entrance to one of the rooms. If the barrier blocks the room containing the goal object, one agent consistently moves the wall to the opposite side, allowing the agent with the previously demonstrated goal preference to access the room. In contrast, during other trials, a different agent shifts the wall to prevent the goal-oriented agent from entering the room.

Test Trials. In the expected trial, the goal-oriented agent approaches the agent who helped by moving the wall (the helping agent). Conversely, in the unexpected trial, it approaches the agent who moved the wall to hinder its progress (the hindering agent).

2.4 Goal Attribution

Can AI or infants attribute goal preference to an element that reaches its goal with self-propelled motion, and not to an element who starts moving after contact with a moving object? Besides probing AI's understanding of causal relationships and agency, this task also further validates that the simple geometric shapes in the tasks indeed elicits animacy.

Developmental Background. Research indicates exhibition of self-propelled motions to be key defining features of agency (Cicchino Rakison, 2008); agents have goals and preferences for said goal (Woodward, 1998); and agents move rationally and efficiently towards their goal (Csibra et al., 1999). Infants can readily distinguish between agents and objects and can encode an agent's preference for an object (Woodward, 1998). When an element is consistently pushed toward a target

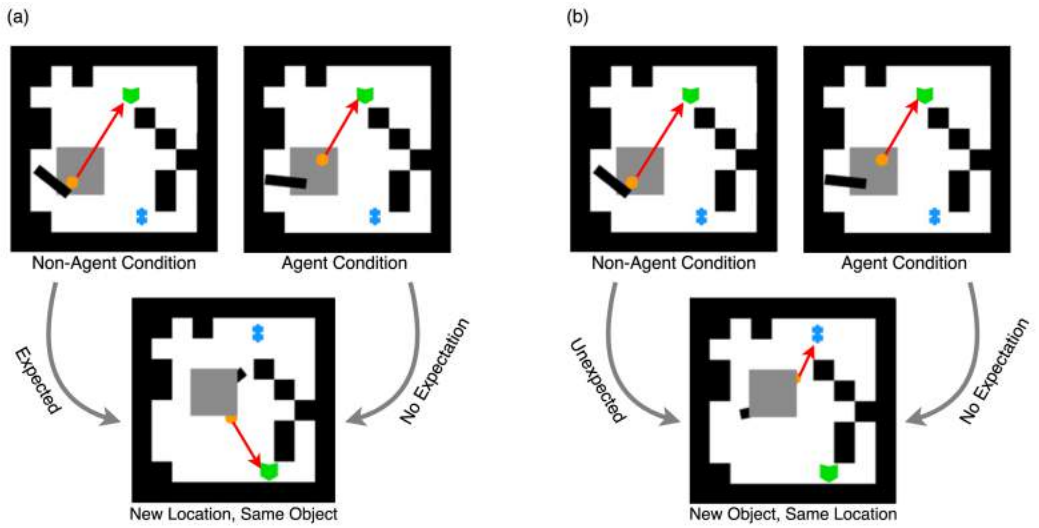

Figure 4: goal attribution

object by a mechanical spinner, can AI systems or infants distinguish that the element's consistent contact with the target is incidental, not a result of its own volition?

Familiarization Trials. The familiarization trials is comprised of a constantly rotating mechanical spinner, an ambiguous element potentially acting as a goal-driven agent or passive object, and two static target objects. Two conditions are presented: in the non-agent condition, the element begins moving after contact with the spinner, in a direction perpendicular to the spinner's arm at the point of contact; in the agent condition, the element, positioned a short distance from the spinner, initiates its own movement. In both scenarios, the element moves straight until colliding with one of the target objects, both of which change color upon impact. The element consistently collides with the same target object at a similar location in each episode. A gray square under the spinner ensures visual consistency between familiarization and test trials.

Test Trials. In the test trials, the target objects' locations are switched. A gray square partially covers the spinner to obscure the element's initial position, ensuring ambiguity in its movement cause. The spinner's starting rotation degree is averaged from both familiarization scenarios. These adjustments create uncertainty about whether the element's movement is due to spinner contact, relying on prior familiarization for agency inference. Shortly after each test trial begins, the element emerges from behind the occluder, moving straight towards either the familiar target object now in a new location (expected in the agent condition, no specific expectation in the non-agent condition), or a different target object in the same quadrant it previously approached (unexpected in the agent condition, no specific expectation in the non-agent condition). Successful AI prediction involves anticipating that the element in the agent condition will pursue the same target object, even in a new location, or finding it less expected for the agent-condition

2.5 Background Training Set

Despite being new to the world, infants bring innate or acquired knowledge to complete the developmental tasks, unlike data-driven deep learning systems which start with limited biases. We craft a background training set to offer machines a generous learning opportunity. This set serves two purposes: firstly, to familiarize machines with the basic setup of the grid world environment and task elements, like the visual features of geometric figures and the structure of each trial. This foundational understanding is essential but not part of the test criteria. Secondly, the set aims to provide a rich environment for learning theory of mind concepts crucial for solving the evaluation tasks, such as understanding that an agent can't pass through an obstacle, it can initiate movement towards a goal, or imitate its social partners.

Like the evaluation set, each background training episode includes 8 familiarization trials followed by 1 test trial. Notably, similar to infants' experiences, there are no negative examples in the training set. While it's challenging to match the complexity of infants' real-world experiences, we believe this training set offers a reasonable foundation for meaningful comparison. We acknowledge that additional training may enhance machine performance.

No-Navigation Preference Task An agent consistently approaches one of two nearby target objects, with their locations varying across trials to indicate object-based, not location-based, goals. Both objects change color upon contact, providing an extra cue for goal-approach behavior and helping machines and infants to differentiate it from social approach behavior. The travel distance is shorter than in evaluation tasks to prevent pattern replication.

Contact causation without navigation A rotating spinner propels a passive entity towards one of two nearby target objects, both changing color upon contact. The target object approached may vary within an episode based on the spinner's initial rotation and the element's relative position, demonstrating that movement influenced by another object does not indicate preference.

Single Object Task An agent navigates around obstacles to efficiently reach the sole target object, which changes color upon contact. This setup helps machines learn about obstacle navigation and goal pursuit. Some episodes include a non-interfering spinner to prevent associating its presence with the absence of agency. Occasionally, the agent returns to its starting position post-interaction.

Contact causation with single object A rotating spinner initiates movement in a static element, propelling it in a straight line until it encounters a wall or target object. This teaches that contact with a moving object can trigger directional movement in a stationary element.

Social Imitation Task During familiarization, two agents sequentially exhibit unique movement patterns, followed by a third agent mimicking the pattern of one it consistently aligns with. The test trial involves the imitating agent approaching its previously mimicked partner. This task differs from the social condition in evaluation: there's no goal object, the "dance" patterns are unique, and both target agents demonstrate their moves in each trial, unlike the alternating pattern in evaluation trials.

Imitative Goal Approach Three agents and a goal object are present. During familiarization, one of two model agents moves uniquely without interacting with the goal object. The main agent then mimics this movement, incidentally contacting the goal object, which changes color. The main agent's path is the only efficient route to the goal due to environmental obstacles. At query, the main agent approaches the goal object, indicating that matching movement trajectories can be coincidental and do not necessarily imply social affiliation.

Helper/Hinder task This tasks employs the same environment as the helper/hinderer task in evaluation. However, the locations of the main agent and the goal object are swapped. So instead of (un)blocking a small room containing the goal object to let (prevent) the goal-seeking agent in, during training the helper (hinderer) (un)blocks the small room to let (prevent) the agent into the bigger room where the goal object is.

True/False belief task Like its evaluation counterpart, this task is consisted of 8 familiarization trials where an agent consistently going back to a room where it last saw a goal object. At test time, a different agent appear and move the goal object around within the same room with or without the agent's presence. The agent is then shown to go back to the same room to retrieve the goal object.

Occlusion by observers A gray square appears in random grid world locations. Initially, it is part of the background, allowing full visibility. In the test, it occludes elements in the same space, familiarizing machines with occlusion use in tasks and teaching them that occluded objects don't disappear. The occluder is utilized in all background training sets but isn't shown in earlier figures to provide an unobstructed view of the task setups.

3 Baseline Model

We use a Transformer Encoder-Decoder model as a baseline for the social cognition tasks. This model is selected for its proficiency in modeling sequential data and because of its ability to separately process two streams of data, making it a convenient pipeline for the familiarization-testing scheme. The primary training objective of our model is to predict the subsequent frame in the test trial given the preceding frames and the familiarization trials. By focusing on next-frame prediction, we aim to develop a model that understands temporal-spatial continuities and causal relationships. Accurate prediction of the next frame requires the model to utilize concepts of intuitive physics and psychology to understand the dynamics between agents and objects, effectively mimicking infant-like reasoning and prediction during a VOE-style experiment.

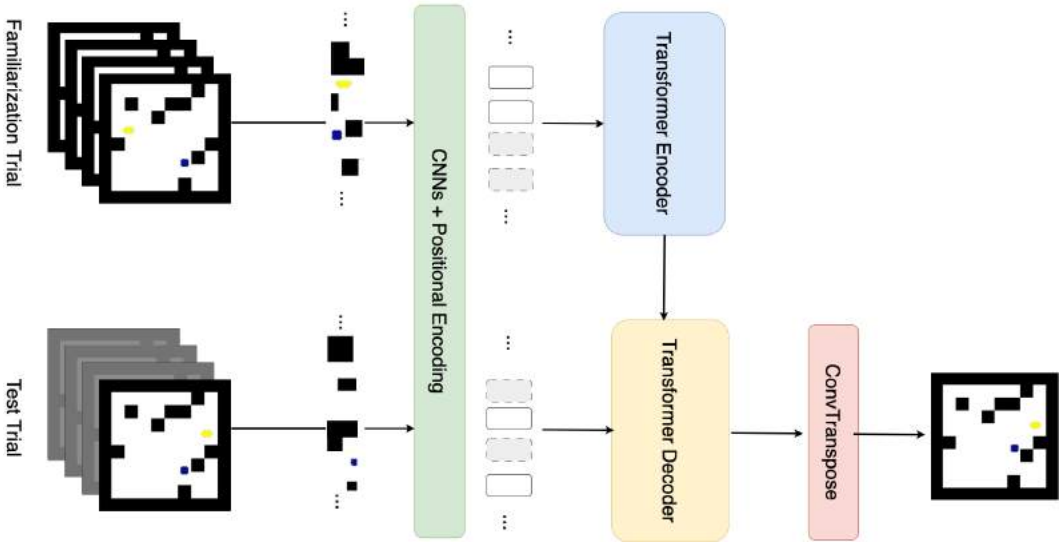

Figure 5: model architecture

During data preparation, each video is segmented into nine trials, from where two trials are randomly sampled at training time: one for familiarization and the other for testing. The encoder processes the familiarization trial, capturing the context for downstream prediction. The decoder processes the frames in the test trial in an auto-regressive manner, combining understanding of the previous frames and the output of the encoder to predict the next frame. Video frames are first sampled at a stride of 6 with a maximal sequence length of 40. In order to maintain temporal resolution and to fit the complete video, we further perform random token dropping which removes 62.5% frames, resulting in a maximum of 15 frames per video. The same token dropping procedure is also performed on the right-shifted target, so the target frame is always a stride of 6 frames away from the last decoder input, even if it is not the relative next frame in the input sequence after random token dropping. Furthermore, each frame is resized to 84 x 84 pixels, and split into 49 patches of 12 x 12 pixels, each with sinusoidal positional encoding consistent with their relative positions in the original video. We use a 128-dimensional embedding space, 8 attention heads, and 5 layers each in the encoder and decoder, balancing processing efficiency and model performance. We use mean square error to estimate training loss. The model is trained on 2 A100 GPUs for a week with a batch size of 48, with learning rate 1e-4, and weight decay of 1e-4.

During evaluation, we pair each of the eight familiarization trials with the test trial and take average of next frame prediction errors. A prediction is correct if the prediction loss is lower on the expected condition of a task than on the unexpected condition. While a model on all BIB1 and BIB2 task is still under training, we will present preliminary training results on a model of all BIB1 task and the goal attribution task in BIB2. These tasks are paired together because of similar video lengths and relevant cognitive functions tested.

Task Name	BC-MLP	BC-RNN	Video-RNN	HBToM	VT	Transformer (NEW)
Preference	26.3	48.3	47.6	99.7	82.1	72.0
Multi-Agent	48.7	48.2	50.3	99.2	49.1	–
Inaccessible Goal	73.3	80.7	71.8	99.7	89.8	90.0
Efficiency – Path	94.0	92.8	99.2	94.9	97.3	97.3
Efficiency – Time	99.1	99.1	99.9	97.2	99.8	100.0
Efficiency – Irrational Agent	73.3	55.7	50.1	96.6	29.5	48.9
Instrumental: No Barrier	98.8	98.8	99.7	98.8	98.7	100.0
Instrumental: Inconsequential Barrier	55.2	78.2	77.0	97.0	96.9	95.8
Instrumental: Blocking Barrier	47.2	56.6	62.5	99.7	82.1	18.0
Goal Attribution	–	–	–	–	–	83.7

Privileged Information	BC-MLP	BC-RNN	Video-RNN	HBToM	VT	Transformer (NEW)
Environment meta-data (element type, coordinates, etc.)	x	x		x	x	
Built-in inductive biases				x		

4 Discussion

Continuing the work in Baby Intuition Benchmark (BIB), which directly compares the capacities of infants and machines to infer others' goals, preferences, and intentions, we introduce a suite of social cognition tasks that encompasses more complex theory of mind reasoning. For example, can AI systems understand members of the same social groups act alike, such as inference of social affiliations and goal attributions. We construct the benchmark in the format of the Violation-of-Expectation paradigm, a classical experiment setting in experiment research, enabling direct comparisons between infants and machines. These benchmarks create valuable opportunities towards developing Artificial Intelligence systems with human-like theory of mind capacities, by identifying and modeling the fundamental cognitive building blocks present in the minds of preverbal infants.

We explore the social cognition tasks along with all BIB tasks with a Transformer model trained with a self-supervised learning paradigm. The non-task specific architecture and training procedure provide a pipeline for future modeling work of developmental experiments with VOE paradigm. This is by far the best performing model trained with no oracle signals or labels (insert table from research log, table description explain why this is the case). It performs better than BIB1 baselines on a few tasks. Notably, it achieves 80% accuracy in the preference task, identifying that other agents have goal objects, instead of goal locations, which all BC, offlineRL, and RNN struggled with in BIB1. It shows interesting performances on the set of instrumental action tasks, with perfect performance on no barrier or inconsequential barrier. Further examination reveals that the model succeed in these tasks trivially. These two tasks require the model to identify that retrieval of the key, an intermediate goal, is only instrumental when all paths leading to the goal is obstructed by a green wall which can be unlocked by the key. An agent is expected to directly approach the goal instead of the key if the green wall is not present or present but do not obstruct the goal. Our Transformer model correctly ignore the goal, but further examinations reveal that the model never learns to approach the goal in the background training set because the key is positioned very close to the agent and thus the short approaching action is omitted with a coarse sampling procedure. Our ongoing work focuses on finding solutions to model the benchmarks with finer temporal resolution with limited computational resources, potentially through longer training time or more aggressive random token dropping.

BIB, AGENT, and other intuitive psychology benchmarks has fostered the development of many new models for commonsense reasoning. Existing models fall into two categories: structured Bayesian models, such as BIPaCK and HBToM in AGENT and deep learning models, such as (RNN- or MLP-based) Behavioral-Cloning, Video RNN in BIB (Ghandi, 2021). As AI systems, probabilistic models often exceed or match human performance on benchmarks, but they rely on task-specific, hand-engineered inductive biases and features (BIPaCK, HBToM). On the other hand, deep learning models adopt an end-to-end approach, offering greater robustness in noisy environments. However, they are data-intensive and generally underperform when confronted with out-of-distribution signals, as evidenced in an attention-based model (VT, workshop paper) and other baseline models of

BIB1. From a cognitive modeling perspective, structured Bayesian models are instrumental in formalizing hypothesis testing when aligned with infants' data. Deep learning models based on artificial neural networks offer a unique, relatively hypothesis-agnostic platform for testing the connectionist hypothesis in the developing minds. Can knowledge about objects, intuitive physics, or even theory of mind be learned purely from data? And if not, what kind of inductive biases help a data-driven learner? This approach may provide insights to understanding the innateness of certain cognitive functions.

References

[1] Astington JW, Pelletier J. The language of mind: its role in learning and teaching. In: Olson DR, Torrance N, eds. The Handbook of Education and Human Development: New Models of Learning, Teaching and Schooling. Oxford: Blackwell; 1996, 593–619.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] Marcus, Gary, and Ernest Davis. Rebooting AI: Building artificial intelligence we can trust. Vintage, 2019.

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

[9] Shu, Tianmin, et al. "Agent: A benchmark for core psychological reasoning." International Conference on Machine Learning. PMLR, 2021.

[10] Woodward, Amanda L. "Infants selectively encode the goal object of an actor's reach." Cognition 69.1 (1998): 1-34.

[11] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

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

[13] Powell, Lindsey J., and Elizabeth S. Spelke. "Human infants' understanding of social imitation: Inferences of affiliation from third party observations." Cognition 170 (2018): 31-48.

[14] Powell, Lindsey J., Adena Schachner, and Elizabeth Spelke. "Infants' generalization of causal and non-causal actions across social groups." Poster presented at International Conference on Infant Studies, Berlin, Germany. 2014.

[15] Gergely, György. "The development of understanding self and agency." Blackwell handbook of childhood cognitive development (2002): 26-46.

[16] Carpenter, Stephen R. "Eutrophication of aquatic ecosystems: bistability and soil phosphorus." Proceedings of the National Academy of Sciences 102.29 (2005): 10002-10005.

[17] Powell, Lindsey J. "Adopted utility calculus: Origins of a concept of social affiliation." Perspectives on Psychological Science 17.5 (2022): 1215-1233.

[18] Powell, Lindsey J., and Elizabeth S. Spelke. "Preverbal infants expect members of social groups to act alike." Proceedings of the National Academy of Sciences 110.41 (2013): E3965-E3972.

[19] Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. "Social evaluation by preverbal infants." Nature 450.7169 (2007): 557-559.

[20] Premack, David, and Ann James Premack. "Infants attribute value±to the goal-directed actions of self-propelled objects." Journal of cognitive neuroscience 9.6 (1997): 848-856.

[21] Kiley Hamlin, J., Karen Wynn, and Paul Bloom. "Three-month-olds show a negativity bias in their social evaluations." Developmental science 13.6 (2010): 923-929.

[22] Hamlin, J. Kiley, Karen Wynn, and Paul Bloom. "Social evaluation by preverbal infants." Nature 450.7169 (2007): 557-559.

[23] Hamlin, J. K. "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 (2015): 1563.

[24] Hamlin, J. Kiley, and Karen Wynn. "Young infants prefer prosocial to antisocial others." Cognitive development 26.1 (2011): 30-39.

[25] Hamlin, J. Kiley. "Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core." Current Directions in Psychological Science 22.3 (2013): 186-193.

[26] Fawcett, Christine, and Ulf Liszkowski. "Observation and initiation of joint action in infants." Child Development 83.2 (2012): 434-441.

[27] Lee, Young-eun, et al. "The development of infants' sensitivity to behavioral intentions when inferring others' social preferences." PLoS One 10.9 (2015): e0135588.

[28] Dunfield, Kristen A., and Valerie A. Kuhlmeier. "Classifying prosocial behavior: Children's responses to instrumental need, emotional distress, and material desire." Child Development 84.5 (2013): 1766-1776.

[29] 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.

[30] Woo, B. M., & Spelke, E. S. (2023). Toddlers' social evaluations of agents who act on false beliefs. Developmental Science.

[31] Woo, B. M., & Spelke, E. S. (2022). Eight-month-old infants' social evaluations of agents who act on false beliefs. Proceedings of the 44th Annual Meeting of the Cognitive Science Society.

[32] Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief?. British Journal of Developmental Psychology, 30(1), 156-171.

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