Cognition 235 (2023) 105406

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Commonsense psychology in human infants and machines

Gala Stojnić [a], Kanishk Gandhi [b], Shannon Yasuda [a], Brenden M. Lake [a,c], Moira R. Dillon [a,*]

[a] Department of Psychology, New York University, New York, NY, USA
[b] Department of Computer Science, Stanford University, Palo Alto, CA, USA
[c] Center for Data Science, New York University, New York, NY, USA

ARTICLE INFO

Keywords:
Intuitive psychology
Commonsense psychology
Action understanding
Infancy
Machine common sense
Artificial intelligence

ABSTRACT

Human infants are fascinated by other people. They bring to this fascination a constellation of rich and flexible expectations about the intentions motivating people's actions. Here we test 11-month-old infants and state-of-the-art learning-driven neural-network models on the "Baby Intuitions Benchmark (BIB)," a suite of tasks challenging both infants and machines to make high-level predictions about the underlying causes of agents' actions. Infants expected agents' actions to be directed towards objects, not locations, and infants demonstrated default expectations about agents' rationally efficient actions towards goals. The neural-network models failed to capture infants' knowledge. Our work provides a comprehensive framework in which to characterize infants' commonsense psychology and takes the first step in testing whether human knowledge and human-like artificial intelligence can be built from the foundations cognitive and developmental theories postulate.

The early-developing ease with which infants know about people (Gergely, Nádasdy, Csibra, & Bíró, 1995; Woodward, 1998), objects (Spelke, 1990; Stahl & Feigenson, 2015), and places (Hermer & Spelke, 1994) is impressive, especially compared with the difficulties machines have had in achieving these simple human competencies (Lake, Ullman, Tenenbaum, & Gershman, 2017; Marcus & Davis, 2019). Such differences between human and artificial intelligence (AI) are critical to address if we aim to create commonsense AI, leading to AI that we better understand and that better understands us.

One of the general challenges of building commonsense AI is deciding what knowledge to start with. A human infant's foundational knowledge is limited, abstract, and reflects our evolutionary inheritance, yet it can accommodate any context or culture in which that infant might develop (Spelke, 2022; Spelke & Kinzler, 2007). If an aim of AI is to build the flexible, commonsense thinker that human adults become, then machines might need to start like adults do, from the same core abilities as infants, whether achieved through learning-driven or engineered approaches (Botvinick et al., 2017).

Over the past several decades, foundational research on infants' commonsense psychology, i.e., infants' understanding of the intentions, goals, preferences, and rationality underlying agents' actions, has suggested that infants attribute goals to agents and expect agents to pursue goals in rationally efficient ways (Baillargeon, Scott, & Bian, 2016; Gergely et al., 1995; Spelke, 2022; Woodward, 1998). The predictions that support infants' commonsense psychology are foundational to human social intelligence (Banaji & Gelman, 2013; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016) and could thus inform better commonsense AI, but these predictions are typically missing from machine-learning algorithms, which instead predict actions directly (e.g., churn, clicks, likes, etc.; Griffiths, 2015), and therefore lack flexibility to new contexts and situations.

Nevertheless, research on infants' commonsense psychology has not yet been evaluated in a framework that could be directly tested against machines'—let alone built into them—because of non-scalable stimuli, varied task demands, isolated questions, and mixed results. For example, experiments on infants' commonsense psychology have exemplified agents and their actions using various displays, from live human actors reaching for everyday objects (Woodward, 1998), to live puppets with or without animate features like eyes or fur (Johnson, Slaughter, & Carey, 1998), to highly minimal animations of simple shapes navigating in 2D or 3D worlds (Csibra, Bíró, Koós, & Gergely, 2003; Csibra, Gergely, Bíró, Koós, & Brockbank, 1999). These experiments have also typically focused on individual questions of, e.g., goal (Woodward, 1998) or rationality (Gergely et al., 1995) attribution, although some work has probed, for example, how infants' inferences about goals and rationality might combine to support notions of consistency, cost, or value (Liu, Ullman, Tenenbaum, & Spelke, 2017; Scott & Baillargeon, 2013).

Different accounts of infants' knowledge about agents have

* Corresponding author.
E-mail address: moira.dillon@nyu.edu (M.R. Dillon).

https://doi.org/10.1016/j.cognition.2023.105406
Received 8 September 2022; Received in revised form 8 February 2023; Accepted 9 February 2023
Available online 16 February 2023

suggested that this knowledge: coheres as a unified set of abstract concepts of causal efficacy, efficiency, goal-directedness, and perceptual access (Spelke, 2022); reflects infants' intuitive understanding of agents' mental states, which direct their efficient actions consistent with their mental states (Baillargeon et al., 2015; Baillargeon et al., 2016); or emerges from individual achievements rooted in infants' own action experience (Woodward, 2009; Woodward, Sommerville, & Guajardo, 2001). From this rich experimental and theoretical tradition thus arises the need for a comprehensive framework in which to characterize infants' knowledge of agents with results on one task comparable with those on another and with results on the suite of tasks comparable across infants and machines. Such a framework can inform both theories of infants' knowledge and the future of human-like AI.

Here we take a critical step in addressing this need. We provide a comprehensive framework for testing infants' commonsense psychology by assessing infants' performance on the "Baby Intuitions Benchmark (BIB)," a suite of six tasks probing commonsense psychology. BIB was designed expressly to allow for testing both infant and machine intelligence alike (Gandhi, Stojnic, Lake, & Dillon, 2021), and fulfilling that intention, here we also directly compare the performance of infants and machines, providing an empirical foundation for building human-like AI.

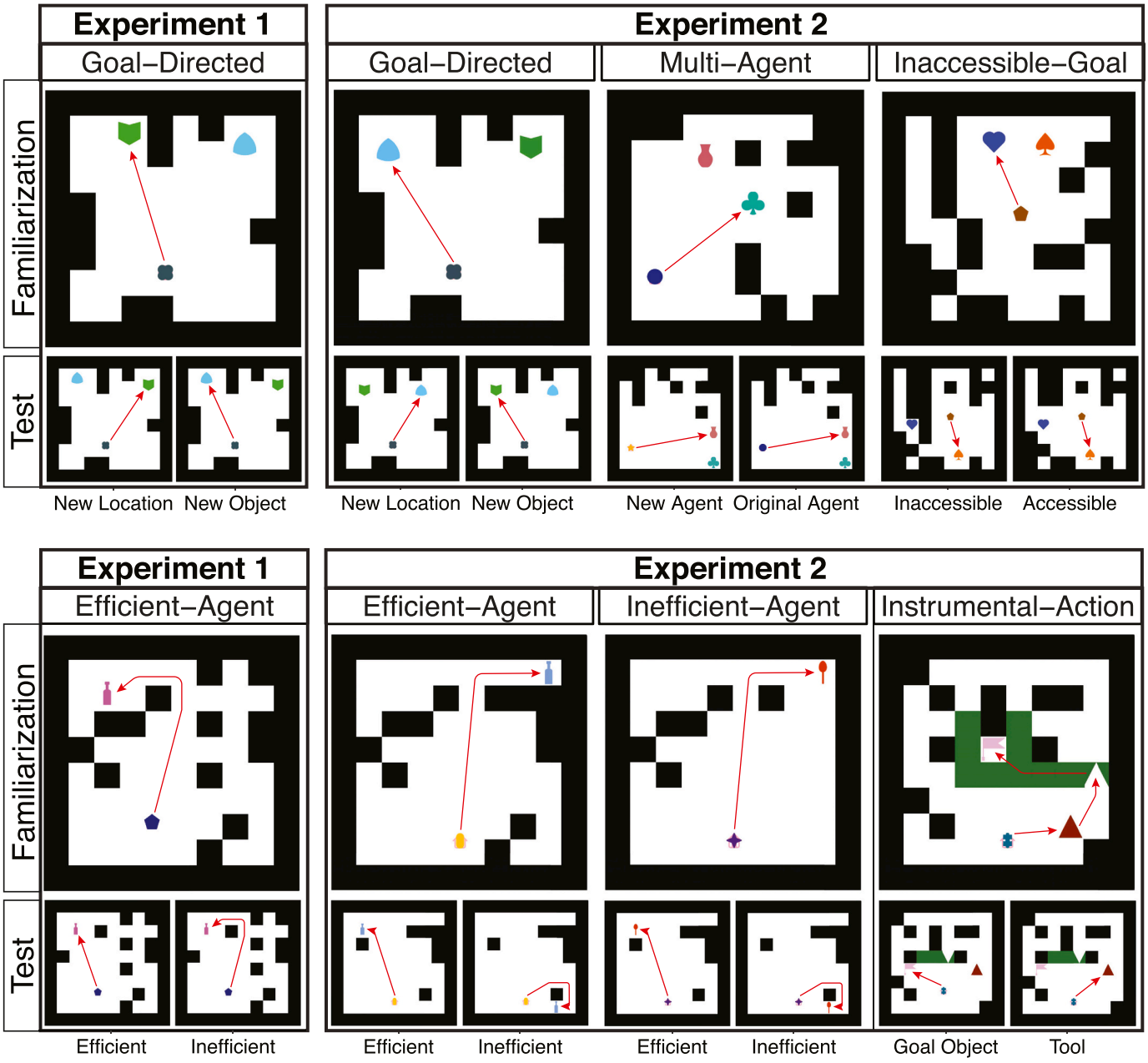

Fig. 1. Schematic of BIB's six tasks used in Experiments 1 & 2 (see also **Fig. S1**). For each task, observers first see eight familiarization trial videos in which an agent acts consistently in terms of its goals, rationality, or instrumentality. The exact make-up of the grid world and the movement of the agent may vary across trials, as described in the main text and **SI**. One example still image per task from a familiarization trial video is shown here. Observers then see expected and unexpected test trial videos (with the order of these trials varying for infants). Example still images of both test trial videos per task are shown here. All of the videos are available at: https://osf.io/r98je/.

1. General methods

1.1. Materials

BIB's tasks include short silent animated videos with simple visuals (Heider & Simmel, 1944), like basic shapes without eyes or limbs, undertaking basic movements in a grid world (Figs. 1 and **S1**). This design allowed for the stimuli's scalable procedural generation, which is required for testing machine-learning algorithms, and emphasized the high-level properties of agents (Csibra et al., 1999; Gao, McCarthy, & Scholl, 2010; Johnson & Gilmore, 2003; Meltzoff, 1995), which challenges the limits and abstraction of an observer's inferential capacity (Kominsky, Lucca, Thomas, Frank, & Hamlin, 2022). This design also presented a novel, overhead navigational context, which required an assumption of agents' full observability of the grid world and its contents (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Luo & Baillargeon, 2007; Luo & Johnson, 2009; Rabinowitz, Perbet, Song, Zhang, & Botvinick, 2018).

Importantly, all of BIB's tasks are presentationally consistent, allowing for comparisons across tasks, without concerns of attributing null effects to varying visual, memory, or other task demands. Instead of focusing on one principle of commonsense psychology, moreover, BIB's tasks focus on three possible attributions to agents' actions that an observer could make—goal attribution, rationality attribution, and instrumentality attribution—thereby addressing whether and how such principles of commonsense psychology might cohere.

Using BIB's environment (Gandhi et al., 2021), we procedurally generated the video stimuli to test infants and computational models and chose the clearest examples of the particular principles of commonsense psychology targeted by each task (Figs. 1 and **S1**). The first three tasks focus on an observer's attribution of goals to agents' actions. The *Goal-Directed Task* captures the idea that agents' goals are directed towards objects, not locations. Observers watch an agent repeatedly move to the same one of two objects in approximately the same location in an unchanging grid world during familiarization. At test, observers may be more surprised when the agent moves to a new object in that grid world after the locations of the two objects switch (Woodward, 1998). The *Multi-Agent Task* asks whether goals are specific to agents. Observers watch an agent move to the same one of two objects during familiarization in a changing grid world, with both objects appearing in varying locations. At test, observers may be more surprised when the original agent versus a new agent moves to a new object (Buresh & Woodward, 2007; Repacholi & Gopnik, 1997). The *Inaccessible-Goal Task* asks whether agents might form new goals when their existing goals become unattainable. Observers watch an agent move to the same one of two objects during familiarization in a changing grid world, with both objects appearing in varying locations. At test, the grid world changes again such that the agent's goal object becomes physically inaccessible. Observers may be more surprised when the agent moves to a new object when its prior goal object is accessible versus inaccessible (Luo & Baillargeon, 2007; Scott & Baillargeon, 2013).

The next two tasks focus on an observer's attribution of rationality to agents' actions. The *Efficient-Agent Task* captures the idea that agents act rationally to achieve goals. Observers watch an agent move to an object efficiently around obstacles in an unchanging grid world during familiarization. At test, the object appears in a location that it had appeared during familiarization, but the grid world has changed such that the obstacles that blocked the object are gone or have been replaced with different obstacles (Gergely et al., 1995; Liu & Spelke, 2017). Observers may be more surprised when the agent moves along a familiar but now inefficient path to the object. The *Inefficient-Agent Task* asks what expectations observers have about agents who initially move inefficiently in a changing grid world. During familiarization, observers watch an agent move along the same paths to an object as the agent in the *Efficient-Agent Task*, but this time there are no obstacles in the agent's way, so the agent's movements to the object are inefficient. At test, the environment changes as in the *Efficient Agent Task*. Observers may either be more surprised when the agent continues to move inefficiently to the object (Liu & Spelke, 2017) or may have no expectations about whether that agent will move efficiently or inefficiently to the object (Gergely et al., 1995).

The last task focuses on an observer's attribution of instrumentality to agents' actions. The *Instrumental-Action Task* captures the idea that agents should only take instrumental actions when necessary. During familiarization, observers watch an agent move first to a key, which it uses to remove a barrier around an object in varying locations, and then to that object. At test, observers may be more surprised when the agent continues to move to the key, instead of directly to the object, when the barrier is no longer blocking the object (Sommerville & Woodward, 2005; Woodward & Sommerville, 2000).

All of the stimuli videos are available at: https://osf.io/r98je/, and additional details about each task are included in the **SI**.

BIB's task structure adopts the "violation-of-expectation" looking-time paradigm often used to test infants (Spelke, 1985; Téglás et al., 2011). Observers see a series of familiarization trials that serve to set up an expectation followed by an expected outcome that is perceptually dissimilar to the familiarization but is conceptually consistent and an unexpected outcome that is perceptually similar to the familiarization but is conceptually surprising. This task structure has been used in recent machine-learning benchmarks focusing on common sense (Piloto, Weinstein, Battaglia, & Botvinick, 2022; Shu et al., 2021; Smith et al., 2019) and is advantageous because it both protects against low-level heuristic-based solutions (Spelke, 1985) and allows for an algorithm's quantitative measure of surprise to be compared with a well-established psychological measure of surprise (Piloto et al., 2022; Stahl & Kibbe, 2022).

2. Infant methods

2.1. Infant design and analyses

In Experiment 1, we collected infants' responses to two of BIB's six tasks, the *Goal-Directed Task* and the *Efficient-Agent Task*. Mixed-model linear regressions with raw looking time as the dependent variable, outcome (expected versus unexpected) as a fixed effect, and participant as a random-effects intercept evaluated infants' performance on each task, and an additional regression examined infants' overall performance across both tasks. To obtain *p*-values, we ran Type 3 Wald tests on the results of each regression. Experiment 1 focused on these two tasks because the common sense they measured has had consistent findings in the prior literature on infants' action understanding (Baillargeon et al., 2016; Gergely & Csibra, 2003; Spelke, 2022; Woodward, 2009). Experiment 1 thus aimed to provide initial evidence of infants' commonsense psychology, as elicited by BIB's highly minimal displays, in BIB's fully observable, overhead navigational context, and with BIB's multiple tasks presented to infants online.

Experiment 2 followed a preregistered design and analysis plan (https://osf.io/p6kba) with replications of the two tasks in Experiment 1 with several improvements, including: automated trial progression; balancing of the side of the goal object across participants in the *Goal-Directed Task*; and matching of the test-trial lengths within participants in the *Efficient-Agent Task*. Infants were tested on these two tasks as well as on BIB's other four tasks outlined above that were not included in Experiment 1.

Following Experiment 1, Experiment 2 evaluated infants' performance on each task with planned mixed-model linear regressions and Type 3 Wald tests with raw looking time as the dependent variable, outcome (expected versus unexpected) as a fixed effect, and participant as a random-effects intercept. Additional planned regressions examined infants' overall performance across all six tasks and directly compared their performance on the two tasks focused on agents' rational actions.

2.2. Infant participants

In Experiment 1, typically developing 11-month-old infants ($N = 26$, $M_{age} = 11.13$ months, *Range* = 10.42 months – 11.83 months; 12 girls) born at ≥37 weeks gestational age were included. They completed the *Goal-Directed Task*, the *Efficient-Agent Task*, or both, with half of the infants receiving each task first, totaling $N = 48$ individual testing sessions and $N = 24$ sessions per task. An additional four sessions were excluded because infants did not complete the session.

In Experiment 2, typically developing 11-month-old infants ($N = 58$, $M_{age} = 11.06$ months, *Range* = 10.50 months – 11.50 months; 31 girls) born at ≥37 weeks gestational age were included. Each infant completed at least one of BIB's tasks, totaling $N = 288$ individual testing sessions. Following our preregistration, data collection stopped when 32 infants ($M_{age} = 11.09$ months, *Range* = 10.50 months – 11.50 months; 17 girls) completed all six of BIB's tasks. Tasks were presented in a semi-randomized order using 32 fixed orders that averaged to each task being presented 5.33 times in each ordinal position (*range*: 4–7 times). All included sessions for each task contributed to the analyses reported here. The final sample sizes for each task were: *Goal-Directed Task*, N = 48; *Multi-Agent Task*, $N = 49$; *Inaccessible-Goal Task*, $N = 47$; *Efficient-Agent Task*, N = 47; *Inefficient-Agent Task*, N = 49; *Instrumental-Action Task*, $N = 48$. The results from the 32 infants who completed all six of BIB's tasks were consistent with the results reported here and so are reported in the **SI**.

An additional 37 sessions were excluded because of preregistered exclusion criteria, including: looking time < 1.5 s to least one test trial and/or two familiarization trials with or without the infant completing the session (16); poor video quality and/or technical failure (18); and caretaker interference (3). An additional two sessions were excluded post hoc for extreme values (> 40 s) to one test outcome, which could artificially inflate the calculation of the sample's variance. These extreme values were identified through examination of a histogram of the raw looking times across all of the sessions and across all of the tasks by two researchers masked to the task and outcome represented by each value. Exclusions were consistent across tasks: *Goal-Directed Task*, 5; *Multi-Agent Task*, 6; *Inaccessible-Goal Task*, 9; *Efficient-Agent Task*, 7; *Inefficient-Agent Task*, 5; *Instrumental-Goal Task*, 7. The total exclusion rate was 11.9%.

Participating families received a $5 Amazon gift card after each testing session and received a bonus gift card of $30 if they completed all six sessions. Prior to participation in session one, we obtained informed consent from the infant's legal guardian, and we confirmed consent before each subsequent session. The use of human participants for this study was approved by the Institutional Review Board on the Use of Human Subjects at our university.

2.3. Infant procedure

Infants were tested online on Zoom. In the first ten minutes of the first testing session, the experimenter explained to caretakers the instructions for setting up their device and for positioning the infant in front of the screen. We asked caretakers to close their eyes and not communicate with the infant during the stimuli presentation. The experimenter, masked to what trial was being presented and the order of the test trials, coded infants' looking to the stimuli live from the start of each video and controlled the progression of stimuli using PyHab (Kominsky, 2019) and slides.com. Each trial video was preceded by a 5 s attention grabber (a swirling blob accompanied by a chiming sound, centered on the screen) to focus the infant's attention to the screen, and each video froze after the agent reached an object. The last frame of the video remained on the screen until infants looked away for 2 s consecutively or for a maximum of 60 s. Testing sessions were recorded through the Zoom recording function, capturing both the infant's face and the screen presenting the stimuli.

Following our preregistration, a different researcher, masked to the study outcome, what trial was being presented, and the order of the test trials, recoded 48 randomly chosen sessions (25%) from the 32 infants who completed all six tasks. The reliability between the first and second coder was very high (ICC = 0.98).

3. Infant results

Infants' performance on Experiment 1's two tasks is displayed in Fig. 2**.** Infants' looking time varied by task, with longer looking to the *Efficient-Agent* versus *Goal-Directed Task* ($F(1, 71) = 9.34$, $p = .003$), reflecting the longer test-trial lengths in the *Efficient-Agent Task* (see **SI**). Overall, infants looked longer to the unexpected versus expected outcomes ($F(1, 66) = 11.34$, $p = .001$), and there was no task by outcome interaction ($F(1, 66) = 0.30$, $p = .585$). Infants were surprised (looked longer) when an agent moved to a new object in the *Goal-Directed Task* ($F(1, 23) = 4.73$, $p = .040$), and they were surprised when an efficient agent later took an inefficient path to an object in the *Efficient-Agent Task* ($F(1, 23) = 2.60$, $p = .016$).

Infants' performance on Experiment 2's six tasks is also displayed in Fig. 2. Infants' looking time varied by task ($F(5, 341) = 2.78$, $p = .018$), reflecting the different test-trial lengths of the different tasks (see **SI**). Overall, infants did not look longer to unexpected versus expected outcomes ($F(1, 341) = 2.27$, $p = .133$), but a task by outcome interaction suggested that different tasks elicited different patterns of infants' looking ($F(5, 341) = 2.23$, $p = .051$).

We first considered infants' performance on Experiment 2's three tasks that focused on goal attribution: the *Goal-Directed*; *Multi-Agent*; and *Inaccessible-Goal Tasks*. First, consistent with the results in Experiment 1, infants were surprised when an agent moved to a new object in the *Goal-Directed Task* ($F(1, 47) = 4.09$, $p = .049$). Infants presented with a new agent in the *Multi-Agent Task*, however, did not show a difference in surprise when that agent versus the original agent moved to a new object ($F(1, 48) = 3.41$, $p = .071$; with longer looking times to the expected outcome). Infants in the *Inaccessible-Goal Task* also did not show a difference in surprise when an agent moved to a new object when its goal object was accessible versus inaccessible ($F(1, 46) = 0.02$, $p = .891$).

We next considered infants' performance on the two tasks that focused on rationality attribution: the *Efficient-Agent* and *Inefficient-Agent Tasks*. First, consistent with the results in Experiment 1, infants were surprised when an efficient agent later took an inefficient path to an object in the *Efficient-Agent Task* ($F(1, 46) = 7.72$, $p = .008$). Infants in the *Inefficient-Agent Tas*k did not show a difference in surprise when an inefficient agent continued to move inefficiently to an object at test ($F(1, 48) = 2.51$, $p = .119$). But, when comparing infants' performance in the *Efficient-Agent* and *Inefficient-Agent Tasks* directly, there was no significant task by outcome interaction ($F(1, 132) = 0.49$, $p = .484$): We did not find evidence that infants' surprise at the inefficient agent's later inefficient action was different from their surprise at the efficient agent's later inefficient action.

Finally, we considered infants instrumentality attribution through their performance on the *Instrumental-Action Task*. Infants did not show a difference in surprise when the agent moved to the tool as opposed to its goal object when the tool was no longer needed to achieve the goal ($F(1, 47) = 0.03$, $p = .853$).

4. Infant discussion

Infants' successful performance in the *Goal-Directed* and *Efficient-Agent Tasks* in both Experiments 1 and 2 suggest that they expect agents' actions to be goal directed towards objects, not locations, and that they expect agents' goal-directed actions to be rationally efficient. These results also show that infants' common sense about the underlying causes of agents' actions are accessible when testing infants online and are highly abstract: Infants' expectations are elicited by BIB's minimal displays and are generalizable to BIB's novel, overhead navigational context.

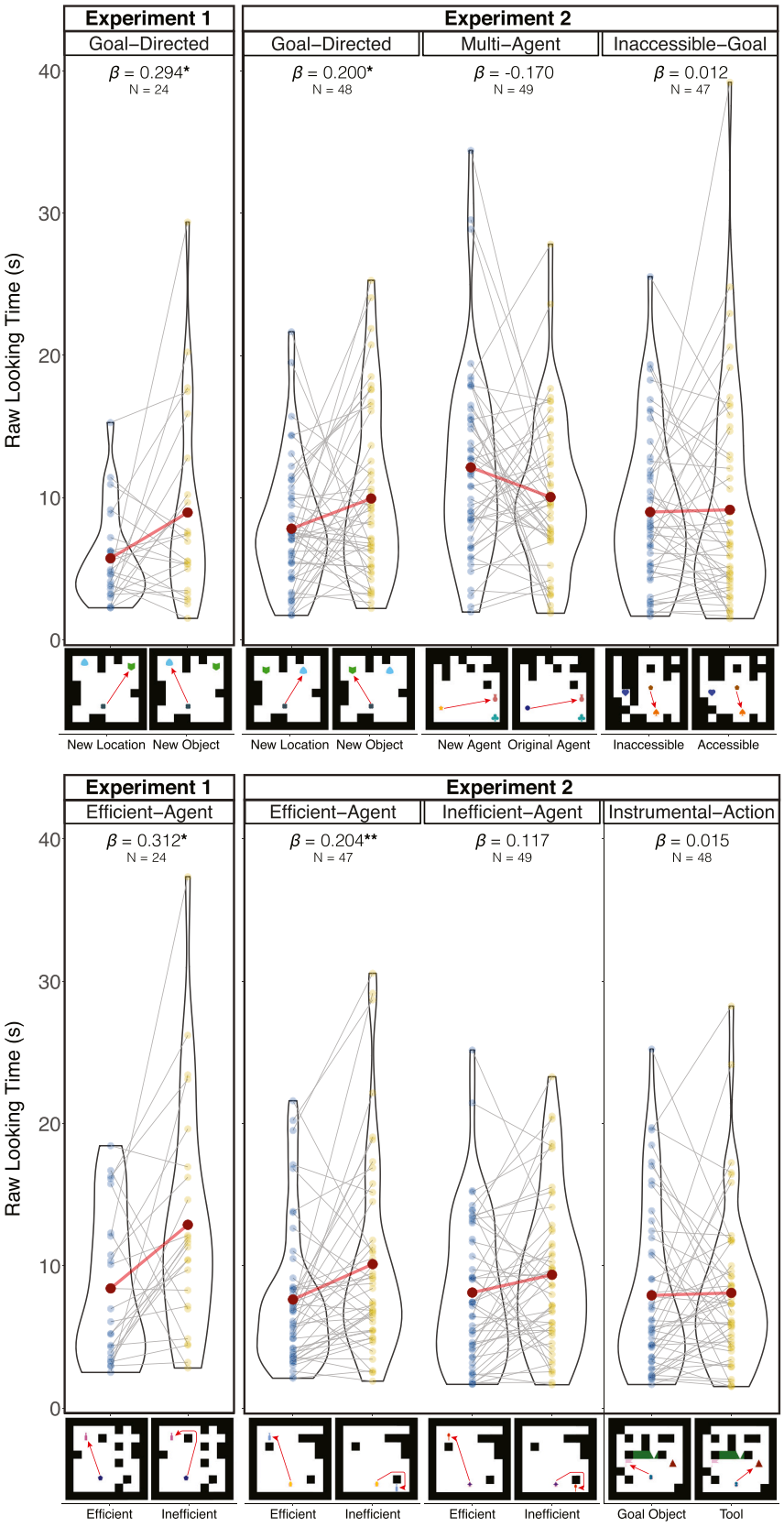

Fig. 2. Infants' raw looking times to the two outcomes in each of BIB's tasks in Experiments 1 & 2. Gray lines connect the individual looking times (represented by blue and yellow dots) of each infant to each outcome. Red dots connected by red lines indicate the mean looking times to each outcome for each task. Beta coefficients are effects sizes in terms of standard deviations, and statistical analyses are reported in the main text (*$p < .05$, **$p < .01$). (For interpretation of the references to colour in this figure legend, please see the online version.)

This latter suggestion is especially striking given infants' success on the *Efficient-Agent Task* since obstacles in the grid world blocked an agent's direct access to the goal object. Given infants sensitivity to and use of agents' perceptual access to objects when making inferences about agents' actions (Luo & Baillargeon, 2007; Luo & Johnson, 2009), infants evidently appreciated BIB's blocking obstacles as only physical, not perceptual. With BIB's context providing no information that these obstacles limit an agent's perceptual access, infants may have interpreted the obstacles as something that agents could "see over" or "see through." Future studies could explore how infants appreciate the geometric, physical, and perceptual affordances of such overhead navigational environments.

Infants' pattern of performance on BIB thus enriches our understanding of their commonsense psychology and raises new questions about the abstract principles that might be inherent to that common sense. Building on questions of infants' sensitivity to agents' physical and perceptual access to objects, future versions of the *Goal-Directed Task* could reveal how having an agent move around obstacles to a goal object, instead of taking only straight paths—actions providing additional cues to agency (Johnson, Shimizu, & Ok, 2007; Luo & Baillargeon, 2005)—might bolster infants' goal attribution in that task. Introducing significant changes to the arrangement of obstacles across the familiarization and test environments in the *Goal-Directed Task*, moreover, could explore the effects of context changes on goal attribution (Liu & Spelke, 2017; Sommerville & Crane, 2009). These latter results might also shed light on infants' failures in some of BIB's other tasks. For example, infants may have failed in the *Inaccessible-Goal Task* because the arrangement of obstacles changed from familiarization to test, including in a way that affected one object's physical accessibility. Infants may have found a change in the object's accessibility itself surprising, or they may not have generalized the agent's goal to this new test environment with significantly different physical affordances because they interpreted this change as indicating two different places in which the agent was acting (Sommerville & Crane, 2009). The *Multi-Agent Task* similarly changed the arrangement of obstacles from familiarization to test, although infants may have failed in this task simply because of heightened attention to the new agent, who appeared for the first and only time in the expected outcome (prior studies showing agent-specific goal attribution had presented the new agent in both test outcomes; Buresh & Woodward, 2007).

Changes to the affordances of the environment from familiarization to test may also explain the pattern of findings in the *Inefficient-Agent Task*, which did not differ from the patterns of findings in the *Efficient-Agent Task*. In particular, previous literature suggests both that infants *do not* expect an agent who had previously moved inefficiently to later move efficiently when an obstacle present during familiarization is removed from the test environment (Gergely et al., 1995; Skerry, Carey, & Spelke, 2013) and that infants *do* expect a previously inefficient agent to later move efficiently if the test environment introduces a new obstacle (Liu & Spelke, 2017). The changes in the number and location of the obstacles across the *Inefficient-Agent Task's* familiarization and test environments may have weakly elicited, or elicited in only some infants, this latter, "default" prediction about rationally efficient goal-directed actions for inefficient agents in the *Inefficient-Agent Task* (Liu & Spelke, 2017). Future versions of the *Inefficient-Agent Task* could thus focus specifically on the effects of different kinds of changes in the context and in the environment's affordances on infants' rationality attribution.

Finally, given infants' successes in previous tasks probing their understanding of instrumental actions, infants may have failed in BIB's *Instrumental-Action Task* because they could not understand the tool object's causal efficacy (Sommerville, Hildebrand, & Crane, 2008) or the agent's ultimate goal. Specifically, prior findings suggesting that infants recognize agents' instrumental actions (e.g., the use of a tool) relied on tools whose causal efficacy was familiar to infants (e.g., pulling a cloth to bring a toy within reach; Piaget, 1953; Sommerville & Woodward, 2005) or on novel tools with which infants were first given direct experience (Sommerville et al., 2008). The tool infants saw in the *Instrumental-Action Task* was both novel and not something they were given experience with. Future versions of the *Instrumental-Action Task* might thus introduce state-changes, such as colour changes, to the contacted tools and objects, which, in previous studies, have made the causal efficacy of otherwise novel and inscrutable actions appreciable to young infants (Liu, Brooks, & Spelke, 2019; Skerry et al., 2013).

5. Model methods

5.1. Model design and analyses

To examine whether infants' intelligence about agents might be reflected in state-of-the-art machine intelligence, we compared infants' performance on BIB in Experiment 2 to the performance of three learning-driven neural-network models. Following prior work (Gandhi et al., 2021; Rabinowitz et al., 2018), the models formed predictions about an agent's actions at test based on its actions during familiarization. To obtain a continuous measure of surprise as a correlate of infants' looking time, we calculated the models' prediction error for each frame of each outcome and considered the frame with the maximum error. To compare model and infant performance, we then calculated the *Z*-scored mean surprisal score to each outcome for each model and the Z-scored mean looking time to each outcome for infants. Z-scores were calculated within task. For an unplanned quantitative comparison of the overall similarity between the infants' and each models' performance, we evaluated the root mean squared error (RMSE) across BIB's six tasks using the mean Z-score to the unexpected outcome. We also included a comparison between infants' performance and a "baseline," which we gave a surprisal score of "0" for all tasks.

Finally, to confirm that the models' performance on the specific trials presented to infants was representative of their performance more generally and not due to any idiosyncrasies of the particular videos shown to infants, we also evaluated the models' accuracy on BIB's full dataset (Gandhi et al., 2021). Because those results were consistent with the models' performance on the infant videos and with prior work (Gandhi et al., 2021), they are reported in the **SI**.

5.2. Model specifications

Learning-driven neural network models have accelerated recent advances in AI (Lecun, Bengio, & Hinton, 2015; Rabinowitz et al., 2018), and so we chose to compare such models' performance on BIB to infants'. Approaches like reinforcement learning (Sutton & Barto, 2018) and inverse reinforcement learning (Ng & Russel, 2000), for example, have succeeded in learning to control agents and in understanding the actions of agents, but these approaches cannot be used with BIB because they require privileged information, including the ability to actively control agents in the test environment and, in the case of reinforcement learning, receive a reward. Infants engage with stimuli like BIB's through passive observation, and so we based our modeling on the "Theory of Mind Net (ToMnet)" architecture from Rabinowitz et al. (2018), which is a neural network designed specifically for passive observation that has been shown to make inferences about an agent's underlying mental states from its behavior.

With this architecture, we tested three models from two classes: behavioral cloning (BC) and video modeling (Gandhi et al., 2021). The models' schematized architectures are presented in Figs. 3 and **S2**. Two BC models predicted how an agent would act using the background training as examples of state and action pairs (see *Model Training* below). To predict the agent's next action in a test trial, BC combined information from the learned features from the previous frame of a test-trial video along with the learned features in the set of familiarization-trial videos. Video modeling used a similar strategy, architecture, and training procedure, but it aimed to predict the entire next frame of the

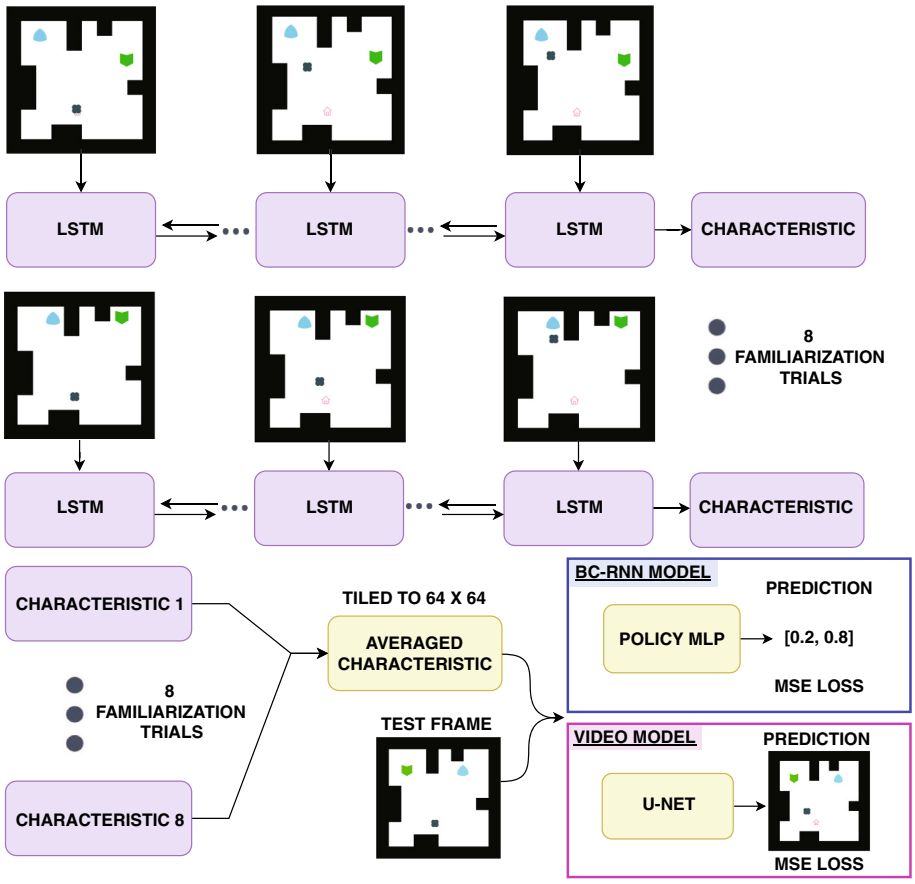

Fig. 3. Architecture of the video and BC RNN models (Gandhi et al., 2021; Rabinowitz et al., 2018). An agent-characteristic embedding was inferred from the familiarization trials using a recurrent net. This embedding, with a frame from the test trial, was used to predict the next action of the agent in case of the BC model and the next frame of the video using a U-net (Ronneberger, Fischer, & Brox, 2015) in the case of the video model.

test-trial video rather than just the agent's next action.

The two BC models differed in their encoding of the familiarization trials. One BC model relied on a simple multi-layer perceptron (MLP) to encode pairs of states and actions independently (**Fig. S2**), and the other BC model relied on a more complex, bi-directional recurrent neural network (RNN) to sequentially encode pairs of states and actions (Fig. 3). The states were encoded with a convolutional neural network (CNN), which was pretrained using Augmented Temporal Contrast (ATC) (Stooke, Lee, Abbeel, & Laskin, 2020). **Table S1** provides the CNN specifications and the ATC data augmentation details. For both the MLP and RNN encoders, the model obtained a characteristic embedding (Rabinowitz et al., 2018) of an agent by first aggregating the embeddings across frames (using the average for the MLP and the last step for the RNN) for each familiarization trial and by second averaging across familiarization trials. When aggregating frames, the videos were randomly sub-sampled to use up to 30 frames. To predict the future actions of the agent, defined as the continuous change in position based on the video (at 3 frames per second), the models combined the characteristic embedding with the current state of the environment (also encoded with the CNN). See **Table S2** for the specifications of the BC models.

The one video model sequentially encoded each familiarization trial by passing up to 30 frames through a CNN and then combining them with a bi-directional RNN. The model obtained a characteristic embedding of an agent by averaging the RNN embeddings. The model combined the characteristic embedding with the current state of the environment (specified by the current frame of the video) to predict the next frame of the video (at 3 frames per second) using a U-net architecture (Ronneberger et al., 2015).

5.3. Model training

Prior to being tested, the models were trained on thousands of background examples provided by the BIB dataset (Gandhi et al., 2021) of BIB-like agents exhibiting simple behaviors in a grid world. While the training set included individual components of the test set (e.g., agents' movement to objects, agents' consistent object goals, barriers, tools, etc.; see below), success on the test set required models to flexibly combine representations across the different training tasks. Moreover, since training included only expected outcomes, training with labeled videos was not possible. The training otherwise used the same familiarization/test task design as the test set.

In one training task, an agent moved to one object in varying locations in the grid world. In a second training task, two objects were presented in varying locations in the grid world but always very close to the agent; the agent consistently moved to one of the two objects. In a third training task, the agent moved to one object in varying locations in the grid world; at varying points during the familiarization, that agent was substituted by another agent. Finally, in a fourth training task, a green barrier surrounded an agent and a key; the agent retrieved the key to let itself out of the blocked area to move to an object.

We included five runs of each model type with the runs initialized randomly and trained until they converged on the background training. The BC models were trained to minimize mean squared error, and the video model was trained to minimize mean squared error in pixel space. Twenty percent of the background training trials were left out as a validation set, and the models were successful at the validation set in predicting agents' actions on all of the background training tasks, with low prediction errors. For example, the MSE error for the BC models on

the validation set was about 0.03 which is 0.8% of the maximum possible prediction error (4.00). The only exception was that the BC RNN model performed an order of magnitude less well compared to the BC MLP model on the training task in which two objects were presented very close to the agent and the agent consistently moved to just one (see **SI**).

6. Model results

Fig. 4 displays the *Z*-scored means of the models' surprisal scores to the expected and unexpected outcomes for each task (see **SI** for additional details). The Z-scored means of infants' looking times in the tasks of Experiment 2 are also displayed. Model performance shows little resemblance to infant performance.

First, to evaluate machines' goal attribution relative to infants', we compared infants and models on the *Goal-Attribution Task*. Unlike infants, who attributed to agents goal objects, not goal locations, the models either attributed to agents goal locations (BC MLP) or neither goal objects nor goal locations (BC RNN, video model). Next, to evaluate machines' rationality attribution relative to infants', we compared infants and models on the *Efficient-Agent* and *Inefficient-Agent Tasks*. While models attributed rational action to agents in the *Efficient-Agent Task* (to an even greater degree than did infants), models did not attribute rational action to previously inefficient agents who act in new environments in the *Inefficient-Agent Task*. Here the models' performance was nearly orthogonal to infants', who did attribute rational action to previously inefficient agents who act in new environments.

The comparisons between machine and infant performance on BIB's other three tasks revealed no instances in which the models demonstrated positive predictions about agents' actions missing from infants' predictions. In particular, while infants' may have been relatively more surprised at the appearance of the new agent in the expected outcome of the *Multi-Agent Task*, as described above, the models did not show a difference in surprise across the two outcomes. In the *Inaccessible-Goal Task*, the video model did appear to be more surprised when the agent moved to a new object when its goal object was accessible, unlike the infants, but given this model's failure on the *Goal-Directed* and *Multi-Agent Tasks*, its performance is unlikely to reflect an understanding of agents' goal-directed actions towards objects. For example, the model may have learned that the obstacles in the grid world block objects and that agents move to objects. This would lead to a lower surprisal score when an agent moved to the one accessible object compared with when it moved to either one of the accessible objects. Similarly, in the *Instrumental-Action Task* the models seemed to have succeeded where the infants did not, showing greater surprise when the agent moved to the key when it was unnecessary to do so. But, closer investigation of the models' performance shows that this apparent success is limited to test trials in which the green barrier was absent versus present and inconsequential (see **SI**). A true understanding of instrumental actions would generalize across the presence or absence of the green barrier at test. The models thus did not understand agents' instrumental actions.

Finally, the RMSE analysis revealed high values for all infant and model comparisons: BC RNN: 0.319; BC MLP: 0.492; video model: 0.297, suggesting little similarity between infant and model performance. Indeed, these RMSE values were higher than the one obtained by comparing infants' performance to "baseline" surprisal scores of "0" for all tasks: 0.143.

7. Model discussion

BIB was expressly designed to allow for testing both infant and machine intelligence alike (Gandhi et al., 2021), providing an empirical foundation for building human-like AI. While the performance of the models tested here has not previously been compared with human performance (let alone with infant performance), and while and models like these are limited in their capacity for flexible generalization to out-of-distribution novel test displays compared with the displays used for their training (a generalization BIB requires and infants excel at), such models have nevertheless accelerated recent advances in AI (Lecun et al., 2015; Rabinowitz et al., 2018). Our comparison reveals that the state-of-the-art "machine theory of mind" captured in such models is indeed missing key principles of commonsense psychology that infants possess.

In particular, while infants expect agents' goal-directed actions to be towards objects, not locations, models either have no expectations or expect those actions to be towards locations, not objects. And, while infants expect both previously efficient and inefficient agents to exhibit rational and efficient goal-directed actions towards objects in new environments, models only expect previously efficient agents to act efficiently in new environments. Finally, where we were unable to find any predictions that infants might have about the goals of new agents, about agents' goal objects in new environments, or about novel instrumental actions, models show no additional commonsense psychology.

Our approach of directly comparing infant and machine intelligence allows us to specify what principles of commonsense psychology are present in infants yet missing in machines, thereby inspiring new directions in engineering AI. For example, alternative models based on Bayesian inverse planning have been applied successfully to tasks like BIB by making more explicit abstract inferences about mental states (Baker et al., 2017; Baker, Saxe, & Tenenbaum, 2009; Shu et al., 2021). Nevertheless, extending the Bayesian approach to BIB in particular and to videos in general is not straightforward: A video format does not by itself provide the identification of the agents or objects present in the scene (let alone any relations among them). Recent approaches based on inverse reinforcement learning (Sim & Xu, 2019; Yu, Yu, Finn, & Ermon, 2019) could also be promising, but, as reviewed above, they require online, active sampling from the testing environment, and BIB's environment, like much of infants' experience, involves passive viewing. It thus remains an open challenge for learning-driven systems to acquire sufficiently rich, abstract structure from BIB's training to match infant commonsense intelligence. Nevertheless, setting infant common sense as a benchmark for machine common sense promises to give AI the foundations of human intelligence.

8. General discussion

BIB includes six highly minimal but presentationally consistent tasks focusing on three high-level principles of commonsense psychology: goal attribution; rationality attribution; and instrumentality attribution. Infants' successes on BIB suggest they have a highly abstract notion of agents' actions as goal-directed towards objects and a principle of rationality that leads to default expectations of agents' efficient actions towards goals. These results are consistent with the rich literature on infants' commonsense psychology (Baillargeon et al., 2015; Baillargeon et al., 2016; Spelke, 2022; Woodward, 2009; Woodward et al., 2001) and synthesize the literature's findings in a unified framework that can be directly compared with—and perhaps built into—machine intelligence. In addition, BIB uniquely reveals that infants appreciate agents' actions in a novel, overhead navigational context, here recognizing obstacles as physical but not perceptual barriers to action.

Infants' failures on BIB suggest that changes to the contexts in which goals are first demonstrated may have significant impacts on infants' goal and rationality attribution (Liu & Spelke, 2017; Sommerville & Crane, 2009). For example, infants may not generalize an agent's goal to a test environment with even minimal or inconsequential changes relative to the environment in which the goal was initially demonstrated if those changes suggest that agents are acting in a new place. Regardless of how infants might come to understand the geometry of BIB's environment, their sensitivity to and use of *where* an agent is for goal and rationality attribution is apparent. Future studies might thus investigate infants' use of such geometry for recognizing places based on their shape or navigability even before infants can navigate on their own (Deen

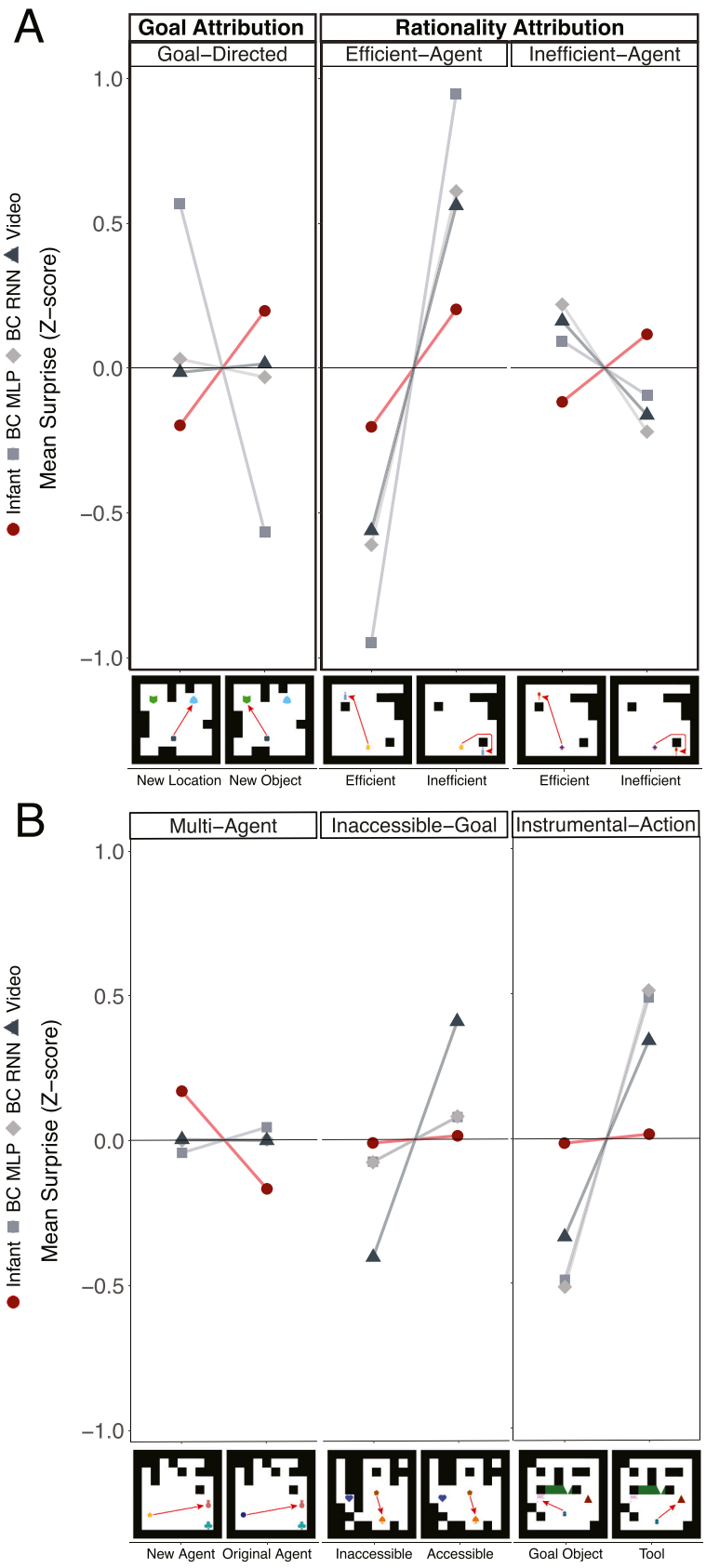

Fig. 4. Z-scored means of the models' surprisal scores (each model is shown with a different shape and in a different shade of gray) and the Z-scored means of infants' looking times (shown in red) to the expected and unexpected outcomes in each of BIB's six tasks in Experiment 2. Models differ from infants in terms of infants' successful goal and rationality attribution (**A**), and models show no additional commonsense psychology missing from infants' performance (**B**). (For interpretation of the references to colour in this figure legend, please see the online version.)

et al., 2017; Kosakowski et al., 2021).

Future work exploring infants' knowledge about the world could extend our general approach to investigate other aspects of infant commonsense psychology. Because BIB's tasks are procedurally generated and presentationally consistent, for example, new tasks could easily be incorporated into BIB's dataset. Future studies might explore expectations of agents' notions of cost and value (Jara-Ettinger et al., 2016; Liu et al., 2017) or recognition of agents' actions that might signal potential social partnerships (Meltzoff, 2007; Powell & Spelke, 2013; Schachner & Carey, 2013; Tomasello, 2018). While we show that learning-driven neural-network approaches already fall short of infant's common sense on BIB's existing tasks, such expectations will nevertheless become increasingly important for AI too as it becomes further embedded in real-world, multi-agent settings that demand common sense. Extending our approach can ultimately inform comprehensive accounts of infants' knowledge not only about agents, but also about objects (Lin, Stavans, & Baillargeon, 2022; Spelke, 1990; Stahl & Feigenson, 2015) and places (Hermer & Spelke, 1994), allowing us to more fully describe the origins and development of human common sense and provide an avenue for building the future of human-like AI.

BIB called for an interanimating research program between developmental cognitive science and artificial intelligence. The present work demonstrates that such a program is both possible and generative for both fields. Our work provides a first step in this productive dialogue between the cognitive and computational sciences to test whether knowledge can be built, in human or machine, from the foundations that cognitive and developmental theories postulate.

Credit author statement

GS, KG, BL, and MRD conceptualized the study. GS, KG, and SY curated the data. KG and SY analyzed the data. MRD and BL acquired funding and supervised the study. MRD wrote the original draft. GS, KG, SY, and BL reviewed and edited the draft.

Data availability

Experiment 2 with infants was preregistered on the Open Science Framework (OSF) prior to data collection, and the preregistration is available at https://osf.io/p6kba. The data, code, and materials related to all of the infant testing and the comparison between infant and machine performance are available on the OSF at: https://osf.io/htjc2/. The code related to the model testing is available at: https://github.com/kanishkg/bib-models.

Acknowledgements

This work was supported by a National Science Foundation CAREER Award (DRL1845924; to MRD) and a DARPA grant on Machine Common Sense (HR001119S0005; to MRD and BML). We thank Eli Mitnick for assistance with data collection, Koleen McCrink, David Moore, Lisa Oakes, and Victoria Romero for their feedback on the project's general aims, and Brian Reilly for his feedback on the project and manuscript. Finally, we thank the generous families who volunteered their time to participate in this research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105406.

References

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology, 67*(April), 159–186. https://doi.org/10.1146/annurev-psych-010213-115033

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K., … Bian, L. (2015). Psychological and sociomoral reasoning in infancy. In *, 1. APA handbook of personality and social psychology, volume 1: Attitudes and social cognition* (pp. 79–150). https://doi.org/10.1037/14341-003

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour, 1*(4), 1–10. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Banaji, M., & Gelman, S. A. (2013). *Navigating the social world: What infants, children, and other species can teach us*. Oxford University Press.

Botvinick, M., Barrett, D. G. T., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., … Hassabis, D. (2017). Building machines that learn and think for themselves. *Behavioral and Brain Sciences, 40*, Article e255. https://doi.org/10.1017/S0140525X17000048

Buresh, J. S., & Woodward, A. L. (2007). Infants track action goals within and across agents. *Cognition, 104*(2), 287–314. https://doi.org/10.1016/j.cognition.2006.07.001

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science, 27*(1), 111–133. https://doi.org/10.1207/s15516709cog2701_4

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition, 72*(3), 237–267. https://doi.org/10.1016/S0010-0277(99)00039-6

Deen, B., Richardson, H., Dilks, D. D., Takahashi, A., Keil, B., Wald, L. L., … Saxe, R. (2017). Organization of high-level visual cortex in human infants. *Nature Communications, 8*. https://doi.org/10.1038/ncomms13995

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.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science, 21*(12), 1845–1853. https://doi.org/10.1177/0956797610388814

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7*(7), 287–292. https://doi.org/10.1016/S1364-6613(03)00128-1

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*(2), 165–193. https://doi.org/10.1016/0010-0277(95)00661-H

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21–23. https://doi.org/10.1016/j.cognition.2014.11.026

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

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature, 370*, 57–59. https://doi.org/10.1038/370057a0

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve utility Calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences, 20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Johnson, M. H., & Gilmore, R. O. (2003). Object-centered attention in 8-month-old infants. *Developmental Science, 1*(2), 221–225. https://doi.org/10.1111/1467-7687.00034

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze following in 12-month-olds. *Developmental Science, 1*(2), 233–238. https://doi.org/10.1111/1467-7687.00036

Johnson, S. C., Shimizu, Y. A., & Ok, S. J. (2007). Actors and actions: The role of agent behavior in infants' attribution of goals. *Cognitive Development, 22*(3), 310–322. https://doi.org/10.1016/j.cogdev.2007.01.002

Kominsky, J. F. (2019). PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behavior and Development, 54*(January), 114–119. https://doi.org/10.1016/j.infbeh.2018.11.006

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. *Cognitive Development, 63*(May), Article 101213. https://doi.org/10.1016/j.cogdev.2022.101213

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2021). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*. https://doi.org/10.1016/j.cub.2021.10.064

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. https://doi.org/10.1017/S0140525X16001837

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lin, Y., Stavans, M., & Baillargeon, R. (2022). Infants' physical reasoning and the cognitive architecture that supports it. In O. Houdé, & G. Borst (Eds.), *Cambridge handbook of cognitive development* (pp. 244–268). Cambridge University Press.

Liu, S., Brooks, N. B., & Spelke, E. S. (2019). Origins of the concepts cause, cost, and goal in prereaching infants. *Proceedings of the National Academy of Sciences of the United States of America, 116*(36), 17747–17752. https://doi.org/10.1073/pnas.1904410116

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition, 160*, 35–42. https://doi.org/10.1016/j.cognition.2016.12.007

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science, 358*(6366), 1038. https://doi.org/10.1126/science.aag2132

Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? - psychological reasoning in 5-month-old infants. *Psychological Science, 16*(8), 601–608. https://doi.org/10.1111/j.1467-9280.2005.01582.x

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition, 105*(3), 489–512. https://doi.org/10.1016/j.cognition.2006.10.007

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science, 12*(1), 142–149. https://doi.org/10.1111/j.1467-7687.2008.00741.x

Marcus, G., & Davis, E. (2019). *Building machines that learn and think like people*. Pantheon Books.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intened acts by 18-month-old children. *Developmental Psychology, 31*(5), 838–850.

Meltzoff, A. N. (2007). "Like me": A foundation for social cognition. *Developmental Science, 10*(1), 126–134. https://doi.org/10.1111/j.1467-7687.2007.00574.x

Ng, A. Y., & Russel, S. (2000). Algorithms for inverse reinforcement learning, *1. International Conference on Machine Learning*.

Piaget, J. (1953). The origins of intelligence in children. *International Journal of Psychoanalysis, 35*, 373–375.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-022-01394-8

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. https://doi.org/10.1073/pnas.1304326110

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., & Botvinick, M. (2018). Machine theory of mind, *10. International Conference on Machine Learning* (pp. 6723–6738).

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33*(1), 12–21. https://doi.org/10.1037/0012-1649.33.1.12

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention, 9*, 16591–16603.

Schachner, A., & Carey, S. (2013). Reasoning about "irrational" actions: When intentional movements cannot be explained, the movements themselves are seen as the goal. *Cognition, 129*(2), 309–327. https://doi.org/10.1016/j.cognition.2013.07.006

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological Science, 24*(4), 466–474. https://doi.org/10.1177/0956797612457395

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., … Gutfreund, D. Ullman (2021). AGENT: A benchmark for core psychological reasoning. *International Conference on Machine Learning*, 9614–9625.

Sim, Z. L., & Xu, F. (2019). Another look at looking time: Surprise as rational statistical inference. *Topics in Cognitive Science, 11*(1), 154–163. https://doi.org/10.1111/tops.12393

Skerry, A. E., Carey, S. E., & Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proceedings of the National Academy of Sciences, 110*(46), 18728–18733. https://doi.org/10.1073/pnas.1312322110

Smith, K. A., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J. B., & Ullman, T. D. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in Neural Information Processing Systems, 32*, 1–11.

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. https://doi.org/10.1111/j.1467-7687.2008.00787.x

Sommerville, J. A., Hildebrand, E. A., & Crane, C. C. (2008). Experience matters: The impact of doing versus watching on Infants' subsequent perception of tool-use events. *Developmental Psychology, 44*(5), 1249–1256. https://doi.org/10.1037/a0012296

Sommerville, J. A., & Woodward, A. L. (2005). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition, 95*(1), 1–30. https://doi.org/10.1016/j.cognition.2003.12.004

Spelke, E. S. (1985). Preferential-looking methods as tools for the study of cognition in infancy. In G. Gottlieb, & N. A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 323–361).

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14*(1), 29–56.

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

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science, 10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science, 348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799

Stahl, A. E., & Kibbe, M. M. (2022). Great expectations: The construct validity of the violation-of-expectation method for studying infant cognition. *Infant and Child Development*. https://doi.org/10.1002/icd.2359

Stooke, A., Lee, K., Abbeel, P., & Laskin, M. (2020). Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*(6033), 1054–1059. https://doi.org/10.1126/science.1196404

Tomasello, M. (2018). How children come to understand false beliefs: A shared intentionality account. *Proceedings of the National Academy of Sciences of the United States of America, 115*(34), 8491–8498. https://doi.org/10.1073/pnas.1804761115

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Woodward, A. L. (2009). Infants' grasp of others' intentions. *Current Directions in Psychological Science, 18*(1), 53–57. https://doi.org/10.1111/j.1467-8721.2009.01605.x

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science, 11*(1), 73–77. https://doi.org/10.1111/1467-9280.00218

Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). Action. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality : Foundations of social cognition* (pp. 149–169). MIT Press.

Yu, L., Yu, T., Finn, C., & Ermon, S. (2019). Meta-inverse reinforcement learning with probabilistic context variables. *Advances in Neural Information Processing Systems, 32*, 1–12.