Title: Commonsense Psychology in Human Infants and Machines

Authors: Gala Stojnic[1], Kanishk Gandhi[2], Shannon Yasuda[1], Brenden M. Lake[1,3], Moira R. Dillon[1]*

* Correspondence to:

Moira R. Dillon, moira.dillon@nyu.edu

Affiliations

[1]Department of Psychology, New York University; New York, NY, USA

[2]Department of Computer Science, Stanford University; Palo Alto, CA, USA

[3]Center for Data Science, New York University; New York, NY, USA

Data Availability Statement

All data, code, and materials related to the infant testing and comparison between infant and machine performance are available on the Open Science Framework at: [ACCESSIBLE TO REVIEWRS]. All code related to the model testing is available at: [ACCESSIBLE TO REVIEWRS].

Abstract

Artificial intelligence (AI) promises to take the flawed intelligence of humans *out of* machines. Why, then, might we want to put the inchoate intelligence of human infants *into* machines? While infants seem to intuit others' underlying intentions merely by observing their actions, AI systems, in contrast, fall short in such commonsense psychology. Here we put infant and machine intelligence into direct dialogue for the first time through their performance on the Baby Intuitions Benchmark (BIB), a comprehensive suite of tasks probing commonsense psychology. Following a preregistered design and analysis plan, we collected 288 individual responses of 11-month-old infants to BIB's six tasks and tested three state-of-the-art learning-driven neural-network models from two different model classes. Infants' performance revealed their comprehensive understanding of agents as rational and goal-directed, but the models failed to capture infants' knowledge. These striking differences between human and artificial intelligence are critical to address to build machine common sense.

Keywords: intuitive psychology; action understanding; infancy; machine common sense; artificial intelligence

Introduction

The early-developing ease with which infants know about objects[1,2], people[3,4], and places[5] is impressive, especially compared to the difficulties machines have had in achieving these simple human competencies[6,7]. 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[8,9], yet it can accommodate any context or culture in which that infant might develop. If an aim of AI is to build the flexible, commonsense thinker that human adults become, then machines might need to start from the same core abilities as infants, whether achieved through learning-driven or engineered approaches[10]. Nevertheless, comprehensively characterizing infants' knowledge is difficult, as behavioral experiments with infants typically focus on just one or two abilities in one content area, rely on diverse materials and methods (e.g., live actors versus animations), and are presented to different groups of infants[11,12].

Here we present a first step in building commonsense AI by putting infant and machine intelligence into direct dialogue for the first time through their performance on the Baby Intuitions Benchmark (BIB), a comprehensive suite of tasks probing commonsense psychology[13]. BIB focuses on an observer's ability to make accurate predictions about agents' underlying intentions merely by observing their actions. These predictions are foundational to human social intelligence[14,15] but are typically missing in AI, which instead predicts the actions directly (e.g., churn, clicks, likes, etc.[16]). BIB includes short silent animated videos presenting simple visuals[17], including basic shapes without eyes or limbs, undertaking basic movements in a grid world. This design allows for procedural generation of BIB's stimuli and emphasizes the high-level properties of agents and objects[18–21], challenging the limits of an observer's inferential capacity and reflecting the kind of abstract knowledge that infants possess. The videos' structure adopts the "violation-of-expectation" looking-time paradigm often used to test infants[22,23] which includes a series of familiarization events that serve to set up an expectation, followed by — in either order — 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 commonsense[24–27] and is advantageous because it both protects against low-level heuristic-based solutions[22] and allows for an algorithm's quantitative measure of surprise to be compared to a well-established psychological measure of surprise[27]. While infants are only presented with videos of the familiarization and test events, models may train on BIB's background training videos[13], which include thousands of examples of BIB-like agents exhibiting simple behaviors in a grid world (e.g., an agent moving to a single object; see **Materials and Methods** and **SI**). Importantly, the test videos require models to generalize outside of the training distribution, combining multiple behaviors that exist in isolation or in a simplified form during training. Because the background training provides only expected outcomes, moreover, supervised learning on labeled videos is not possible.

BIB includes six separate tasks inspired by the rich empirical literature on infants' knowledge about agents (**Figure 1**, see **Materials and Methods** and **SI**). Using BIB's environment[13], we procedurally generated the video stimuli to test infants and machines and chose the clearest examples of the particular principles of commonsense psychology targeted by each task. The *Goal-Directed Task* captures the idea that agents' goals are usually objectdirected. Observers watch an agent repeatedly move to the same one of two objects in approximately the same location in a grid world during familiarization. At test, observers should be more surprised when the agent moves to a new object after the locations of the two objects switch[3]. The *Multi-Agent Task* captures the idea that goals might be specific to agents. Observers watch an agent move to the same one of two objects during familiarization. At test, a new agent appears, and observers should be more surprised when the original agent versus the new agent moves to a new object[28,29]. The *Inaccessible-Goal Task* captures the idea that agents might form new goals when their existing goals are unattainable. Observers watch an agent move to the same one of two objects during familiarization. At test, observers should be more surprised when that agent moves to a new object when its goal object is accessible versus inaccessible[30]. The *Efficient-Agent Task* captures the idea that agents act rationally to achieve goals. Observers watch an agent move efficiently around obstacles to an object during familiarization. At test, observers should be more surprised when that agent moves inefficiently to the object[4]. The *Inefficient-Agent Task* asks what expectations observers have about agents who move inefficiently to objects. Observers watch an agent in the familiarization 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. At test, observers should have no expectations about whether that agent will move efficiently or inefficiently to the object[4,31]. The *Instrumental-Action Task* captures the idea that agents should only take instrumental actions when they are necessary. During familiarization, observers watch an agent move first to a key, which it uses to remove a barrier around an object, and then to that object. At test, observers should 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[32,33].

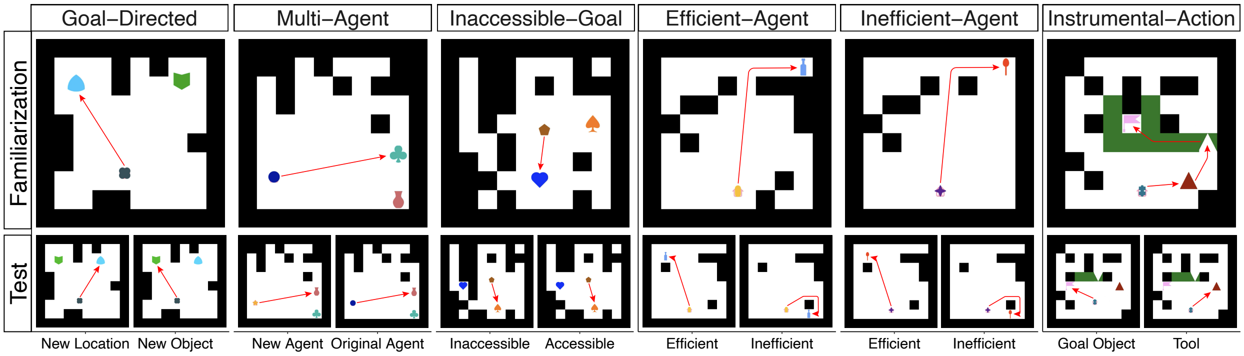

Figure 1. Schematic of BIB's six tasks. For each task, observers first see eight familiarization videos in which an agent acts consistently in terms of its rationality and goal-directedness. The exact make-up of the grid world and the movement of the agent varies across trials, as described in the main text and **SI**. One example familiarization trial per task is shown here. Observers then see an expected and unexpected test video (with the order of these videos varying for infants). Examples of both test trials are shown here. All of the videos are available at: [ACCESSIBLE TO REVIEWRS].

Results

Infant Performance on BIB

Following a preregistered design and analysis plan [ACCESSIBLE TO REVIEWRS], we collected 288 individual responses of 11-month-old infants to BIB's six tasks (*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; see **Materials and Methods**). Planned 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. 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. Unplanned, post hoc analyses explored the effects of the order in which infants saw certain tasks and their looking during a task's familiarization phase. Additional planned and unplanned analyses are reported in the **SI** along with the results of our pilot test of infants on versions of the *Goal-Directed* and *Efficient-Agent Tasks*.

Infants' performance on BIB's six task is displayed in **Figure 2**. Infants' looking time varied by task ($F(5, 341) = 2.78$, $p = .018$; reflecting the different test-trial lengths of the different tasks, **SI**), and, overall, infants did not look longer to unexpected versus expected outcomes ($F(1, 341) = 2.27$, $p = .133$). A task by outcome interaction suggested that, overall, different tasks elicited different patterns of infants' looking ($F(5, 341) = 2.23$, $p = .051$). Because BIB's tasks are presentationally consistent, focusing on several components of human reasoning about agents, and because we presented its tasks to some of the same infants, our findings, coupled with the rich existing literature on infants' action understanding, paint a comprehensive picture of infants' commonsense psychology about agents.

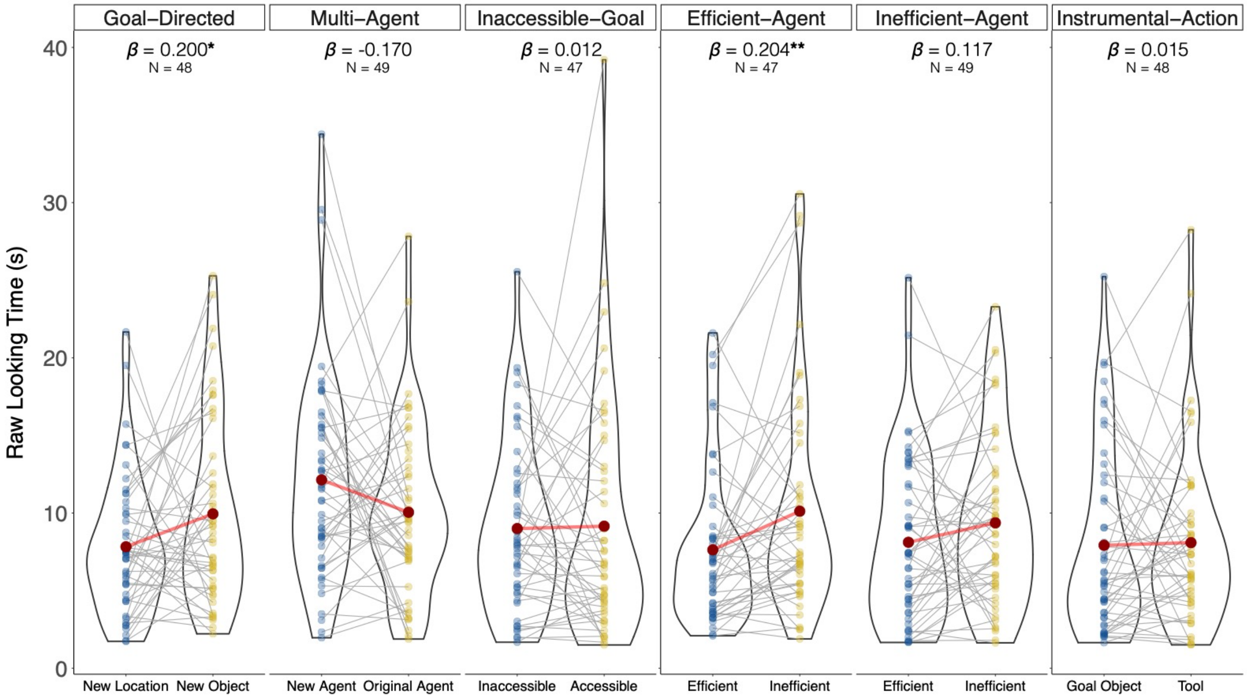

Figure 2. Infants' raw looking times to the two outcomes in each of BIB's six tasks. Gray lines connect the individual looking times (represented by blue and yellow dots) of each infant to each

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 the statistical analyses are reported in the main text (*$p < .05$, **$p < .01$).

We first consider infants' performance on the *Goal-Directed*, *Multi-Agent*, and *Inaccessible Goal Tasks*, the three tasks that focus on representations of agents' goal-directed actions. Overall, the results from these three tasks suggest that infants expect agents' goal-directed actions to be towards objects, not locations, and infants' expectations are limited to individual agents in environments with consistent constraints.

First, consistent with prior findings[3] and with our pilot sample (**SI**), infants were surprised (looked longer) when an agent moved to a new object in the *Goal-Directed Task* ($F(1, 47) = 4.09$, $p = .049$). Prior findings are mixed as to whether infants expect goals to be specific to or shared between agents[28,29,34,35]. Accordingly, infants presented with a new agent in the *Multi-Agent Task* looked longer when that agent versus the original agent moved to a new object ($F(1, 48) = 3.41$, $p = .071$). Infants may attend equally when either a new or a familiar agent approaches a new object for the first time[9], and the visual attention elicited by the new agent, who appeared for the first and only time in that outcome, may explain infants' longer looking (prior studies had presented the new agent in both test outcomes[28]). Finally, prior findings suggest both that infants recognize when an object is inaccessible to an agent[30,36] and that infants do not carry over their expectations of an agent's goal-directed actions to new environments[37]. Accordingly, in the *Inaccessible-Goal Task*, infants showed no difference in surprise when an agent moved to a new object when its goal object was accessible versus inaccessible ($F(1, 46) = 0.02$, $p = .891$). Infants may indeed have recognized that the agent's goal object was inaccessible in the test environment, and they may have thus considered this new constraint indicative of a new environment and not carried-over over any predictions about the agent's goal-directed actions.

We next consider infants' performance on the *Efficient-Agent* and *Inefficient-Agent Tasks*, the two tasks that focus on representations of agents' rational actions. Overall, the results of these two tasks suggest that infants have expectations that rational, goal-directed agents will act efficiently to achieve goals. They also suggest that infants may revert to default expectations about efficiency for agents who act in varying environments and whose actions are ambiguously goal-directed.

First, consistent with prior findings[4] and with our pilot sample (**SI**), infants were surprised when an efficient agent later took an inefficient path to an object in the *Efficient-Agent Task* ($F(1, 46)= 7.72$, $p = .008$). Prior findings suggest both that infants have no expectations about the subsequent actions of an agent who had previously moved inefficiently[4,38] and that infants expect such an inefficient agent to later move efficiently if there is a new obstacle in the test environment[31]. Infants' performance on the *Inefficient-Agent Task* may reflect both previous findings. In particular, infants showed no 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 ($F(1, 132) = 0.49$, $p = .484$), suggesting that infants' surprise at the inefficient agent's later inefficient action was no different from their surprise at the efficient agent's later inefficient action. While the *Inefficient-Agent Task* minimally varied the arrangements of obstacles across familiarization and test environments, and we saw no effects of the presence or absence of obstacles on infants' performance ($F(1, 45) = 0.03$, $p = .872$; **SI**), some infants may have nevertheless considered the constraints of the test environment new enough compared to the familiarization environments to predict that the agent would act efficiently in the test environment[31]. A post hoc analysis, moreover, suggests the order in which infants saw these two tasks may have affected their performance (**Figure S1**). Infants who saw the *Efficient-Agent Task* first (N = 23) were surprised when the efficient agent later acted inefficiently but showed no surprise when the inefficient agent later acted inefficiently (*Efficient-Agent Task*: $M_{efficient}$ = 7.89 s; $M_{inefficient}$ = 11.51 s; *Inefficient-Agent Task*: $M_{efficient}$ = 7.32 s; $M_{inefficient}$ = 7.74 s). In contrast, infants who saw the *Inefficient-Agent Task* first (N = 22) were surprised when the agents in both tasks later acted inefficiently (*Efficient-Agent Task*: $M_{efficient}$ = 7.74 s; $M_{inefficient}$ = 9.24 s; *Inefficient-Agent Task*: $M_{efficient}$ = 8.32 s; $M_{inefficient}$ = 11.05 s). Because of BIB's minimal cues to agency, it is thus possible that infants who encountered BIB's version of an efficient agent first may have been better able to subsequently differentiate that agent's actions from the inefficient agent's actions.

Finally, we consider infants performance on the *Instrumental-Action Task*. The results from this task suggest that infants' knowledge of a particular object's causal efficacy as a tool for instrumental action[39] or an extended familiarization to the task's more complex displays[40] may be required for infants to recognize instrumental actions.

First, 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[33,41] or on novel tools to which infants were first given direct experience[39]. The tool infants saw in the *Instrumental-Action Task* was both novel and not something they were given experience with. Accordingly, infants were not surprised when the agent moved to the tool as opposed to its goal object when the tool was no longer needed to achieve the goal ($F(1, 47) = 0.03$, $p = .853$). A post hoc analysis may provide further insight into infants' performance. In particular, previous studies on infants' recognition of instrumental actions relied on a looking-time paradigm in which test trials started only after an individual infant's looking time had decreased to half its initial levels. Twenty-one infants achieved this decrease in looking time across the fixed number of eight familiarization trials. Consistent with prior findings, those infants were surprised when the agent moved to the tool instead of directly to its goal object ($M_{goal\ object}$ = 5.25 s; M_{tool} = 7.14 s). Infants who did not show this looking-time decrease (N = 27), in contrast, were instead surprised when the agent moved directly to its goal object ($M_{goal\ object}$ = 10.00 s; M_{tool} = 8.83 s). This relation between looking decrease during familiarization and performance at test was present only in the *Instrumental-Action Task*, not in the other tasks (*p*s > .126, see **SI**).

In sum, infants' performance on BIB reveals that they have strong expectations that agents will exhibit rational and efficient goal-directed actions towards objects. These results are consistent with findings in the prior literature on infants' action understanding, but they extend this literature to demonstrate that infants' knowledge is abstract enough to include cases in which agents and objects are conveyed through BIB's highly minimal displays. Addressing mixed findings in the prior literature, moreover, infants' performance on BIB suggests that they may similarly attend to both new and familiar agents who demonstrate new goals, like moving to new objects, and infants may fail to recognize instrumental actions when such actions are novel or causally opaque. Finally, infants do not carry over their expectations about the goals of agents' actions to new environments, and infants may revert to default expectations about the efficiency of goal-directed actions for agents who act in new environments or agents whose previous actions were ambiguously goal-directed.

Machine Performance on BIB

To examine whether infants' intelligence about agents might be reflected in state-of-the-art machine intelligence, we next directly compare infants' performance on BIB to the performance of three learning-driven neural-network models. These models span two classes, behavioral cloning and video modeling, and the behavioral cloning model includes both multi-layer perceptron and recurrent neural network architectural variants. All three models' architectures are designed to form implicit mental state inferences from behavior[13,42] (**Figure S2**). Prior to being tested, the models were trained on BIB's background training videos, as described above, in the **Materials and Methods**, and in the **SI**. While the performance of such learning-driven models has not previously been compared to human performance (let alone to infant performance) and models like these are limited in their capacity for flexible generalization to out-of-distribution novel test displays compared to the displays used for their training (a generalization BIB requires and infants excel at), our comparison nevertheless tests whether the standard "machine theory of mind" captured in such models might be missing key principles of commonsense psychology about agents that infants possess.

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. For example, the behavioral cloning models produced the maximum possible surprisal score if they predicted the agent would move upward and to the left but the agent actually moved downward and to the right. 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. Because those results are consistent with the models' performance on the infant videos and with prior work[13], we report them in the **SI**.

Figure 3 displays the Z-scored means of the models' surprisal scores over four runs for each model to the expected and unexpected outcomes for each task (see the **SI** for additional details about the models' performance). The Z-scored means of infants' looking times are also displayed for comparison. Overall, the models, unlike the infants, did not attribute to agents goal-directed actions towards objects or a principle of rationality that leads to default expectations of agents' efficient actions towards goals. Infants' commonsense intelligence about agents thus includes key features missing in standard forms of machine common sense.

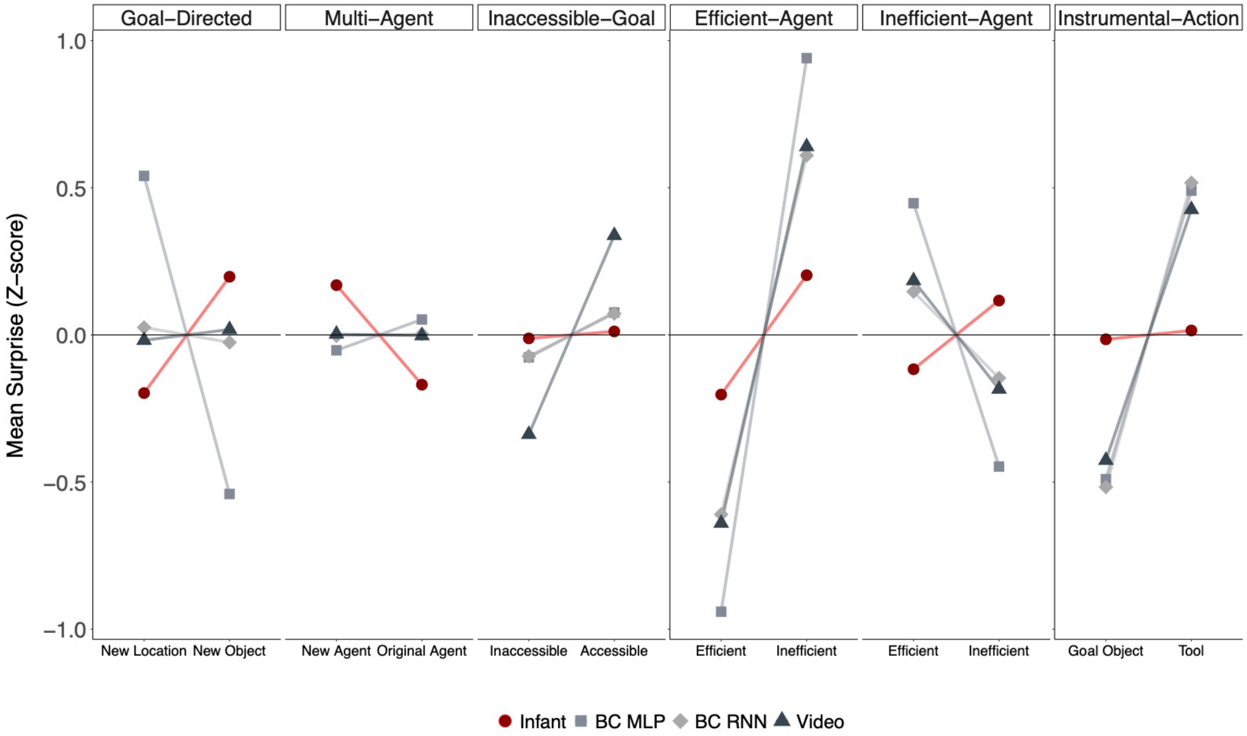

Figure 3. *Z*-scored means of the models' surprisal scores (shown in shades of gray) and the infants' looking times (shown in red) to the expected and unexpected outcomes in each of BIB's six tasks.

The simplest way to evaluate any differences between the model and infant performance is to consider the direction of the surprise, not its degree: Are models and infants surprised by the same outcome or by opposite outcomes? First, the models were more surprised in the *Goal-Directed Task* when the agent moved to a new location versus to a new object. This expectation is the opposite of infants'. Second, while the video model appeared to succeed in the *Inaccessible-Goal Task*, given its 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 black barriers 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 to when it moved to either one of the accessible objects. Third, the models were more surprised when an efficient agent later moved inefficiently in the *Efficient-Agent Task*. The expectation underlying this surprise is shared by infants, but it only captures one component of infants' expectations of agents' rational action. In particular, the models were also surprised when the inefficient agent later took an *efficient* path in the *Inefficient-Agent Task*, expecting instead that it would continue to act inefficiently. While the results with infants on this task were mixed, in general, infants were more surprised at the opposite outcome, i.e., when the inefficient agent later took an *inefficient* path. Finally, while the models appeared to be more surprised when an agent moved to the tool instead of directly to its goal object in the *Instrumental-Action Task*, a 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**). The models thus did not understand agent's instrumental actions.

Discussion

Although learning-driven neural-network models have accelerated recent advances in AI[42,43], three such models tested here on BIB fell short of capturing infants' commonsense psychology about the rational, goal-directed actions of agents. Alternative models based on Bayesian inverse planning have been applied successfully to tasks like BIB by making more explicit, abstract inferences about mental states[26,44,45], but extending the Bayesian approach to BIB in particular and videos in general is not straightforward. Approaches based on inverse reinforcement learning[46,47] could also be promising, but they require online, active sampling from the testing environment, while infants do not. It remains an open challenge for learning-driven systems to acquire sufficiently rich, abstract structure from BIB's training and match infant commonsense intelligence. Nevertheless, setting infant common sense as a benchmark for machine common sense will give AI the foundations of human intelligence.

Future work exploring infants' knowledge about the world could extend our approach to investigate other aspects of infant commonsense psychology, including, for example, infants' expectations of agents' notions of cost and value[15,48] or what actions might signal to infants potential social partnerships[49–52]. Such competencies will become increasingly important for AI as well, as AI systems become further embedded in real-world, multi-agent settings that demand common sense. Putting machines and infants into direct dialogue will also give us a comprehensive account of infants' knowledge not only about agents, but also about objects[1,2] and places[5], allowing us to more fully describe human common sense.

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 human knowledge can be built from the foundations our cognitive and developmental theories postulate.

Materials and Methods

Infant Participants

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. The 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.

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 parental 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 by two researchers masked to the task and outcome represented by each value. Exclusions were rather 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 New York University.

Materials

We procedurally generated video stimuli using BIB's environment[13] and chose the clearest examples of the particular principles of commonsense psychology targeted by each task. For example, while some of the trials in the full dataset of BIB's *Goal-Directed Task* had slight variations in the distance between the agent and its goal object versus a new object at test, the trial we used presented the two objects at exactly the same distance from the agent. All of the stimuli videos are available at: [ACCESSIBLE TO REVIEWRS]. Each of BIB's tasks consisted of a familiarization phase and a test phase. The familiarization phase included a succession of eight trials that introduced the main elements of the visual displays used in the test phase and served to set up any expectations for the test phase.

Goal-Directed Task: During familiarization, an agent repeatedly moved to one of two objects in a grid world. The agent's starting position was fixed across trials, and the locations of the objects were correlated with their identities such that the goal object and non-goal object appeared in approximately the same location across trials. The test used two object locations that had been used during one familiarization trial, but the identity of the objects at those locations was switched. There were two outcomes: the agent moved to the object that had been its goal during the familiarization (expected); or, the agent moved to the object that had not been its goal during familiarization (unexpected). Each test trial lasted 3 s. The following variables were counterbalanced in different versions of the stimuli: 1) the goal object (2); 2) the side of the goal object during familiarization (2); 3) the order of the test trials (2). This yielded eight versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Multi-Agent Task: During familiarization, an agent consistently moved to one object over the other, as above, but objects appeared at widely varying locations in the grid world. The test presented the two objects in new locations and either a new agent who approached the original agent's non-goal object (expected/no expectation) or the original agent who approached its non-goal object (unexpected). Each test trial lasted 5 s. The following variables were counterbalanced in different versions of the stimuli: 1) the goal object (2); 2) the agent during familiarization (2); 3) the order of test trials (2). This yielded eight versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Inaccessible-Goal Task: During familiarization, an agent consistently moved to one object over the other, as above, and objects appeared at widely varying locations in the grid world. The test presented the two objects in new locations and two possible outcomes. The agent's goal object was either inaccessible, blocked on all sides by fixed black barriers, and the agent moved to its non-goal object (expected/no expectation). Or, both of the objects remained accessible, and the agent moved to its non-goal object (unexpected). Each test trial lasted 3 s. The following variables were counterbalanced in different versions of the stimuli: 1) the goal object (2); 2) the order of test trials (2). This yielded four versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Efficient-Agent Task: During familiarization, an agent consistently moved along an efficient path to its goal object around fixed black obstacles in the gird world. The test included two possible scenarios. In one scenario, there was no obstacle between the agent and the goal object. The agent either moved directly to its goal object, following a straight line (expected) or moved along one of the same curved paths it moved during familiarization even though that path was now inefficient (unexpected). The goal object was farther from the agent's starting position in the expected trial compared to the unexpected trial and so the time it took for the agent to move to its goal was matched across trials. In the other scenario, a new obstacle appeared between the agent and its goal object, and the agent either moved on the most efficient, curved path to it (expected) or moved on a less efficient curved path (unexpected). Again, the goal object was farther from the agent's starting position in the expected trial compared to the unexpected trial and so the time it took for the agent to move to its goal was matched across trials. Each test trial lasted 5 s. The following variables were counterbalanced in different versions of the stimuli: 1) the presence of an obstacle at test (2); 2) the order of test trials (2). This yielded four versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Inefficient-Agent Task: This task used identical stimuli to the *Efficient-Agent Task* except for two changes. First, the shapes and colors (and so the identities) of the agent and object were different across tasks. Second, in the *Inefficient-Agent Task* the obstacles present during familiarization were absent, so nothing blocked the agent's straight-line path to its goal object. The agent thus appeared inefficient during familiarization. As in the *Efficient-Agent Task*, at test, the agent either moved on the most efficient path to its goal (expected/no expectation) or on an inefficient path (unexpected/no expectation). Each test trial lasted 5 s. The following variables were counterbalanced in different versions of the stimuli: 1) the presence of an obstacle at test (2); 2) the order of test trials (2). This yielded four versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Instrumental-Action Task: The familiarization included five main elements: an agent; a goal object; a key; a lock; and a green removable barrier. During familiarization, the green barrier initially restricted the agent's access to the object. The agent removed the barrier by collecting and then inserting the key into the lock. The agent then moved to the object. The test phase presented two different scenarios. In one scenario there was no green barrier and in the other scenario there was an inconsequential green barrier that did not block the object. In both scenarios, the agent moved directly to the object (expected) or to the key (unexpected). Each test trial lasted 2 s. The following variables were counterbalanced in different versions of the stimuli: 1) whether there was no green barrier or an inconsequential green barrier at test (2); 2) the order of test trials (2). This yielded four versions, which were assigned equally and randomly across the 32 participants who received all six tasks.

Procedure for Testing Infants

Infants were tested online on Zoom. In the first ten minutes of the first testing session, the experimenter explained to parents the instructions for setting up their device and positioning the infant in front of the screen. We asked parents 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 and controlled the progression of stimuli using PyHab[53] 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.

As preregistered, 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 = .98).

Model Specifications and Training

Inspired by previous work[13,42] and adapted to be put into dialogue with the infant data collected here, we tested three learning-driven neural-network models from two classes: behavioral cloning (BC) and video modeling. BC aimed to predict the future actions of an agent, and video modeling aimed to predict the future frames of a video. Each model also varied in terms of how it encoded the familiarization trials. The models' schematized architecture is presented in **Figure S2**.

We tested two BC models with different ways of encoding the familiarization trials. One way of encoding relied on a simple multi-layer perceptron (MLP) to encode pairs of states and actions independently, and the other way of encoding relied on a more complex, bi-directional recurrent neural network (RNN) to sequentially encode pairs of states and actions. The states (frames) were encoded with a convolutional neural network (CNN), which was pretrained using Augmented Temporal Contrast (ATC)[54]. **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[42] of an agent by averaging the embeddings at each time step, randomly subsampling as needed to use no more than 30 frames. To predict the future actions of an agent, defined in a continuous space based on the video (at 3 frames per section), the models combined the characteristic embedding with the current state of the environment (also encoded with the

CNN). The BC models were trained to minimize mean squared error. See **Table S2** for the specifications of the BC models.

We tested one video model (**Figure S2**). This model sequentially encoded each familiarization trial by passing up to 30 frames through a CNN and combining them with a bi-directional RNN. The model obtained a characteristic embedding of an agent by averaging the RNN embeddings. To predict the future state of the agent, 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[55]. The model was trained using a mean squared error in pixel space.

The models were trained on thousands of examples of BIB-like agents exhibiting simple behaviors in a gird world. The training tasks were provided by the BIB dataset[13] and used the same familiarization/test task design as the test set, except that there were only expected outcomes. 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 gird 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. While the training set thus included individual components of the test set (e.g., agents' movement to objects, agents' consistent object goals, barriers, tools, etc.), success on the test set required models to flexibly combine representations across the different training tasks.

We included four runs of each model type with the runs initialized randomly and trained until they converged on the background training. 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 (4.0). 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.

References

1. R. Baillargeon, E. S. Spelke, S. Wasserman, Object permanence in five-month-old infants. *Cognition*. **20**, 191–208 (1985).

2. A. E. Stahl, L. Feigenson, Observing the unexpected enhances infants' learning and exploration. *Science.* **348**, 91–94 (2015).

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

4. G. Gergely, Z. Nádasdy, G. Csibra, S. Bíró, Taking the intentional stance at 12 months of age. *Cognition*. **56**, 165–193 (1995).

5. L. Hermer, E. S. Spelke, A geometric process for spatial reorientation in young children. *Nature*. **370**, 57–59 (1994).

6. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people. *Behav. Brain Sci.* **40** (2017).

7. G. Marcus, E. Davis, *Rebooting AI: Building artificial intelligence we can trust* (Pantheon Books, New York, NY, 2019).

8. E. S. Spelke, K. D. Kinzler, Core Knowledge. *Dev. Sci.* **10**, 89–96 (2007).

9. E. S. Spelke, *What Babies Know: Core Knowledge and Composition* (Oxford Cognitive Development Series, Oxford University Press, 2022).

10. M. Botvinick, D. G. T. Barrett, P. Battaglia, N. de Freitas, D. Kumaran, J. Z. Leibo, T. Lillicrap, J. Modayil, S. Mohamed, N. C. Rabinowitz, D. J. Rezende, A. Santoro, T. Schaul, C. Summerfield, G. Wayne, T. Weber, D. Wierstra, S. Legg, D. Hassabis, Building machines that learn and think for themselves, *Behav. Brain Sci.* **40,** (2017).

11. M. Sheskin, K. Scott, C. M. Mills, E. Bergelson, E. Bonawitz, E. S. Spelke, L. Fei-Fei, F. C. Keil, H. Gweon, J. B. Tenenbaum, J. Jara-Ettinger, K. E. Adolph, M. Rhodes, M. C. Frank, S. A. Mehr, L. Schulz, Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* **24**, 675–678 (2020).

12. M. C. Frank, E. Bergelson, C. Bergmann, A. Cristia, C. Floccia, J. Gervain, J.K. Hamlin, E.E. Hannon, M. Kline, C. Levelt, C. Lew-Williams, A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy.* **22**, 421-435 (2017).

13. K. Gandhi, G. Stojnic, B. M. Lake, M. R. Dillon, Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Adv. Neural Inf. Process. Syst.* **34,** (2021).

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

15. J. Jara-Ettinger, H. Gweon, L. E. Schulz, J. B. Tenenbaum, The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.* **20**, 589–604 (2016).

16. T. L. Griffiths, Manifesto for a new (computational) cognitive revolution. *Cognition*. **135**, 21–23 (2015).

17. F. Heider, M. Simmel, An experimental study of apparent behavior. *The American Journal of Psychology.* **57**, 243-259 (1944)**.**
18. G. Csibra, G. Gergely, S. Bíró, O. Koós, M. Brockbank, Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition*. **72**, 237–267 (1999).
19. S. C. Johnson, Detecting agents. *Philosophical Transactions of the Royal Society of London*. **358**, 549-559 (2003).
20. T. Gao, G. McCarthy, B. J. Scholl, The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychol. Sci.* **21**, 1845–1853 (2010).
21. A. N. Meltzoff, Understanding the intentions of others: Re-enactment of intened acts by 18-month-old children. *Dev. Psychol.* **31**, 838 (1995).
22. E. S. Spelke, "Preferential-looking methods as tools for the study of cognition in infancy" in *Measurement of audition and vision in the first year of postnatal life: A methodological overview*, G. Gottlieb, N. A. Krasnegor, Eds. (1985), pp. 323–361.
23. E. Téglás, E. Vul, V. Girotto, M. Gonzalez, J. B. Tenenbaum, L. L. Bonatti, "Pure reasoning in 12-month-old infants as probabilistic inference. *Science.* **332**, 1054-1059 (2011).
24. R. Riochet, M. Y. Castro, M. Bernard, A. Lerer, R. Fergus, V. Izard, E. Dupoux, IntPhys: A benchmark for visual intuitive physics reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).
25. K. A. Smith, L. Mei, S. Yao, J. Wu, E. Spelke, J. B. Tenenbaum, T. D. Ullman, Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
26. T. Shu, A. Bhandwaldar, C. Gan, K. Smith, S. Liu, D. Gutfreund, E. Spelke, J. Tenenbaum, T. Ullman, Agent: A benchmark for core psyhological reasoning. *Int. Conf. Mach. Learn.* in *Proceedings of Machine Learning Research.* **139**, 9614-9625 (2021).
27. L. Piloto, A. Weinstein, T. B. Dhruva, A. Ahuga, M. Mirza, G. Wayne, D. Amos, C. Hung, M. Botvinick, Probing Physics Knowledge Using Tools from Developmental Psychology. Preprint at https://arxiv.org/pdf/1804.01128.pdf (2018).
28. J. S. Buresh, A. L. Woodward, Infants track action goals within and across agents. *Cognition*. **104**, 287–314 (2007).
29. B. M. Repacholi, A. Gopnik, Early reasoning about desires: Evidence from 14- and 18-month-olds. *Dev. Psychol.* **33**, 12–21 (1997).
30. R. M. Scott, R. Baillargeon, Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychol. Sci.* **24**, 466–474 (2013).
31. S. Liu, E. S. Spelke, Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*. **160**, 35–42 (2017).
32. A. L. Woodward, J. A. Sommerville, Twelve-month-old infants interpret action in context. *Psychol. Sci.* **11**, 73–77 (2000).
33. J. A. Sommerville, A. L. Woodward, Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition*. **95**, 1–30 (2005).

34. J. Topal, G. Gergely, A. Erdohegyi, G. Csibra, A. Miklosi, Differential sensitivity to human communication in dogs, wolves, and human infants. *Science.* **325**, 1269–1272 (2009).

35. J. Topal, A, Miklosi, Z. Sumegi, A. Kis, Response to comments on "Differential sensitivity to human communication in dogs, wolves, and human infants". *Science.* **329**, 142-142 (2010).

36. Y. Luo, R. Baillargeon, Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*. **105**, 489–512 (2007).

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

38. A. E. Skerry, S. E. Carey, E. S. Spelke, First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proc. Natl. Acad. Sci.* **110**, 18728–18733 (2013).

39. J. A. Sommerville, E. A. Hildebrand, C. C. Crane, Experience Matters: The impact of doing versus watching on infants' subsequent perception of tool-use events. *Dev. Psychol.* **44**, 1249–1256 (2008).

40. E. Tan, J. K. Hamlin, Mechanisms of social evaluation in infancy: A preregistered exploration of infants' eye-movement and pupillary responses to prosocial and antisocial events. *Infancy*. **27**, 255–276 (2022).

41. J. Piaget, *The Origins of Intelligence in the Child* (Routledge, New York, 1953).

42. N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, M. Botvinick, Machine theory of mind. *Int. Conf. Mach. Learn.* in *Proceedings of Machine Learning Research* **80**, 4218-4227 (2018).

43. Y. Lecun, Y. Bengio, G. Hinton, Deep learning. *Nature*. **521**, 436–444 (2015).

44. C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* **1**, 1–10 (2017).

45. C. L. Baker, R. Saxe, J. B. Tenenbaum, Action understanding as inverse planning. *Cognition*. **113**, 329–349 (2009).

46. K. Xu, E. Ratner, A. Dragan, S. Levine, C. Finn, Learning a prior over intent via meta-inverse reinforcement learning. *Int. Conf. Mach. Learn.* in *Proceedings of Machine Learning Research* **97**, 6952-6962 (2019).

47. L. Yu, T. Yu, C. Finn, S. Ermon, Meta-inverse reinforcement learning with probabilistic context variables. *Adv. Neural Inf. Process. Syst.* **32** (2019).

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

49. A. N. Meltzoff, "Like me": A foundation for social cognition. *Dev. Sci.* **10**, 126–134 (2007).

50. M. Tomasello, How children come to understand false beliefs: A shared intentionality account. *Proc. Natl. Acad. Sci.* **115**, 8491–8498 (2018).

51. A. Schachner, S. Carey, Reasoning about "irrational" actions: When intentional movements cannot be explained, the movements themselves are seen as the goal. *Cognition*. **129**, 309–327 (2013).

52. L. J. Powell, E. S. Spelke, Preverbal infants expect members of social groups to act alike. *Proc. Natl. Acad. Sci.* **110**, E3965–E3972 (2013).

53. J. F. Kominsky, PyHab: Open-Source Real Time Infant Gaze Coding and Stimulus Presentation Software. *Infant Behavior & Development* **54**, 114-119 (2019)

54. A. Stooke, K. Lee, P. Abbeel, M. Laskin, Decoupling representation learning from reinforcement learning, *Int. Conf. Mach. Learn.* in *Proceedings of Machine Learning Research*. **139**, 9870-9879 (2021).

55. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical Image Computing and Computer-Assisted Intervention*. **18**, 234-241 (2015).