Theory of AI Mind: How adults and children reason about the "mental states" of conversational AI

Griffin Dietz, Joseph Outa, Lauren Lowe, James A. Landay, Hyowon Gweon
{dietz, joouta, laurenkl, landay, gweon}@stanford.edu
Department of Psychology, Stanford University, USA
Computer Science Department, Stanford University, USA

Abstract

Conversational AI devices are increasingly present in our lives and even used by children to ask questions, play, and learn. These entities not only blur the line between objects and agents—they are speakers (objects) that respond to speech and engage in conversations (agents)—but also operate differently from humans. Here we use a variant of a classic false-belief task to explore adults' and children's attributions of mental states to conversational AI versus human agents. While adults understood that two conversational AI devices, unlike two human agents, may share the same "beliefs" (Exp.1), 3- to 8-year-old children treated two conversational AI devices just like human agents (Exp.2); by 5 years of age, they expected the two devices to maintain separate beliefs rather than share the same belief, with hints of developmental change. Our results suggest that children initially rely on their understanding of agents to make sense of conversational AI.

Keywords: theory of mind; artificial intelligence; false belief

Introduction

Artificial intelligence (AI) is present in many aspects of our lives. Among various forms of AI technologies, conversational AI in voice user interfaces (VUIs) is particularly accessible to lay users: Over 85% (Pew Research Center, 2021) and 51% (ThinkNow, 2020) of U.S. households have VUIs on smartphones or smart speakers (e.g., Google Home), respectively. Critically, their users are not constrained to adults; many children are growing up with these technologies at home, actively using them to ask questions, play, and learn.

Yet, conversational AI occupies an unusual space in the broader landscape of entities in our environment. In many ways, conversational AI devices (e.g., smart speakers) *look* and *behave* like inanimate, non-living objects; they lack bodily features such as faces, arms, and legs, and cannot move on their own. In other ways, however, they present themselves as intelligent agents; despite their limitations, they respond to speech, answer questions, and engage in conversations with human users. How do people—especially young children—conceptualize these novel entities that blur the distinction between objects and agents? More specifically, what does their mental model of AI—what it is, what it knows, and how it knows—look like?

The current study investigates how human users, both adults and children, reason about the "minds" of conversational AI. Just as humans interact with other people based on a lay (rather than scientific) theory of how mental states give rise to behaviors (i.e., theory of mind, intuitive psychology; Wellman, 1992; Gopnik & Wellman, 1992; Jara-Ettinger et al., 2016), the ways in which humans interact with AI agents might also be supported by a theory-like understanding: an intuitive theory (or a mental model) of AI minds (Flanagan et al., 2023; Gweon et al., 2023). Yet, an accurate understanding of conversational AI's "mind" also differs from our mental model of the human mind in important ways; for instance, while two individual human agents have separate minds of their own, two conversational AI devices can be "connected" (e.g., two smart speakers in the same household) and share access to the same knowledge base. Thus, acquiring an accurate understanding of such properties requires going beyond one's existing theory of mind.

There are clear practical reasons for the need to understand how people represent and reason about AI minds. Having incorrect mental models of AI can raise usability issues and create educational—or even ethical—concerns, especially for young users (Eslami et al., 2016; French & Hancock, 2017). To facilitate an accurate understanding of conversational AI and anticipate potential ethical challenges in human-AI interactions, we should first understand how humans, particularly young children, perceive, conceptualize, and reason about intelligent agents (Gweon et al., 2023).

Beyond these practical reasons, however, there are also important scientific reasons to study people's mental models of AI and how it develops in early childhood. Humans, even at a very young age, have an understanding of agents as entities with perceptual capacities and mental states (Spelke, 2022; Csibra et al., 2003; Woodward, 1998), and their ability to reason about others' mental states to predict, explain, and learn from others' behaviors also continues to develop throughout early childhood (Wellman, 2014; Phillips et al., 2021; Gweon, 2021). Prior work on children's ability to distinguish between the living and non-living (Wellman & Estes, 1986; Gelman, 1990; Rosengren et al., 1991; Inagaki & Hatano, 1996) and their intuitive understanding of biology more broadly (Hatano & Inagaki, 1994; Carey, 1985) also suggests that children in their preschool years continue to show marked changes in how they reason about biological entities that do not possess canonical features of agents (e.g., trees and plants), such as self-locomotion or facial features.

Prior work suggests that targets of mental-state attribution often extend to non-living kinds such as puppets (e.g., Wellman et al., 2001; Yu & Wellman, 2022; Asaba et al., 2019)

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

and robots (e.g., Jipson & Gelman, 2007; Bernstein & Crowley, 2008). Beyond perceptual and psychological properties, children attribute freedom of choice to humanoid robots, consider them as deserving of moral evaluation (Flanagan et al., 2019, 2021), and trust information from robots similarly to the way they trust humans (Brink & Wellman, 2020). However, these studies involved target entities that *look, behave, and move* like humans. While even minimalistic agents (e.g., geometric shapes) can elicit robust mental-state attribution (Heider & Simmel, 1944) and have often been used in developmental research (e.g., Kominsky et al., 2022; Hamlin, 2014; Gergely & Csibra, 2003), they still exhibit critical cues for animacy such as self-locomotion and goal-directed behavior. Unlike these agents, conversational AI devices—at least in their appearance—are indistinguishable from other objects.

Emerging literature on conversational AI, however, offers some insights. Although most developmental research with conversational AI relies on observations of children interacting with these entities (Druga et al., 2017; Xu & Warschauer, 2020; Oranç & Ruggeri, 2021), some studies have used interviews/surveys that ask children—or their parents–to describe them (Druga et al., 2017; Xu & Warschauer, 2020; Girouard-Hallam et al., 2021; Hoffman et al., 2021; Oranç & Ruggeri, 2021; Bharadwaj, 2022) or answer questions about their properties (Flanagan et al., 2019; Flannery et al., 2013). These interviews find that children may assign mental, social, or moral attributes to voice assistants (Girouard-Hallam et al., 2021), attribute preferences or emotions (Xu & Warschauer, 2020), ask questions pertaining to the self or environment (Oranç & Ruggeri, 2021), or even develop emotional ties with these agents (Hoffman et al., 2021).

In particular, Flanagan et al. (2023) ask 4 to 11 year old children to describe the physical, moral and socio-cognitive capabilities of three different interactive technologies—a Roomba vacuum, an Amazon Alexa, and a Nao humanoid robot. Children's responses cluster into a three-factor structure—having experiences, having minds, and deserving moral treatment—and they show a general tendency to endorse agent-like features: Alexa is conceived as capable of thinking and feeling but not of physical experiences like hunger, Roomba is thought to have physical experiences but no mental states, and Nao is perceived to have emotional and mental capabilities like intentional actions. Critically, across this work, a trend emerges that older children are less likely to exhibit these behaviors.

From this perspective, the uniqueness of conversational AI as objects with agent-like capacities offers an intriguing opportunity to understand how children reason about the minds of novel entities that do not neatly map onto their existing theories. Yet, little work has directly investigated this question using experimental methods that allow comparison between reasoning about humans versus conversational AI. Here, we build on decades of prior work on false-belief understanding to explore how adults and children reason about the "beliefs" of conversational AI devices and whether this reasoning differs from their reasoning about the beliefs of human agents.

The process by which children come to understand that others' beliefs can differ from their own (or what is true about the actual world) has been investigated extensively by various false-belief tasks (Perner et al., 1987; Wellman & Estes, 1986; Wellman & Gelman, 1992; Wimmer & Perner, 1983; Rakoczy, 2022). One classic example is the Sally-Anne task (Baron-Cohen et al., 1985), where children are presented with an agent (e.g., Sally, who placed her ball inside a basket) whose belief is then rendered false by being absent during a world state change (e.g., Anne took the ball and placed it inside a box instead). Children are then asked to predict the agent's action (e.g., where Sally would look for the ball). Success on these tasks has been regarded as evidence that children can use their intuitive theory of mind to reason about others' mental states.

Building on this work, we presented children with novel scenarios that tap into their reasoning about the "beliefs" of people and conversational AI devices. Imagine Person A was told one fact, and Person B was told something that directly contradicts what A knows. In this case, it is intuitively clear that Person A and B have different belief states. Critically however, if A and B were two connected smart speakers in the same house, you might expect Device A to share the same belief as Device B. Importantly, this reasoning requires an understanding of the technological peculiarity of conversational AI; they can be present in any hardware but have access to the same, single knowledge base. Without this understanding, you might treat them as two separate entities and make the same inferences as you would for humans. Would adults understand this property of conversational AI, and how might this develop in early childhood?

Assuming that adults have a fully developed understanding of others' beliefs, they would have no trouble reasoning about Person A and B; also, insofar as they have a basic understanding of how smart speakers and conversational AI work, they would respond differently in the latter scenario, understanding that even Device A would have updated its "beliefs" to match those of Device B. First establishing this pattern in adults (Exp. 1) would allow us to then investigate the development of children's reasoning in both scenarios (Exp. 2).

Experiment 1

In Experiment 1, we presented adults with a variant of a false-belief task that closely resembled our example above. Similar to how the Sally-Anne false belief task (Baron-Cohen et al., 1985) probed participants' action prediction, we asked what the first agent would do given their knowledge. To make these scenarios apply to both humans and conversational AI, the two agents had identical names in both versions. While Siri or Alexa would be most intuitive for conversational AI, it was important to use a novel name that would be plausible for both humans and conversational AI. Thus, we used the name "Paisia" in all of our scenarios (see Methods/Stimuli). To make these agents "act" on their "beliefs," our scenario

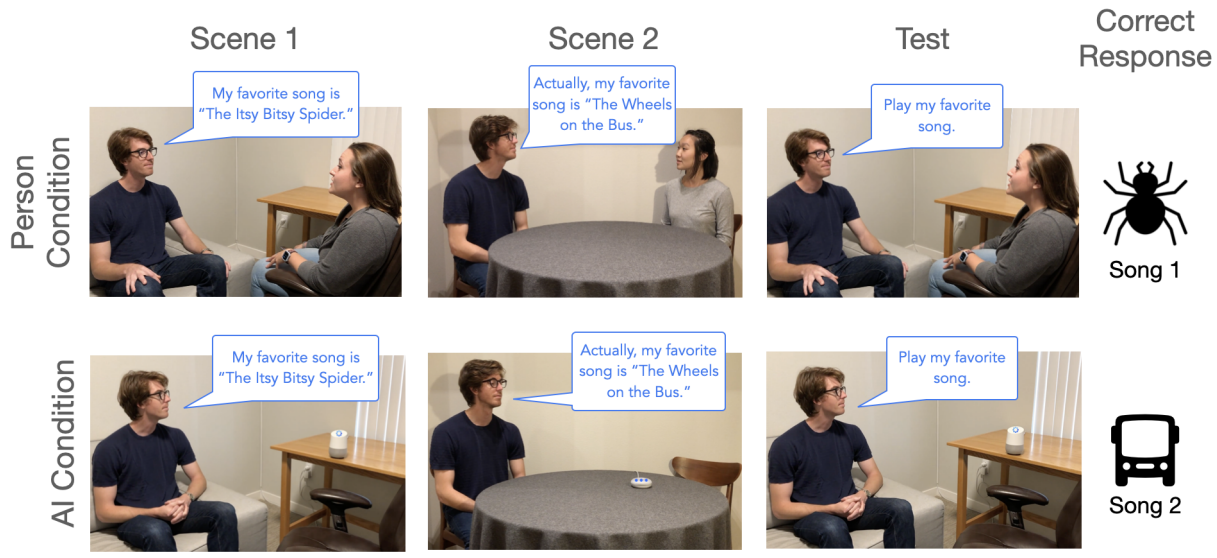

Figure 1: Overview of experimental procedure across the Person and AI conditions.

involved one's favorite song (rather than food) and leveraged the fact that both humans and smart speakers could play one's favorite song.

Participants

We recruited 57 adults from Prolific (Palan & Schitter, 2018), randomly assigned to one of two conditions: Person or AI. All were native speakers of English, 28 were female (28 male, 1 nonbinary), and they had an average age of 37.60 (SD = 12.53). Their experience with smart speakers varied widely; 27 (47%) had at least one smart speaker at home, and 13 of those (23%) had two or more; the remaining 30 participants (53%) did not have any smart speakers at home. The study lasted for about 5 minutes and participants were paid $1.25 in exchange for their time. Three additional participants were excluded for failing attention checks.

Stimuli

The stimuli consisted of video clips featuring a human agent named "Scotty" who interacted with either two human agents (Person condition) or two smart speaker AI devices (AI condition) in two different rooms (a dining room and a home office). In both conditions, Scotty first talked to one person or device named "Paisia" in one room, then walked to a different room to speak with another person or device, also named "Paisia." Despite having identical names, the two Paisias were easily visually distinguishable in both conditions. In the Person condition, the two Paisias were two different adult female actors with distinct hair, skin color, etc.; in the AI condition, the two Paisias were two Google Home smart speakers that were different in shape, size, and color, although they had the same voice to reflect the attributes of commercial smart speakers in the same home.

During the videos, Paisia played the first few verses of the song "The Wheels on the Bus" or "The Itsy Bitsy Spider" either as a person using a smartphone (Person condition) or as a conversational AI device playing music through its speakers (AI condition). These specific songs were chosen with child participants in mind. The order of the rooms and songs were counterbalanced across participants.

Procedure

Introduction: Participants were told that they would watch a video about a person named Scotty who would talk to two different people/smart speakers, both named Paisia. They were then asked two initial attention check questions about the people and places to be shown in the videos. Participants then watched two video clips in which Scotty interacts with one Paisia (Scene 1) and then with another Paisia (Scene 2), each with an additional attention check question, followed by a Test clip and the key test question (see Figure 1). While participants watched different videos depending on the condition, the script was identical between conditions.

Scene 1: In a home office, Scotty said to Paisia: "Hey Paisia, do you remember my name?" Paisia responded, "Your name is Scotty." Scotty then said: "Hey Paisia, my favorite song is The Wheels on the Bus," and Paisia replied: "Okay, I'll remember that you said your favorite song is The Wheels on the Bus." Scotty then asked: "Hey Paisia, can you please play my favorite song?" and Paisia played the song's first verse. Note that Scotty always started his utterance with "Hey Paisia" in much the same way users of conversational AI use "wake words" to trigger the device. Paisia responded using the exact response Google Assistant gives to these inputs.

After watching this video clip, participants answered an attention check question: "In the office, what song did Scotty tell Paisia was his favorite?" Participants were given four options—The Wheels on the Bus, The Itsy Bitsy Spider, Old

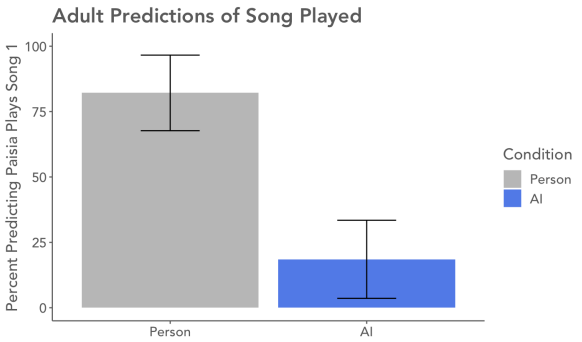

Figure 2: Results from Experiment 1, showing the proportion of adult participants who chose Song 1 (the song from Scene 1) in each condition. Song 1 was the correct answer in the Person condition and incorrect answer in the AI condition. Error bars show 95% CI.

McDonald Had a Farm, and If You're Happy And You Know It—and the order of presentation for these response options was randomized for each participant. Then participants were prompted to watch the second video (Scene 2).

Scene 2: The Scene 2 video was very similar to Scene 1. Written text in the survey implied that the scenes were temporally contiguous, with Scene 2 immediately following Scene 1. In Scene 2, participants were told Scotty entered the dining room to talk to the other Paisia. The script was identical to the first video except that Scotty told the second Paisia: "Hey Paisia, *actually*, my favorite song is The Itsy-Bitsy Spider," and Paisia played this song instead. The same attention check question was asked after the Scene 2 video.

Importantly, the songs used in the two scenes were counterbalanced. Thus we refer to the song used in Scene 1 as Song 1 (i.e., Wheels on the Bus in the above example) and the song used in Scene 2 as Song 2 (i.e., Itsy-Bitsy Spider in the above example).

Test: In the critical test video, Scotty returned to the office (shown in Scene 1) to talk to the first Paisia. He asked: "Hey Paisia, can you please play my favorite song?" The video cut off before Paisia responded, and participants were presented with the final test question: "What song will this Paisia play?" Participants selected from options—The Wheels on the Bus (Song 1) and The Itsy Bitsy Spider (Song 2)—and were asked to explain their choice using a text box.

Results and Discussion

Our primary question was whether adults differentiate the individual minds of people from the shared "minds" (i.e., data source) of smart speaker devices, and whether experience with such devices moderates this effect. We ran a logistic regression with condition and ownership of connected smart speakers as predictors for choosing Song 1. As expected, participants chose different songs depending on the condition ($z = 3.10$, $p = 0.002$); participants in the Person condition were more likely to predict that Paisia would play Song 1 than those in the AI condition (see Figure 2). Furthermore, participants in the Person condition reliably chose Song 1 above chance ($p < 0.001$; 95% CI: [0.64, 1.00], Binomial test against 50%) and those in the AI condition showed an opposite pattern, choosing Song 1 below chance ($p < 0.001$; 95% CI: [0.00, 0.35]). However, our logistic regression did not find an interaction effect between condition and ownership of multiple connected smart speakers ($z = 0.01$, $p = 0.99$).

These results were in line with our prediction: Adults showed opposite response patterns in the Person versus AI conditions. In the Person condition—although responses did not reach ceiling—participants expected two human agents to have beliefs of their own, such that one agent's existing belief would not change even when new, conflicting information was given to another agent. In contrast, in the AI condition, participants understood that the two smart speakers may share the same "mind," expecting the first agent to update its "belief" based on the information provided to the second agent.

Notably, adults' experience with connected smart speakers (indirectly measured by whether they have more than 2 smart speakers at home) did not influence their responses, suggesting additional sources of experience may have scaffolded adults' intuitions (e.g., devices at work or general information about how these devices function). Collectively, our results suggest that adults already have a mental model of conversational AI in smart devices that differ substantially from their mental model of human agents.

Experiment 2

Given that Experiment 1 established the predicted pattern of responses in adults, in Experiment 2 we used the same task to investigate whether and at what age children begin to distinguish human minds from AI minds.

We explored a wide age range to identify both the development of belief attribution (in the Person condition) and the effect of experience with conversational AI/smart speakers (in the AI condition). To capture this developmental change, we increased our sample size in this second experiment, chose age 3 as our minimum age (Wellman & Gelman, 1992; Wellman et al., 2001), and recruited children up to age 8, anticipating that by early school years children may begin to distinguish AI mind from human mind.

Given that children's responses in standard false belief scenarios undergo a noticeable change between 3 to 5 years of age, we predicted a similar trend in the Person condition: an increase in their tendency to choose Song 1 in the Person condition (i.e., increasing choice of Song 1 with age). In the AI condition, we considered two possibilities. First, given prior work (Xu & Warschauer, 2020; Lee et al., 2019; Oranç & Ruggeri, 2021; Hoffman et al., 2021), children might treat the AI agents as they treat human agents; thus, they might show a similar increase in their tendency to choose Song 1, especially between ages 3 and 5, even though this is, in fact, the *wrong* answer in the AI condition. Second, as children gain

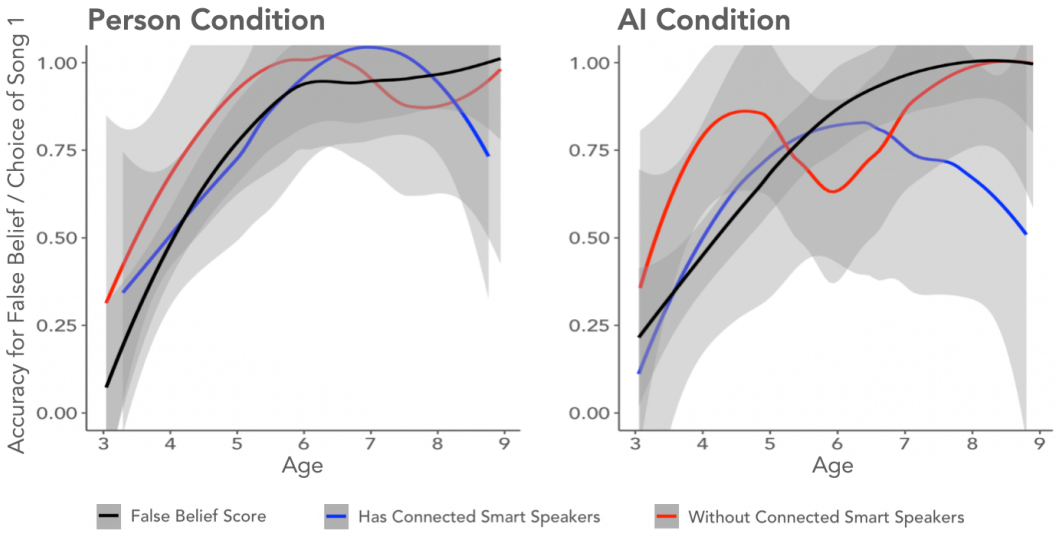

Figure 3: Comparison of children's accuracy on a false belief battery with their choice of song in our task, both scaled to fall between 0 and 1. The black line represents the average score on the false belief questions. The red line shows responses of children who do not have connected smart speakers at home and the blue line shows responses of those who do. We note that for the AI condition Song 2 is the accurate response so we'd be looking for a downward trend with age in this condition.

more experience with smart speakers and conversational AI, they might begin to understand that the two Paisias in the AI condition may share a common "mind." This might manifest as an increasing trend in choice for Song 2, similar to adults.

While we did not have strong a priori evidence about when children would start showing the latter pattern (i.e., increasing choice for Song 2), we anticipated that this tendency might be related to children's overall exposure to, and experience with, smart speakers. Thus we obtained information from parents about children's experience with such devices.

Participants

We recruited 131 children from 3 to 8 years of age ($M = 6.00$, $SD = 1.76$) to participate in an online study session conducted via Zoom (76 female, 52 male, 2 non-binary, 1 declined to answer). 51 of 131 children did not have a smart speaker at home, 28 had one, and 52 had more than one. Families received a $5 gift card in exchange for their time.

We excluded 15 additional children due to failing to answer the two attention check questions accurately (see Procedure; $N = 3$), not completing the study ($N = 7$), technical difficulties (e.g., issues with Zoom; $N = 2$), or having recently watched a sibling participate in the same study ($N = 1$). Due to the importance of understanding the conversation in the videos and the storylines in the false belief battery, we also excluded an additional two children due to limited proficiency in English ($N = 2$), meaning their parent reportedthat the child's native language was not English, the child did not speak English "all" or "most" of the time at home, and the child had not been attending school in an English-language classroom for at least two years.

Stimuli

The videos were identical to the ones used in Experiment 1. Children participated synchronously over Zoom and all questions were read aloud by an experimenter to the child participants. Like Experiment 1, the key test question was a binary choice between Song 1 and Song 2. Additionally, a series of questions aimed to assess children's theory of mind was added to the end of the study.

Procedure

The procedure for Experiment 2 was very similar to that of Experiment 1. Children were told they were going to watch a video about a person named Scotty, and that Scotty would talk to two "Paisias," who were shown on the screen. To avoid confusion especially for young children, the experimenter emphasized that both people/smart speakers were named Paisia and that Scotty would talk first to one Paisia and then to the other Paisia. We then played Scene 1 and Scene 2 videos, and after each video, the experimenter asked two questions to ensure that children could remember the key events in the videos: "Who was Scotty talking to?" and "In the [room], what did Scotty tell Paisia was his favorite song?". Unlike in Experiment 1, we did not provide response options for these attention check questions. If children forgot Paisia's name, we reminded them "Scotty was talking to Paisia." If children forgot the song that Paisia played, we excluded them from the study. At the end of the test video, the experiment asked: "What song will this Paisia play: the Wheels on the Bus or the Itsy Bitsy Spider?" and followed up with a question to explain their choice.

After the main task, we also assessed children's understanding of others' mental states using a more traditional

method: a theory of mind battery based on a storybook, adapted from prior work for remote testing (Richardson et al., 2018; Gweon et al., 2012). Due to time constraints, we selectively administered the following 5 items: diverse desires and diverse beliefs (to establish that children could reliably answer these questions), and 3 false belief items for the main analysis. The storylines were read aloud with images and animation to support understanding and attention.

Results and Discussion

We designed this task as a variant of the false belief task, but it has several features that deviate significantly from classic scenarios. To establish that this task does tap into children's false belief understanding, we first explored the similarity between children's responses to the key test question in the Person condition and their scores on standard false belief scenarios (average accuracy over items): There was a strong correlation between these data ($\rho = 0.66, p < 0.001$).

One way to test our second prediction is with a binomial regression with age (continuous), condition (categorical), and household ownership of connected (i.e., two or more) smart speakers (binary) as predictors of song choice. Our prediction can be expressed as a 3-way interaction between age, condition, and ownership of multiple smart speakers: Older children with connected smart speakers at home may be more likely to respond differently to the test question in the AI condition than in the Person condition. While we did find a significant effect of age ($z = 2.29$, $p = 0.02$), we did not find evidence for a three way interaction ($z = -0.23$, $p = 0.82$).

Critically however, plotting the data in each condition reveals a pattern that begins to emerge in the oldest participants (see Figure 3); while children's responses in the Person condition aligns closely with their ToM scores regardless of smart speaker experience, responses in the AI condition start to diverge between age 7–9 depending on their smart speaker experience. Thus, while these tasks do seem to tap into children's false belief understanding, we do not find conclusive evidence for their ability to distinguish human from AI minds, suggesting that an adult-like understanding of conversational AI may emerge later than we had anticipated.

General Discussion

In this study, we explored how adults and children reason about the minds of conversational AI. Using a novel variant of a classic false-belief task, Experiment 1 established that adults have a mental model of conversational AI that differs substantially from their mental model of human agents; they understood that the cloud-based, connected nature of smart speakers allows them to "share" the same epistemic state.

Experiment 2 found a striking similarity between children's responses in the Person condition of our task and standard false-belief questions, lending support for our task as a variant of the false-belief task. However, children's responses in the AI condition were clearly different from adults. Unlike adults who treated these devices differently from human agents, our findings suggest that children still reason about these devices' "beliefs" (i.e., knowledge base) in much the same way as they reason about the beliefs of human agents. Despite the developmental transition in the Person condition, we did not find clear evidence that children within this age range are able to differentiate the separate minds of people from the connected "minds" of smart speaker devices. This is particularly surprising given that the two devices in the AI condition had identical voices.

One might wonder how to reconcile our finding with existing work on children's reports of smart speakers and virtual assistants. This prior research demonstrates a change in children's reports on the capabilities and characteristics of virtual assistants as they get older, suggesting a shift in how they conceive of conversational AI (Druga et al., 2017; Xu & Warschauer, 2020; Girouard-Hallam et al., 2021; Hoffman et al., 2021; Oranç & Ruggeri, 2021; Flanagan et al., 2023). In particular, recent data from children aged 4 to 11 suggests that the decrease in children's tendency to attribute moral intent is particularly pronounced in older children (Flanagan et al., 2023). Consistent with these findings, our findings suggest that although children begin to understand how conversational AI differs from humans in key aspects, at least through age 8, this realization may not extend to shifts in reasoning about how these devices update their knowledge base.

Interestingly however, we do find a suggestive trend in our data that raises the possibility that this understanding is just about to emerge, at least in our sample. While the results from younger children suggest that children's reasoning about AI minds may build upon their understanding of agents, the understanding of AI minds as "shared" may be driven by the amount of experience with these devices as well as access to formal and informal education about how these devices work (which may, indeed, vary depending on socioeconomic status and other demographic or cultural factors). While typical cognitive development research tends to recruit from primarily middle-class families, samples for online research can, in principle, be more diverse in SES than traditional in-lab studies (see Sheskin et al., 2020, but also Lourenco & Tasimi, 2020). Compared to national statistics—51% of US households have smart speakers (ThinkNow, 2020)—61% of children in our study had one or more smart speakers at home, suggesting that our sample demographics may not deviate far from the national average. To further explore how an understanding of AI minds develops in childhood, we are currently running a large-scale extension of this study that includes older children while also being mindful of sample diversity.

As children enter upper elementary and middle school, they will begin to interact with AI technology more and more. It is therefore important to consider the educational implications of those experiences in future system design. This work demonstrates that by early school years, children still have an "inaccurate" mental model of AI systems that are in fact reflecting their "accurate" mental model of agents, suggesting a need for further AI education that helps to facilitate an understanding of how AI learns and what it knows.

Acknowledgments

We are grateful for our research assistant, Ava Deconcini, and members of the Social Learning Lab (SLL) for valuable feedback on the project. This work was supported by the McDonnell Scholars Award, NSF-2019567, NSF-2120095, and NSF-2042489 sub-award. The experiment was approved by Stanford's Institutional Review Board.

References

Asaba, M., Li, X., Yow, W. Q., & Gweon, H. (2019). A friend, or a toy? Four-year-olds strategically demonstrate their competence to a puppet but only when others treat it as an agent. In *Cogsci* (pp. 98–104).

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Bernstein, D., & Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences*, *17*(2), 225–247.

Bharadwaj, N. (2022). "ok google, how tall is the sky?" how children use and understand digital assistants.

Brink, K. A., & Wellman, H. M. (2020). Robot teachers for children? young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology*, *56*(7), 1268.

Carey, S. (1985). Conceptual change in childhood.

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.

Druga, S., Williams, R., Breazeal, C., & Resnick, M. (2017). "hey Google is it OK if I eat you?" Initial explorations in child-agent interaction. In *Proceedings of the 2017 conference on interaction design and children* (pp. 595–600).

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First i "like" it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2371–2382).

Flanagan, T., Rottman, J., & Howard, L. (2019). Do children ascribe the ability to choose to humanoid robots? In *Cogsci* (pp. 302–308).

Flanagan, T., Rottman, J., & Howard, L. H. (2021). Constrained choice: Children's and adults' attribution of choice to a humanoid robot. *Cognitive Science*, *45*(10), e13043.

Flanagan, T., Wong, G., & Kushnir, T. (2023). The minds of machines: Children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Developmental Psychology*.

Flannery, L. P., Silverman, B., Kazakoff, E. R., Bers, M. U., Bontá, P., & Resnick, M. (2013). Designing scratchjr: support for early childhood learning through computer programming. In *Proceedings of the 12th international conference on interaction design and children* (pp. 1–10).

French, M., & Hancock, J. (2017). What's the folk theory? reasoning about cyber-social systems. *Reasoning About Cyber-Social Systems (February 2, 2017)*.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive science*, *14*(1), 79–106.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292.

Girouard-Hallam, L. N., Streble, H. M., & Danovitch, J. H. (2021). Children's mental, social, and moral attributions toward a familiar digital voice assistant. *Human Behavior and Emerging Technologies*, *3*(5), 1118–1131.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child development*, *83*(6), 1853–1868.

Gweon, H., Fan, J., & Kim, B. (2023). Socially intelligent machines that learn from humans and help humans learn. *Philosophical Transactions of the Royal Society A*. doi: 10.1098/rsta.2022.0048

Hamlin, J. K. (2014). The origins of human morality: Complex socio-moral evaluations by preverbal infants. *New frontiers in social neuroscience*, 165–188.

Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition*, *50*(1-3), 171–188.

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

Hoffman, A., Owen, D., & Calvert, S. L. (2021). Parent reports of children's parasocial relationships with conversational agents: Trusted voices in children's lives. *Human Behavior and Emerging Technologies*, *3*(4), 606–617.

Inagaki, K., & Hatano, G. (1996). Young children's recognition of commonalities between animals and plants. *Child development*, *67*(6), 2823–2840.

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.

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child development*, *78*(6), 1675–1688.

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. *Cognitive Development*, *63*, 101213.

Lee, S., Kim, S., & Lee, S. (2019). "what does your agent look like?" A drawing study to understand users' perceived persona of conversational agent. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–6).

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: conducting science during covid-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

Oranç, C., & Ruggeri, A. (2021). "alexa, let me ask you something different" children's adaptive information search with voice assistants. *Human Behavior and Emerging Technologies*, *3*(4), 595–605.

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British journal of developmental psychology*, *5*(2), 125–137.

Pew Research Center. (2021, April). *Mobile fact sheet.* Pew Research Center.

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ... Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, *44*, e140.

Rakoczy, H. (2022). Foundations of theory of mind and its development in early childhood. *Nature Reviews Psychology*, *1*(4), 223–235.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature communications*, *9*(1), 1–12.

Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development*, *62*(6), 1302–1320.

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., ... others (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

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

ThinkNow. (2020, May). *Voice controlled products.* ThinkNow.

Wellman, H. M. (1992). *The child's theory of mind.* The MIT Press.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), 655–684.

Wellman, H. M., & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child development*, 910–923.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual review of psychology*, *43*(1), 337–375.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

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

Xu, Y., & Warschauer, M. (2020). What are you talking to?: Understanding children's perceptions of conversational agents. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).

Yu, C.-L., & Wellman, H. M. (2022). Young children treat puppets and dolls like real persons in theory of mind research: A meta-analysis of false-belief understanding across ages and countries. *Cognitive Development*, *63*, 101197.