How do children's perceptions of machine intelligence change when training & coding smart programs?

Stefania Druga
University of Washington
Seattle, Washington, USA
st3f@uw.edu

Amy J. Ko
University of Washington
Seattle, Washington, USA
ajko@uw.edu

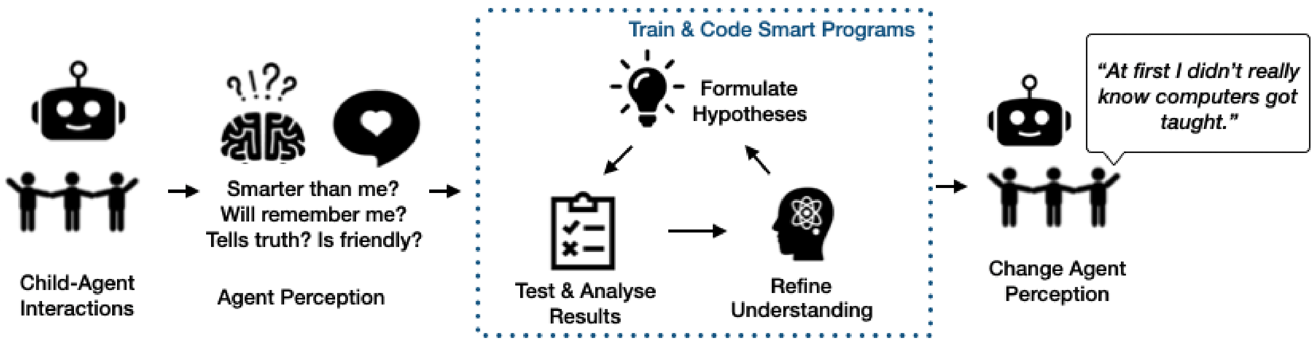

Figure 1: Findings summary: Children (7 to 12 years old) engage in the scientific method when training & coding smart programs and become more skeptical of certain abilities of smart devices.

ABSTRACT

Children are increasingly surrounded by AI technologies but can overestimate smart devices' abilities due to their lack of transparency. Drawing on the sense-making theory, this study explores how children come to see machine intelligence after training custom machine learning models and creating smart programs that use them. Through a 4-week observational study in after-school programs with 52 children (7 to 12 years old), we found that children engage in the scientific method while training, coding and testing their smart programs. We also found that children became more skeptical of certain abilities of smart devices as they shifted their attribution of agency from the devices to the people who program them. These changes in perception happened both through individual interactions with agents and prompted debates with peers. Based on these results, we conclude with discussions on strategies for promoting children's sense-making practices and sense of agency in the age of machine learning.

IDC '21, June 24–30, 2021, Athens, Greece

ACM ISBN 978-1-4503-8452-0/21/06…$15.00
https://doi.org/10.1145/3459990.3460712

CCS CONCEPTS

• **Human-centered computing → Usability testing**; **User centered design**; **Interface design prototyping**; *Field studies*; *Sound-based input / output*; **Human computer interaction (HCI)**; User studies; • **Social and professional topics → Children**.

KEYWORDS

AI literacy, Children, Machine Intelligence, Child-Agent Interaction

ACM Reference Format:
Stefania Druga and Amy J. Ko. 2021. How do children's perceptions of machine intelligence change when training & coding smart programs?. In *Interaction Design and Children (IDC '21), June 24–30, 2021, Athens, Greece.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3459990.3460712

1 INTRODUCTION

Many children are spending more time engaging with artificial intelligence. This engagement with what we will call *smart agents* is likely to increase, as there is significant growth in smart toys and more than 50% of North American households alone are expected to have a dedicated voice-assistant by 2022 [67].

Researchers have begun to examine children's experiences with these agents. For example, smart toys might influence children's perception and attribution of intelligence, their moral choices, or their behavior through play [21, 76, 81]. Prior work has shown that children see agents as friendly and truthful, and older children (7-10 years old) especially consider the agents to be more intelligent than themselves [21]. However, what may seem initially to be a playful interaction between a child and the smart agents can trigger events of significant consequence, such as children being spied on after their connected toys were hacked [49]. Many of these devices have proven to be easy to compromise [3, 74, 78], and some companies designing these technologies utilize questionable practices [42, 54].

The unequal access to smart agents in the home also amplifies digital divides, with only some children learning to make sense of how smart toys and devices function [7, 19]. Prior work has demonstrated that parental attitudes, socio-economic status, and cultural differences play a significant role in how children attribute agency, intelligence, and socio-emotional traits to the agents [20, 22]. Other studies have shown that children often misunderstand agents and tend to overestimate their abilities, either because children do not understand how these agents work, or because artifacts like toys and phones can talk, express emotions, and engage with youth in ways other humans would: with persuasive and charismatic modes of engagement [24, 55, 83]. In this context, we recognize the need for inclusive Artificial Intelligence (AI) literacy efforts to prepare a generation of children growing up with AI. We define AI literacy as the ability to critically decide if, when, and how to use smart devices.

Explorations of AI literacy applications in education are challenging since the mechanisms and opportunities of AI are unfamiliar to most people outside computer science. AI literacy is also considered a vital part of computational thinking [17, 71] and there are arguments to include AI literacy as part of the CS curricula in K12 level [20, 44, 53]. In parallel, several studies explored how youth can learn more about AI by interacting with pre-trained models [33, 75] or training their models [44, 84]. Vartiainen et. al found that young children (3-9 years old) reason about the relationship between their bodily expressions and the output of an interactive image prediction tool and engage in an emergent process of teaching and learning from the machine [75]. Kahn et al. found that high schoolers in developing countries enjoy to created block-based programs using pre-trained AI models but do not always understand how these models work [33]. Zimmermann et al. showed that youth with no programming experience can incorporate AI classifiers into athletic practice by building models of their physical activity on a mobile app [84]. However, none of these prior studies explored how children changed their perception of AI abilities after engaging in AI programming and training activities.

In this study, we plan to address this gap by focusing on one specific aspect of AI literacy: when learning to program smart agents, how do children's perceptions of smart agents' intelligence change? Britto et al. observed that families are assigning meaning and intelligence to smart technologies even before starting to use them and this process bears weight on the decision to adopt them or not [11]. Turkle notes how smart toys in particular have changed the way children evaluate the "liveliness" of a machine. Rather than assessing machines solely based on intelligence, children have now begun to also inquire whether their smart toys can feel and convey emotions [72, 73].

Prior work on general programming suggests some possible changes. For example, Scaife and Duuren found evidence that the "programmability" of technology can shift children's theories of intelligence about computers away from the device and toward the programmers of the device [23, 62]. While these studies were investigating traditional programming, in our study we investigate if similar phenomena can be observed when children are using AI-based training and programming.

We focus our study on two research questions:

- RQ1: How do children make sense of machine intelligence when training and coding smart programs?
- RQ2: How do children's perception of machine intelligence change before and after building smart programs?

To answer these questions, we ran a 4-week study in both public and private after-school programs and community centers with 52 children (7 to 12 years old), observing children's sense-making and measuring their shifts in machine intelligence perception (see Fig. 1). Our investigation makes three contributions to the understanding of AI literacy in children. First, we provide empirical evidence of how children engage in sense-making practices when training and coding smart programs. Second, we present how children's perceptions of machine intelligence change after participating in the study. Finally, we discuss how the theoretical model of sense-making is relevant to developing AI literacy in children.

2 BACKGROUND

While research on children's programming is extensive and research on smart agents is growing, research on children's AI literacy is still sparse. Here we review this work and the theories relevant to exploring these interactions.

2.1 Smart Agents, Avatars and Social Robots

Research on human interactions with smart agents shows a clear tendency of anthropomorphism. For example, humans often anthropomorphize objects and are capable of engaging socially with machines [32, 47, 56]. This is especially true of robots and embodied agents [25, 34, 68]. The more life-like an agent is in terms of embodiment (the physical form of the AI), physical presence, social presence, and appearance, the more persuasive it can become [4, 59, 65]. Vollmer et al. showed that robots could even exert peer pressure over children [76]. In their experiment, 7- to 9-year-old children had a tendency to echo the incorrect, but unanimous, responses of a group of robots to a simple visual task [76]. Smart agents also lead children to overestimate the intelligence of these devices, trusting them, and deferring to them when making decisions [21].

Despite the clear potential for undue influence, smart agents are becoming increasingly embedded in homes. Voice interfaces are not only available in smart home assistants, but are also being embedded in toys which are more familiar to children. For instance, in the case of the My Friend Cayla doll, the device uses a non-encrypted Bluetooth connection to a smart-phone application for triggering the speech functionality, posing safety and privacy risks that children do not understand [49]. Parents are likely equally uninformed; some parents are explicitly using smart devices as parenting tools for setting limits or managing kids' schedules [8, 14, 82]. These tendencies lead us to question how much children could be influenced by smart agents as they becomes more personified, embodied and able to lead conversations, and what role machine intelligence perception can play in preventing this undue influence.

2.2 Children Learning to Train Smart Programs

Numerous efforts over the past few decades have attempted to engage children in learning to code. Early efforts viewed coding as a path to more robust general problem solving skills [50]; more recent efforts in North America and Europe have sought to add or integrate computer science concepts into primary and secondary school curricula, often with goals of equitable economic attainment (e.g., csforall.org). Most of these efforts have focused on introducing basic concepts in programming languages and algorithms, where the behavior of computer programs is determined primarily by the programmer [29, 35].

Only recently have these modern efforts at CS literacy begun to explore AI literacy [41]. Early work revealed that such learning requires youth to understand the role of data in determining machine behavior [46], showing that children are capable of developing a new schema when they can physically test and debug their assumptions. However, recent work has found that, just as with procedural programming, AI programming can result in behavioral complexities that are immensely difficult to debug and comprehend [69]. A core challenge then becomes to make the agents' data-driven reasoning more transparent. Studies analyzing adults' mental models of AI technologies found that even a 15 minute tutorial with an experimenter can significantly increase the soundness of participants' mental models. This phenomenon was consistent in studies on intelligent music recommender systems [38] and the effects of different AI errors [5].

A few initiatives today aim to introduce children to AI either by allowing them to program using pre-trained models [33], incorporate AI classifiers into athletic practice [84] or explore how object recognition works by building custom prediction models [44]. In parallel, prior research on data literacy for children shows the importance of linking statistical inference to personal data [16, 43], and is based on the idea that children may have an easier time understanding how data is used in AI when they are familiar with its origins and meaning [13, 30, 37, 57].

Our study builds on this prior work by allowing children to not only train their own prediction models using their own meaningful examples of text or images, but to also use these custom models in a familiar visual coding environment and program their own smart programs. We specifically want to understand if youth change their perception of machine intelligence in this process.

2.3 Theories of Sense-Making

There are many ways to study how children might come to comprehend the behavior of smart agents. Prior work on program understanding has often focused on cognitive approaches, providing learners with interactive representations of program behavior (e.g., [36]). The machine learning, AI, and HCI communities have followed similar trends in pursuing explainable AI, aiming to invent representations that help people reason accurately about AI behavior [1].

In this work, we take a different theoretical stance, instead approaching children's AI literacy through the lens of sense-making. Within this frame, we define sense-making as a process by which people encounter situations or contexts that are unfamiliar, and then need to process and understand in order to move forward [18]. People form new knowledge from engaging in complex and information rich situations in which they may not always have expertise. The learning sciences further consider how learners make sense of quantitative change in complex systems [80], how learners reason with large sets of data [60], and what role argumentation plays in knowledge formation [63].

Within this literature, sense-making is both an individual *and* social process, occurring in small groups, organizations, communities, and societies. Therefore, we examine children's individual machine-intelligence sense-making processes in the context of the social processes of training and coding smart programs. This approach has strong precedence in recent work on computing education with children, such as children's computer programming in low-vision populations [70], supporting children as data scientists through Scratch programming [15], children's development of spatial reasoning through programming [27], and the family ecology in which children interact with smart agents [31, 64]. This prior work suggests a need to consider not only how children perceive smart agents' intelligence and abilities, but also the social context surrounding children's sense-making practices when interacting with smart technologies.

3 METHOD

To understand how children make sense of machine-intelligence when training and coding smart programs and how their perception of machine intelligence changes, we structured our study in the following order: perception game, observations of children in 3 learning activities, perception game, analyze pre/post perception game responses and observations to understand changes in children's perception of machine-intelligence. Fig. 2 overviews the study design.

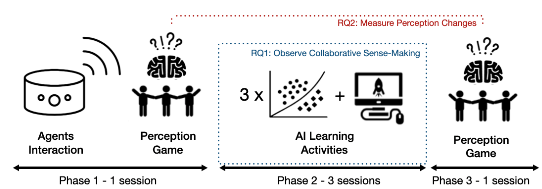

Figure 2: Study Overview

3.1 Selection and Participation of Children

We specifically chose many different locations for our study workshops to include a diverse population of students. The workshops took place in the following locations in Massachusetts, USA: an after-school program in a Spanish-English bilingual public school with mostly children of immigrant families (Public After-School Program), a non-profit community the center housed in a former church (Church Community Center); an after-school program in a private school in Cambridge (Private After-School Program), a private STEM center in Lexington (STEAM Community Center). In total, we had 52 children of ages 7-12 years old, with 16 girls and 36 boys, 28 younger children (7-9 years old) and 24 older children (10-12 years old).

3.2 Study Procedure

Our study comprised of three phases: 1) initial encounters with agents and perception game (pre), 2) programming and training AI and 3) perception game (post).

Phase 1: Initial Encounters. The goal of our first phase was to introduce children to smart agents and programming, while establishing a baseline of children's perceptions of machine intelligence. We started by introducing to children three different embodied intelligent agents: Jibo robot, Anki's Cozmo robot and Amazon's Alexa home assistant. First, the researchers would demonstrate the vocal commands for activating each agent (e.g. "Hey Jibo" or "Alexa") and some of its capabilities. Then the children were left to explore on their own. After the initial play and interaction, children were also encouraged to program the agents using the existing commercial coding apps developed for each agent. At the end of the session children answered questions about the agents intelligence and abilities as part of the Pre-Perception game (described shortly).

Phase 2: Programming AI. Next, we introduced children to the Cognimates AI platform (described shortly) where they could learn how to train, code and test a series of smart programs. To guide this introduction, we created the set of learning activities with starter coding projects.

Phase 3: Post-Perception Game. In this final session, we repeated the Perception Game from Phase 1, gathering a post-measure of children's perceptions of machine intelligence had changed, supporting our second research question. Because not all children attended the final session, not all the children completed the perception game. Additionally, in the case of the public after-school program, we were not able to collect any data because of cancellations due to snow. When we did meet the children again a few days later, we only conducted interviews and had a final discussion about what they learned, which concluded with a certificate of participation award.

3.3 Study Materials: AI Platform, Learning Activities & Perception Game

In this section we will present the Cognimates AI platform we used in the study, the learning activities we used to guide instruction and the perception game we used to measure children's shift in perceptions.

3.3.1 *Cognimates AI platform.*

For this study, we needed a platform that would allow children to both train, test and program with their custom AI models. There are many professional AI training tools we could have adopted, but because our study was focused on changes in perception of intelligence, and the children in the study had no prior exposure to programming AI, the tools needed to be highly scaffolded for learning. Therefore, we built a platform that integrated the model training, programming, and testing into a single platform, giving learners multiple views of the same training data. This followed the Bifocal Modeling (BM) framework [9], which suggests that representing the same science experiment data in different examples synchronously helps children more quickly abstract and infer information about this data.

Our Cognimates AI platform had two main components: the TeachAI page and the Codelab. The Teach AI page (Fig. 3.1) provided children with opportunities to train machine learning models with their own data. The Codelab (Fig. 3.2) was the section of the platform where children could write interactive programs using a rich collection of visual blocks, building up AI behaviors to gather user input, classify it, and respond. On the Teach AI Page children could train their own classifiers by providing examples of images and text. For example, a child would train an ideal model, e.g., for distinguishing unicorns from narwhals (see Fig.3.1).

Fig.3 portrays a case where a child created a game that would use her custom image classification model "Unicorns vs Narwhals" to detect if her drawings were a narwhal or an unicorn, and also report the confidence score in top corner left (Fig. 3.3). A character would display the model prediction for each drawing (bottom corner left). In this example, we see the importance of providing children with access to the model confidence scores. While both predictions in this example are correct the confidence scores are very low (0.000036 confidence for the "Narwhal" prediction, and 0.00001 for the "Unicorn" prediction). This feedback was important for children so they could improve their models and include more data in their training set. Such experiences also allowed children to become more skeptical of predictions they get not only in their game but also in the real world and understand what goes behind the scenes of the image prediction. Children could add new data to their models either directly from their coding projects by using the dedicated blocks (see 2 in Fig. 3) or by using the Teach AI page (see 1 in Fig. 3).

3.3.2 *Learning Activities.*

During the study, all children completed 3 main learning activities: the "Make me Happy Program" using text training, the "Rock Paper Scissors Program" using image training and the "Smart Home Program" using text and speech training. The children were allowed to choose if they want to do the activities together or individually and were provided printout materials to support the activities. The printout materials would provide children with prompts to lead them to decide what data to include in their training and with code examples that could be used during the programming stage. Researchers would also walk around when children engaged in these learning activities and prompt students with understanding probing questions (i.e. "why do you think it does that?", "how would you fix it?"). After the children finished the 3 learning guides they were encouraged to play and modify other smart programs on Cognimates AI platform.

"Make me Happy Program". We started with very simple text training activities like "Make me Happy Program". In this activity, the students had to teach the computer through the Cognimates "Teach AI" platform how to recognize funny messages or serious messages. Once the model was trained with their examples the students could use it in a pre-coded starter project which would make a character on the screen or one of the robots react to their messages. If the message they gave was classified as "funny" based on the model they trained, the character or robot would be "happy". If the message was classified as "serious", the character or robot would be "thinking". The "Teach AI" text models would require them at least 2 categories (e.g. "funny" and "serious") and five examples of text for each category. The text could be one word or an

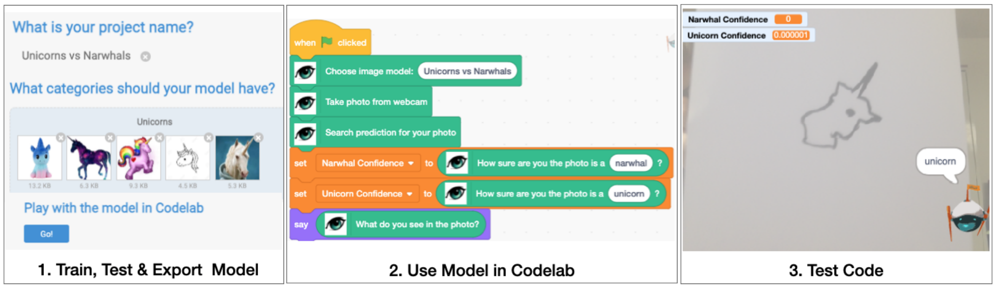

Figure 3: Prediction program for a custom image classification model that can recognize Narwhals and Unicorns

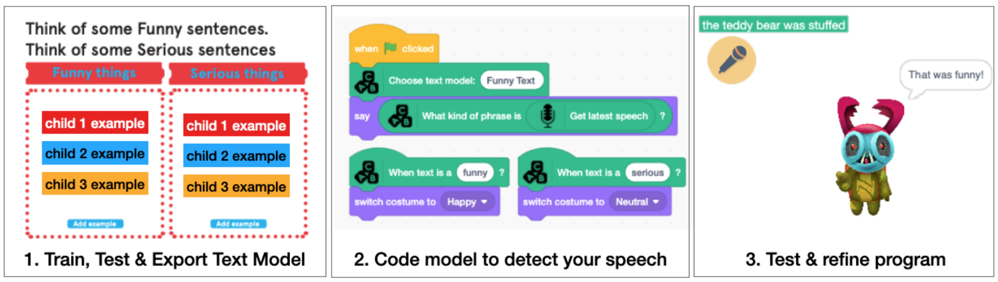

Figure 4: "Make me Happy Program"- Text Training Learning Activity

entire phrase. On average the text models would take 2-3 minutes to train.

"Rock Paper Scissors Program". After learning about text training, we introduced children to image training in the "Rock Paper Scissors (RPS) Program". In this activity, the students had to teach the computer how to recognize images of hands showing "rock", "paper" or "scissors". Once the model was trained with their examples, the students could use it in a coding project to test the RPS program with the computer via the webcam. Once they finished coding the program they would test it together with their friends. If the program would fail to recognize some of their hand gestures they would retrain the model to include the new gesture images.

"Smart Home Program". In this activity, the children had to teach the computer how to recognize different commands for turning the lights on and off. They first trained a text model to recognize different types of commands for "lights on" and "light off". Once the model was trained with their examples the children could use it in a coding project in order to control internet connected lights via voice commands.

Other smart programs. After children completed the 3 main learning activities presented above, they were encouraged to test and modify other smart programs. The most popular projects were the *Jellyfish game*, where a jellyfish floats only if you tell it happy messages (using Speech and Sentiment analysis blocks), the *Good boy* program where a dog reacts with sounds and animations to how you talk to it (using Speech and Text classifier blocks). Children often wanted to modify the projects to make both the characters more expressive and to add new types of messages the characters could react to.

3.3.3 *Perception Game.*

To answer RQ2 about children's shift in perceptions, we used an AI Perception questionnaire adapted from Bartneck et al. [6]. This is an existing instrument that has been frequently used to measure children's anthropomorphism, animacy, perceptions of likeability, perceptions of intelligence, and perception of safety of robots. The original instrument examines perceptions across 24 items. Because 24 items was too numerous for our age group, we adapted the items to specifically focus on a subset of 5 characteristics: *it understands me*, *it is smarter than me*, *it will remember me*, *it tells the truth*, *it is friendly*, and *it likes me*; we also reduced the levels to just three: the two endpoints of each the scale (*yes* and *no*) and a *maybe*. Finally, rather than presenting the instrument as a survey, we presented as a "Perception Game", to more effectively engage younger children. In our game, there were a series of printed statements who share a belief about a smart agent. Before asking the questions, the researchers gave an example of how to respond. We conducted the game separately for each of three agents: Alexa, Cozmo, and Jibo. The children were asked to place a sticker closer to the statement with which they most agreed. At the end of the questions researchers wrote the child's name next to their sticker and take a picture in order to be able to later identify the answers.

3.4 Data Collection and Analysis

Our study resulted in pre- and post- perception game data as well as video recordings of all sessions at all sites. For the qualitative analyses, the first author and a team of five undergraduate students transcribed the videos and also noted comments on children's body language and non-verbal interactions. The final corpus included 100 pages of transcripts (34,300 words). Once all the transcriptions were

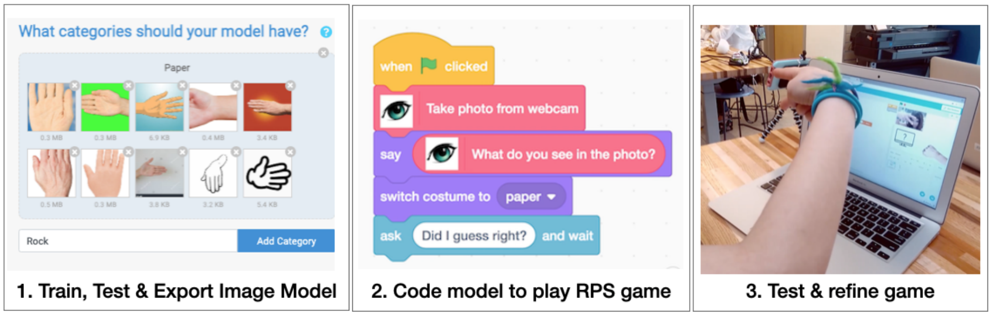

Figure 5: "Rock Paper Scissors Program"- Image Training Learning Activity

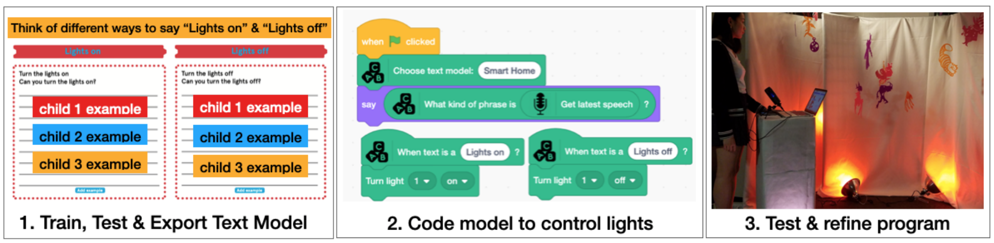

Figure 6: "Smart Home Program"- Text & Speech Training Learning Activity

finished, the authors each reviewed half of the data independently, looking for ways of explaining the three phases of the study. In this process, the authors separately analyzed each transcript using a combination of etic codes developed from our theoretical frameworks and emic codes that emerged from the interviews themselves [45, 52]. We listed all the the sense-making practices [79] that were found in prior studies with kids and science or math learning [80] and identified connections with a series of themes that emerged from our study. After a final coding frame was developed, all the transcripts were coded by the first author. If new codes emerged, both authors discussed discrepancies in the analyses until they reached agreement. The coding frame was changed and the transcripts were reread according to the new structure. The final list of codes, their definitions and presence across the different study sites is presented in Table 1. This process was used to develop categories, which were then conceptualised into broad themes after further discussion. Towards the end of the study no new themes emerged, which suggested that major themes had been identified [10].

4 RESULTS

In this section, we present an overall summary of our perceptions of children's experiences, then discuss our results to RQ1 (how children made sense of agent behavior) and RQ2 (how children's experiences programming AI impacted their perceptions of agents).

4.1 RQ1: How do children make sense of machine intelligence when training smart programs?

Within the rich experiences described in the previous section, we now turn to a more granular analysis of children's collaborative sense-making of agent behavior. Overall, our qualitative analysis revealed a clear pattern of behavior: children engaged in a scientific process of initially formulating hypotheses about a smart object behavior, then they came up with scenarios for testing the hypotheses via interaction with the device or with peers, and finally they refined their understanding of the technology either by affirming their initial hypotheses or coming up with new ones. What varied were the tactics that children used to conduct these empirical investigations. To illustrate this variation, we present several tactics that emerged from our inductive analysis.

4.1.1 ***What type of hypotheses did children propose?*** One major source of variation was how children generated hypotheses to investigate.

Social Judgement Hypotheses. Based on our analysis, some hypotheses appeared to be formed by a social judgement of intelligence, where the children analyzed the agent's social behavior and inferred intelligence from it. They would make these hypotheses while interacting with the devices for explaining why they perform an action or not (e.g., "she did not listen" for explaining why Alexa won't play a song). They would also anthropomorphize the devices when they would hypothesize if they are friendly or trustworthy during the initial AI perception game as seen in the following examples:

Code		Definition	*Study locations* A	B	C	D
Initial Hypotheses	*Social Judgement*	Child analyzed the agent's social behavior & inferred intelligence from it	x	x	x	x
	Funds of Knowledge	Child is using prior experiences to form theories about behavior	x	x	x	x
	Egocentric	Extrapolating from a child's behavior to the agent's behavior	x	x		
	Observational	Objective details of what the agent is doing without social inferences	x	x	x	x
	Agency	Child would question if the device had agency or not			x	x
Test Assumptions	*Edge Cases*	Testing via edge cases to understand the limits of the agent's intelligence			x	x
	Common Cases	Testing via common cases to reveal deeper understanding	x	x	x	x
	Agency	Testing to see if the agent is autonomous or not	x	x	x	x
Refined Hypotheses	*Post-test Behavior*	Used test results to build more complex models of agent's behavior	x	x	x	x
	Social Intelligence	Used judgements of social intelligence to refine models behavior	x	x		
	Programmability	New hypotheses of machine intelligence referencing programming	x	x	x	x
	AI Training	New hypotheses of machine intelligence referencing AI training			x	x

Table 1: List of codes used for transcripts analysis from the different study sites: A - after-school program in a public school, B - non-profit community center, C - after-school program in a private school, D - private STEM center

> *"He just seems like he's in something else right now"* — B., age 12, referring to a Cozmo robot. *"I think he cares about me because when I ask him something, he listens instead of just not even caring about what he says"* — C., age 7, referring to Jibo robot. *"Well, sometimes I ask a question and she says she doesn't know and I'm not completely sure if she's actually telling the truth"* — A., age 7, referring to Alexa. *"She has more of a human personality but she still like doesn't have emotions and the friendliness part"*-Si. age 10, referring to Alexa.

Funds of Knowledge Hypotheses. Our analysis showed that some hypotheses seemed to emerge from funds of knowledge, using prior experiences to form theories about the agent's behavior. This practice is consistent with children's sense making practices in other domains like agent simulations in physics [80] or mathematics [66]. Children referenced not only personal past experiences in interacting with computers or other similar AI agents but also examples and stories they heard about in the media or from their friends and parents.

> *"Uh, maybe they coded something on the computer to tell it, like, tell the computer what to do. Sort of like the computer's brain, computer is the brain"* — C., age 6.5, referring to a Cozmo robot. *"She will remember you because I'm pretty sure just like Siri you can tell her your name, to like ask her to remember you, like who you are. Because you can tell them your name"* — E., age 8, referring to Alexa.
> *"You have to say what text is bad and what text is happy or maybe backhanded, and over time, it'll be able to recognize it without you telling. And, um, I remember seeing a video on the Avengers about why there were such split rates, and uh, the people made a bot"* -Ch., age 7, referring to the text training for sentiment analysis.

Egocentric Hypotheses. In our examination we saw that some hypotheses seemed to emerge from egocentric speculations, extrapolating from the children's ideas about how they would perform a task or solve a problem to the agent's behavior. This was consistent with Papert's findings on body syntonicity, where children project robot geometrical puzzles on their own body to solve a differential mathematics problem in Logo [51].

> *"Well, I've seen lots of pictures and even if I've never seen what, like, a train that has purple stripes, I would just know it's a train by the way it looks, not by its color"* — So., age 8, referring to custom image model trained to recognize trains. *"I think they learn kind of the same and kind of different, because when we learn stuff, we can forget it, but then we can look for it in the real world. But, computers almost never forget it, but if they forget it, they can't look for it in the real world"* — L.,age 7.5, referring to how the Jibo robot learned to recognize faces.

Observational hypotheses. Based on our analysis, we found that some of the children's hypotheses built upon what they had seen the agent do. In this case children would describe details of what the agent was doing without drawing social inferences.

> *"Because it has to recognize every bit, every single thing that's green. If I said 'green', but put this [the green balloon] with like, a background of something else, it might not recognize that because it's supposed to recognize the bigger things as green"* — R., age 7.5, referring to a color sensing coding project. *"I think it works because it says, umm, when you hear good, or happy speech, then, go up and when it hears bad, it just says go down. Then it says when you're out of bounds, make beeping sounds. And when you hit the side, switch directions"* — E., age 8, describing the Jellyfish coding project.

Agency hypotheses. Our analysis showed that participants proposed several hypotheses when asked to evaluate if the agents were smart, trustworthy or human like. Most children proposed these hypotheses during our pre- and post- group discussions about agent intelligence. Children shared beliefs such as:

> *"It's programmed to always tell the truth"* — J., age 8, referring to Jibo. *"Yes, and I think they programmed her so she acts nice"* — L., age 7.5, referring to Alexa. *"Then, the computer would learn, and then it would try to fix it's mistake"* — As., age 7.

These varying sources of hypotheses show that if children believed a smart agent was controlled or programmed by someone else, then they would be more skeptical of its intelligence and human-like abilities. In turn, if the children believed the agent was in control, they would tend to overestimate its abilities to perform human tasks.

4.1.2 ***How did children test their initial assumptions?***

Whereas children's sources of hypothesis were highly varied, our analysis found less variation in how children tested hypotheses, with most directly interacting with agents. What varied were the types of test *cases* that children chose to probe agent behavior.

Testing edge cases. Our analysis showed that children would come up with a variety of edge cases in order to understand the limits of the agents' abilities and intelligence. Children would use all the resources they had in their arsenal to test the agents: from using niche cultural references, to speaking in a different language or trying to find examples of images that are very confusing (e.g., images of dogs with sunglasses). We interpreted this to be similar to practices observed in studies on AI understanding with the use of counterfactual examples [2, 77], children in our study would build on responses they would elicit from the agents in order to identify more and more narrow edge cases.

> *"Alexa, play a legendary Kirby rap on Spotify"* — Ch., age 7, talking to an Alexa smart speaker. *"We are trying to confuse it by getting a puppy that looks kind of like a Kirby that is wearing sunglasses"* — P., age 8.5, referring to their custom image classification game.
> *"We tried to make him say poems but he wouldn't do it"* — Do., age 7.5, referring to a Jibo robot.

Similar to other examples of playful debugging [39], children would take great pride when they would find a case that would confuse or trick the agent and they would share their discovery with their peers.

Testing common cases. We found that this type of testing was used by children when trying to reveal deeper understanding. In the instances where they knew something should work but it did not, children would try to infer the reasons for failure and come up with other similar examples in order to test their assumption:

> *"He's seeing the colors - it's true because I'm showing the balloon and if I take it away, it's false. Show it, true, hide it, false, yeah? So, now, if you show the color, the paddle should move (paddle doesn't move)"* — A.& E., age 8 & 6.5, debugging their color sensing project. *"Cause I put baseball bat, not a baseball, but somehow they must have looked kinda similar, I don't know, and it did it"* — D., age 9, referring to his custom image classification program.

Testing agency. Our analysis showed that children used a series of strategies in order to test the nature of the agents they were interacting with and tried to more accurately place the devices on the animate/inanimate spectrum [26, 28, 34]. Participants would either directly ask questions to the devices about their nature (e.g., "How do you work?", "Who made you?") or they would come up with play scenarios to see if they could get the agent to embody different personas. Sometimes children would physically cover the devices, disconnect them from the internet or move them in the room to test how they would behave, similarly to children's attempts to make sense of social robots like Cog [73].

> *"Alexa, does everything you say really get texted to someone?"* — E., age 8. *"I'm trying to figure out how to make it, um, say, when I say 'I am potato', or, or, I want to say 'Are you a potato?' then it will say 'yes' "* — A., age 7, referring to Jibo robot.
> *"I think it goes to the internet, but if the internet does not have connection, she'll say, okay, it was nice talking with you"* — Si., age 8.5, referring to Alexa.

4.1.3 ***When and how did children refined their understanding?***

The testing practices in the previous section were often a precursor to children refining their understanding of agent intelligence. These moments were particularly observable when children were listening and debating perception questions with other children and after several sessions of coding and training where they gained more insight into how smart agents learn from examples. We observed children make several types of inferences from their hypothesis testing.

Post-test Behavior Hypotheses. Some children used the results of testing to build more complex models of the agent's mechanics:

> *"It can make a mistake. Someone could have made a mistake in programming but it is supposed to"* — So., age 8, referring to an Alexa smart speaker. *"Like, if I taught it face recognition, I would go like, oh, this is the real face, no this is the real face, no this is the real one, and it would be really mixed up and it wouldn't know who is who"* — D., age 9, referring to Cozmo robot.

Judgements of Social Intelligence. Some children used social judgements of intelligence to refine their models of the agent's abilities:

> *"He doesn't even know how to pick up a block when I say pick a block"* — N, age 6, referring to Cozmo robot.
> *"Because the Alexa, sometimes I asked her questions and she doesn't understand and sometimes I ask her and she knows it"* — C., age 6.5. *"Because my mind is going both. It will remember me, but it won't, I just can't"* — G., age 6, referring to Jibo robot.
> *"If the computer knows how to learn, I think it would be easier to make it, um, a robot version of a person, because it can learn like a person, and then it could probably think like a person, move like a person, and act like a person. And then, someday, someone - a person in your house - could be a robot."* — L, age 7.5.

References to Programmability. Many children exhibited more elaborate explanations of agent behavior, grounded in the concrete activities of programming. For example, C., age 6.5, initially said the Alexa smart speaker makes mistakes because *"sometimes she says she doesn't know"*. In the final perception discussion she described the same device in the following way:

> *"Surprised me the most that, at first I didn't really know computers got taught. I thought computers, once they were invented, knew stuff. I didn't know they got taught to do rock paper scissors and all that"* — Em., age 6.5, referring to her experience in the study.
> *"I think its smarter but a person created it so its as smart as the person(referring to creator) but programmed to be smarter"* — said M, age 7.5 *"I know how to use these ["forever" coding blocks]. I've coded using a different program before. Oh my god, this is going to be so cool, I want to use them for the robot now"* — P., age 8.5, referring to his prior coding experience.

References to AI Training. Many children specifically grounded their judgements in their experiences with AI teaching and training. We observed that children would often try to confuse the AI by showing it examples that combine the different things it is trained to recognize (e.g., dog with glasses). The experience of confusing the robot or the computer was primarily attractive for children because it was perceived as fun and because it put them in charge of the process. This led to inferences about the limitations of AI:

"I think I know. Well maybe because that one looks more like a drawing. And it doesn't get it because it's a drawing" — R., age 7, referring to her model prediction result. *"Because it's going to learn what those pictures are going to be"* — L, age 7.5, referring to So.'s model also. *"Cause I don't wanna put more funny words or more boring words, cause if I put like, 2 funny words and 1 boring word, it would probably put funny"* — T, age 7.5, after typing "doing funny homework" in his prediction game. *"Probably it did see me, but it didn't really recognize me, but he can learn to recognize me"* — said a Em., age 6.5, referring to the Cozmo robot.

Peer Support. Children would often support each other in refining their understandings either by explaining how a specific program works or by providing alternative answers to group discussions. Although children tested their initial assumptions about agents either with edge or common cases, when it came to testing the agency of the devices, their testing strategies were based more on peer-aided judgements and examples. The way children explained their reasoning for their answers influenced each other which lead them to internalize new explanations and concepts presented by their peers. For example, here are two exchanges between children facilitating each others' testing:

Discussion 1: *"So how many examples of trains do you need to give it so it can recognize multiple trains?"*—researcher (referring to a custom image classification model). *"165 million"* answered N., age 6. L. and M.,ages 7.5 and 8, added *"10, at least 10"*. Discussion 2: *"I actually don't really know how to use the app"* — C.,age 6.5, referring to a program made for Jibo robot. *"Ah, I got an idea, see if I can get it to look in different places. 'Swipe up', 'swipe right', I'm just trying to see if I can make it look up when I swipe up. By the way, it won't say 'hi' 10 times, but it'll say anything you put in here 10 times, and you can edit this more if you want"*-Ch., age 7, replying to C.'s question about the program.

Overall children refined their understanding of smart agent behavior by evaluating their test findings and coming up with new interpretations for the agents behavior and new judgements to explain their social intelligence. In the final sessions, children will make more complex references to programmability and build on their peer support to refine their AI explanations.

4.2 RQ2: How does children's perception of machine intelligence change?

The previous section showed that children examined and reasoned about agency in diverse ways. In this section we consider how these varied forms of reasoning led to changes in self reported judgements about smart agents' abilities.

As discussed earlier, we measured these changes using pre- and post- answers of the Perception Game during the initial and last session in each location. Unfortunately, due to snow, our Public After-School Program's last session was canceled, and we only report results from the three sites that completed the pre/post. Additionally, at some sites, we only asked 5 of the 8 Perception game questions because the children became too impatient to answer all the questions. Many of the children changed their answers to the perception questions pre- and post- (see Fig.7). Children became more skeptical of the agent's human like abilities, such as remembering them or being friendly. For instance, even when the children said that the agent is friendly or that it will remember them, they would explain it was programmed to do so:

"Because, he looks like he has feelings, but he might not. You can make him, like, sad, happy, surprised, bored."— L. 7 years old. *"He's a robot, so he's probably going to have lots of things programmed into him that he knows and he doesn't have to remember them. Humans have to remember the stuff, but robots don't."* — A. 8 years old.

To understand whether any of these shifts were statistically significant, we did the following. For each of the three completed sites, and for each of the 5 completed questions in the AI Perception game, we performed a Man-Whitney U test with the dependent variable being the answers to questions measured on an ordinal scale ("yes" answer as 1, "maybe" answer as 2, "no" answer as 3) and the independent variable being the pre- and post- conditions. The five questions included two questions about intelligence and legibility attribution (*Is the agent smart? Does it understand me?*), and three questions about socio-emotional attributes *(Will it remember me? Does it care about me? Is it friendly?)*. We used a conservative Bonferroni correction for multiple comparisons, setting our alpha to .05/5=0.01 for each site.

Fig.7 shows the distribution of responses. Of the 5 tests, two were significant across all the students participating in the study. Overall all students in the program were more likely to answer no to the question *Will the agent remember you?* after the program ($U = 355$, $p = 0.00424$). Similarly, a significant change in ranks of the children that initially said that the agents are friendly changed their answers to "no" and "maybe" at the end ($U = 149.5$, $p = 0.01684$). We did not find statistically significant changes in the other measures; the only other trending shift was an increase in the number of students who said "no" to *Is the agent smarter than you?*.

One noticeable trend in the data is that there are more "maybe" answers in the post- than in the pre-(see Fig. 7). This could possibly indicate that, after being exposed to the programmability of AI machines and thinking critically about the machine's agency, the children were reasoning about the complexities and ambiguities of machine intelligence at a higher level in the post- than in the pre-perception discussions.

5 DISCUSSION

Our work contributes several new insights about AI literacy by addressing out initial research questions:

- RQ1: *How do children make sense of machine intelligence when training smart programs?* Our qualitative results show that children engage in the scientific method by formulating hypotheses about machine intelligence, then coming up with scenarios for testing, and finally refining their understanding either by affirming their initial hypotheses or formulating

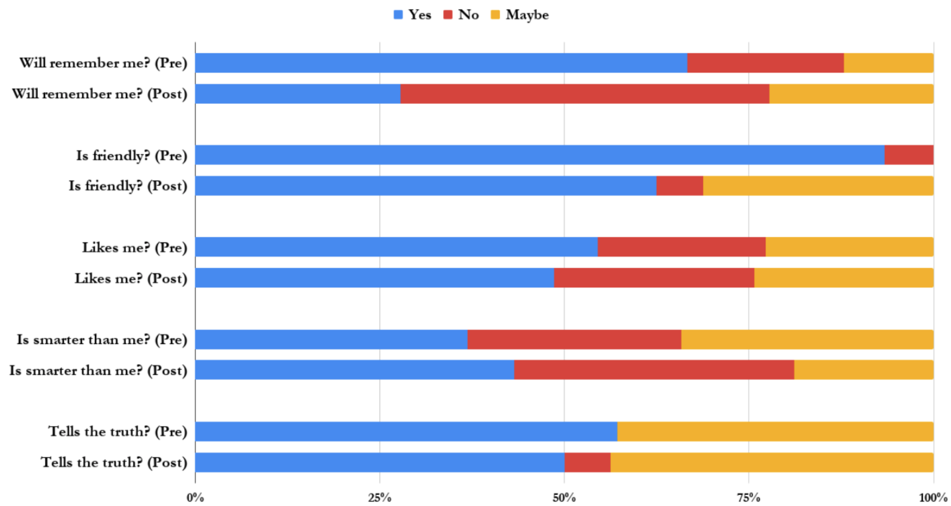

Figure 7: Answers shift from five Perception Game pre- and post- answers in all locations

new ones. In this process children use a diverse set of social sense-making strategies, drawing from their egocentric perceptions of agency and their empirical observations, to make inferences about agency.

- RQ2: *How do children's perception of machine intelligence change before and after training smart programs?* Based on the pre-post shifts we observed, children's perception of machine intelligence trended toward skepticism. Children also decreased their pro-social attitudes toward the smart agent's behavior. We did not observe changes in children's perception of an agent's truthfulness or ability to like them.

Our results suggest that engaging children in programming with AI leads many children to replace conceptions of smart agents as intelligent with new conceptions of smart agents as fallible but helpful. Importantly, these shifts did not occur for all children, nor did they occur in the same directions, suggesting the challenges of promoting a specific conception of machine intelligence through programming.

Limitations. Some limitations in the study complicate the interpretation of our findings. It was not possible to systematically observe every child's interaction with every agent, nor did every child speak in every group; it may be that children who did verbalize more reasoned differently than those who verbalized less. For the interactions we *could* observe, observing a child reason about an agent does not necessarily indicate ground truth for their conceptions; for example, it may be the case that children were reasoning in similar ways but were verbalizing their reasoning differently. We also did not have data for all perception questions and all sites, nor did our sites cover the many possible ways that culture, community, and collaboration might have shaped sense-making. Since our analysis was episodic rather than temporal, sense-making strategies may have been highly variable within individual and group behavior. Therefore, while modest interpretation of our results suggest that the children in our particular intervention demonstrated diverse reasoning strategies and a shift toward skepticism, other populations could reveal new types of sense-making and different shifts in perceptions.

Programmability impacts intelligence perception. Despite these limitations, our results have many implications for interpreting prior work. For example, as we shared earlier, prior studies on smart agents has shown a clear trend of anthropomorphism, especially of embodied agents [34, 47, 68]. Some studies have even shown that embodied agents can exert peer pressure over children [76] and that children can overestimate the intelligence of embodied agents [21]. Our results show that one reason for this susceptibility is that children have not engaged in examining the mechanisms and limits of AI; when children in our study engaged in this examination, their conceptions of smart agents were still anthropomorphizing, but often less trusting in machine intelligence. These findings are consistent with Duuren's results that identified programmability as a key element in children's perception of social robots' abilities [23]. In another experiment, Vollmer et. al found that 7- to 9-year-old children had a tendency to echo the incorrect, but unanimous, responses of a group of robots to a simple visual task [76]. Thus, the trends in prior work may be conditioned upon what experiences children have had with programming with AI.

Sense-making for AI literacy. Our results also have implications for prior work on children developing AI literacy. Prior work has revealed many challenges, including the importance of children understanding the role of data in shaping machine behavior [46] and the persistent challenge of debugging and comprehension [69]. Other studies with adults has explored methods of bridging these comprehension gaps by helping people develop more robust mental models about AI (e.g., [5, 38, 61]). Our findings suggests that similar approaches may work for children, at least when children are engaged in constructing projects that use AI techniques. Our qualitative findings about children's sense-making strategies also suggest new interpretations of prior research on program understanding. Whereas prior work has largely focused on individual, cognitive accounts of program understanding (e.g., [1, 36]), our investigation of program understanding from a constructionist [50] and social sense-making [18] lens suggests that children rely on numerous assets beyond cognition to understand agent behavior. These assets include social strategies for enacting scientific activities such as observation with peers, discussing hypotheses with peers, as well as introspective, egocentric strategies for inferring models of agent behavior.

Platform design choices. Importantly, our result do *not* speak to work on data literacy. For example, prior work has shown that children engaging with and making sense of data itself has its own challenges [37], as does reasoning about statistics [12]. Our design choices in Cognimates intentionally abstracted and scaffolded away these challenges in service of engaging children in examining agency. Different designs and pedagogies would likely be necessary to promote these different literacies.

Guidelines for designers & educators. Of course, all of these findings have implications for both designers of learning technologies and AI literacy teaching methods for children. Our work selected particular scaffolding to support more accurate assessments of intelligence. While our results were not granular enough to point to specific aspects of this scaffolding that contributed to the strategies and shifts in beliefs that we observed, our work does generate concrete hypotheses to investigate in research and practice. For example, one clear trend in our results was that some children attempted to take the perspective of the agent to reason about its capabilities, trying to imagine how it was making decisions to make inferences about it's capabilities. Designers and teachers might therefore consider methods for promoting perspective taking about AI agents, just as similar work on programming language learning has encouraged learners to take the perspective of a compiler [40, 48]. Another clear trend was that children used their experiences in generating training data to make inferences about agent ability. Designers and teachers might explore methods for engaging children in reflecting on the relationship between the training data, the agent's use of that data, and it's resulting behavior.

Future work. While these implications for design are modest, the need for future work is clear. The results in this paper demonstrate the feasibility of promoting more accurate estimations of intelligence, and begin to reveal the mechanisms behind those changes, but many questions remain about how robust—or repeatable—these changes are in different settings, with different instructors, on different platforms, and using different assessments. Future work should explore these variations, but also extend them to longitudinal observations to understand the robustness of these conceptions over time, and the degree to which they transfer to non-learning settings such as home, play, and adulthood.

6 CONCLUSION

After a 4-week observational study in after-school programs, we found programming with AI leads many children to replace conceptions of smart agents as intelligent with new conceptions of smart agents as fallible but helpful. If we can build a robust understanding of how to promote AI literacies, we will be much better positioned to respond to a future in which AI is embedded in children's' everyday lives. By enabling inclusive AI literacy we will help democratize AI education [20, 41], and by increasing children's AI literacy we would allow them to responsibly use smart technologies for creative learning and personal expression [58]. This vision *must* be attained if our children and our children's children are to live in a just and equitable society.

ACKNOWLEDGMENTS

Thank you to Benjamin Kushigian, Regina Cheng, Benjamin Mako Hill for their help in reviewing and providing feedback on this study. We also want to thank Sarah T.Vu, Eesh Likhith, Tammy Qiu, Steven Dale for their contributions to the Cognimates platform and support in running the study workshops.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Jose M Alonso. 2019. Explainable Artificial Intelligence for Kids. In *2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)*. Atlantis Press.

[3] Morgan G Ames. 2015. Charismatic technology. In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives*. Aarhus University Press, 109–120.

[4] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 701–706.

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[6] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[7] Mark Bauerlein. 2011. *The Digital Divide: Arguments for and Against Facebook, Google, Texting, and the Age of Social Netwo rking*. Penguin.

[8] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A Kientz, Jason Yip, and Alexis Hiniker. [n.d.]. Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. ([n. d.]).

[9] Paulo Blikstein. 2012. Bifocal modeling: a study on the learning outcomes of comparing physical and computational models linked in real time. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 257–264.

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[11] Rita Brito, Rita Francisco, Patrícia Dias, and Stephane Chaudron. 2017. Family dynamics in digital homes: the role played by parental mediation in young children's digital practices around 14 European countries. *Contemporary Family Therapy* 39, 4 (2017), 271–280.

[12] Rosemary Anne Callingham. 2006. Assessing statistical literacy: A question of interpretation?. In *International Conference on Teaching Statistics (ICOTS)*. 1–6.

[13] Yu Chi, Wei Jeng, Amelia Acker, and Leanne Bowler. 2018. Affective, behavioral, and cognitive aspects of teen perspectives on personal data in social media: A model of youth data literacy. In *International Conference on Information*. Springer, 442–452.

[14] Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. 2010. Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the sixth symposium on usable privacy and security*. ACM, 15.

[15] Sayamindu Dasgupta and Benjamin Mako Hill. 2017. Scratch community blocks: Supporting children as data scientists. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3620–3631.

[16] Erica Deahl. 2014. Better the data you know: Developing youth data literacy in schools and informal learning environments. *Available at SSRN 2445621* (2014).

[17] Peter J Denning and Matti Tedre. 2019. *Computational thinking*. MIT Press.

[18] Brenda Dervin. 1998. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of knowledge management* 2, 2 (1998), 36–46.

[19] Betsy DiSalvo, Cecili Reid, and Parisa Khanipour Roshan. 2014. They can't find us: the search for informal CS education. In *Proceedings of the 45th ACM technical symposium on Computer science education*. 487–492.

[20] Stefania Druga, Sarah T Vu, Eesh Likhith, and Tammy Qiu. 2019. Inclusive AI literacy for kids around the world. In *Proceedings of FabLearn 2019*. ACM, 104–111.

[21] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 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*. ACM, 595–600.

[22] Stefania Druga, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2018. How Smart Are the Smart Toys?: Children and Parents' Agent Interaction and Intelligence Attribution. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. ACM, New York, NY, USA, 231–240. https://doi.org/10.1145/3202185.3202741

[23] Mike Van Duuren. 1998. Gauging Children's Understanding of Artificially Intelligent Objects: A Presentation of"Counterfactuals. *International Journal of Behavioral Development* 22, 4 (1998), 871–889.

[24] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 5.

[25] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3 (2003), 143–166.

[26] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3 (2003), 143–166.

[27] Krista Francis, Steven Khan, and Brent Davis. 2016. Enactivism, spatial reasoning and coding. *Digital Experiences in Mathematics Education* 2, 1 (2016), 1–20.

[28] Heather M Gray, Kurt Gray, and Daniel M Wegner. 2007. Dimensions of mind perception. *Science* 315, 5812 (2007), 619–619.

[29] Shuchi Grover and Roy Pea. 2013. Computational thinking in K–12: A review of the state of the field. *Educational researcher* 42, 1 (2013), 38–43.

[30] James K Hammerman and Andee Rubin. 2004. Strategies for managing statistical complexity with new software tools. *Statistics Education research journal* 3, 2 (2004), 17–41.

[31] Tim Ingold. 2001. Beyond art and technology: the anthropology of skill. *Anthropological perspectives on technology* (2001), 17–31.

[32] Katherine Isbister and Clifford Nass. 2000. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International journal of human-computer studies* 53, 2 (2000), 251–267.

[33] K Megasari Kahn, Rani Megasari, Erna Piantari, and Enjun Junaeti. 2018. AI programming by children using snap! block programming in a developing country. (2018).

[34] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction* 19, 1 (2004), 61–84.

[35] Caitlin Kelleher and Randy Pausch. 2005. Lowering the barriers to programming: A taxonomy of programming environments and languages for novice programmers. *ACM Computing Surveys (CSUR)* 37, 2 (2005), 83–137.

[36] Amy J. Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad Myers, et al. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 1–44.

[37] Clifford Konold, Traci Higgins, Susan Jo Russell, and Khalimahtul Khalil. 2015. Data seen through different lenses. *Educational Studies in Mathematics* 88, 3 (2015), 305–325.

[38] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

[39] Michael J Lee, Faezeh Bahmani, Irwin Kwan, Jilian LaFerte, Polina Charters, Amber Horvath, Fanny Luor, Jill Cao, Catherine Law, Michael Beswetherick, et al. 2014. Principles of a debugging-first puzzle game for computing education. In *2014 IEEE symposium on visual languages and human-centric computing (VL/HCC)*. IEEE, 57–64.

[40] Michael J Lee and Andrew J Ko. 2011. Personifying programming tool feedback improves novice programmers' learning. In *Proceedings of the seventh international workshop on Computing education research*. 109–116.

[41] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[42] Dorian Lynskey. 2019. 'Alexa, are you invading my privacy?' – the dark side of our voice assistants. https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants/. [Online; accessed 19-January-2020].

[43] Katie Makar and Andee Rubin. 2009. A framework for thinking about informal statistical inference. *Statistics Education Research Journal* 8, 1 (2009).

[44] Radu Mariescu-Istodor and Ilkka Jormanainen. 2019. Machine learning for high school students. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*. 1–9.

[45] Matthew B Miles and A Michael Huberman. 1984. Drawing valid meaning from qualitative data: Toward a shared craft. *Educational researcher* 13, 5 (1984), 20–30.

[46] David Mioduser and Sharona T Levy. 2010. Making sense by building sense: Kindergarten children's construction and understanding of adaptive robot behaviors. *International Journal of Computers for Mathematical Learning* 15, 2 (2010), 99–127.

[47] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[48] Greg L Nelson, Benjamin Xie, and Andrew J Ko. 2017. Comprehension first: evaluating a novel pedagogy and tutoring system for program tracing in CS1. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*. 2–11.

[49] Philip Oltermann. 2017. German parents told to destroy doll that can spy on children. *The Guardian* (2017).

[50] Seymour Papert. 1980. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.

[51] Seymour Papert. 1993. *The children's machine: Rethinking school in the age of the computer.* ERIC.

[52] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*. SAGE Publications, inc.

[53] Blakeley H Payne. 2019. An ethics of artificial intelligence curriculum for middle school students. *MIT Media Lab Personal Robots Group. Retrieved Oct* 10 (2019), 2019.

[54] Rachel Rabkin Peachman. [n.d.]. Mattel Pulls Aristotle Children's Device After Privacy Concerns. "https://www.nytimes.com/2017/10/05/well/family/mattel-aristotle-privacy.html", year = 2017, note = "[Online; accessed 19-January-2020]".

[55] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, 640.

[56] Byron Reeves and Clifford Nass. 1996. *How people treat computers, television, and new media like real people and places.* CSLI Publications and Cambridge university press.

[57] Yim Register and Amy J. Ko. 2020. Learning Machine Learning with Personal Data Helps Stakeholders Ground Advocacy Arguments in Model Mechanics. In *Proceedings of the 2020 ACM Conference on International Computing Education Research* (Virtual Event, New Zealand) *(ICER '20)*. Association for Computing Machinery, New York, NY, USA, 67–78. https://doi.org/10.1145/3372782.3406252

[58] Mitchel Resnick, Amy Bruckman, and Fred Martin. 1996. Pianos not stereos: Creating computational construction kits. (1996).

[59] Maaike Roubroeks, Jaap Ham, and Cees Midden. 2011. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics* 3, 2 (2011), 155–165.

[60] Andee Rubin. 2019. Learning to Reason with Data: How Did We Get Here and What Do We Know? *Journal of the Learning Sciences* (2019), 1–11.

[61] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable aI and users' mental models. (2019).

[62] Michael Scaife and Mike Duuren. 1995. Do computers have brains? What children believe about intelligent artifacts. *British Journal of Developmental Psychology* 13, 4 (1995), 367–377.

[63] Baruch B Schwarz. 2009. Argumentation and learning. In *Argumentation and education*. Springer, 91–126.

[64] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, What's Up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 857–868.

[65] Kazuhiko Shinozawa and Junji Yamato. 2007. Effect of Robot and Screen Agent Recommendations on Human Decision-Making. In *Human Robot Interaction*. InTech.

[66] Robert S Siegler. 2009. Improving the numerical understanding of children from low-income families. *Child Development Perspectives* 3, 2 (2009), 118–124.

[67] Statista. 2019. Digital assistant and voice assistant adoption 2019. (2019).

[68] Fumihide Tanaka, Aaron Cicourel, and Javier R Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17954–17958.

[69] Emmeline Taylor and Katina Michael. 2016. Smart Toys that are the Stuff of Nightmares. *IEEE Technology and Society Magazine* 35, 1 (2016), 8–10.

[70] Anja Thieme, Cecily Morrison, Nicolas Villar, Martin Grayson, and Siân Lindley. 2017. Enabling collaboration in learning computer programing inclusive of children with vision impairments. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. 739–752.

[71] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for k-12: What should every child know about AI?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9795–9799.

[72] S Turkle. 2006. Tamagotchi diary. *The London Review of Books, April* 20 (2006).

[73] Sherry Turkle, Cynthia Breazeal, Olivia Dasté, and Brian Scassellati. 2006. Encounters with kismet and cog: Children respond to relational artifacts. *Digital media: Transformations in human communication* (2006), 1–20.

[74] Junia Valente and Alvaro A Cardenas. 2017. Security & privacy in smart toys. In *Proceedings of the 2017 Workshop on Internet of Things Security and Privacy*. 19–24.

[75] Henriikka Vartiainen, Matti Tedre, and Teemu Valtonen. 2020. Learning machine learning with very young children: Who is teaching whom? *International journal of child-computer interaction* 25 (2020), 100182.

[76] Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. 2018. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science Robotics* 3, 21 (2018), eaat7111.

[77] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[78] Mark Warschauer and Morgan Ames. 2010. Can One Laptop per Child save the world's poor? *Journal of international affairs* (2010), 33–51.

[79] Karl E Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.

[80] Michelle Hoda Wilkerson-Jerde and Uri J Wilensky. 2015. Patterns, probabilities, and people: Making sense of quantitative change in complex systems. *Journal of the Learning Sciences* 24, 2 (2015), 204–251.

[81] Randi Williams, Christian Vázquez Machado, Stefania Druga, Cynthia Breazeal, and Pattie Maes. 2018. My doll says it's ok: a study of children's conformity to a talking doll. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 625–631.

[82] Allison Woodruff, Sally Augustin, and Brooke Foucault. 2007. Sabbath day home automation: it's like mixing technology and religion. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 527–536.

[83] Kun Xu and Matthew Lombard. 2017. Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior* 74 (2017), 152–162.

[84] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K Kane, and R Benjamin Shapiro. 2019. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 121–132.