"It's like I'm the AI": Youth Sensemaking About AI through Metacognitive Embodiment

Eric Greenwald
Ari Krakowski
Timothy Hurt
Kelly Grindstaff
eric.greenwald@berkeley.edu
akrakowski@berkeley.edu
thurt@berkeley.edu
kelly.grindstaff@berkeley.edu
University of California Berkeley's Lawrence Hall of Science
Berkeley, California, USA

Ning Wang
nwang@ict.usc.edu
University of Southern California, Institute for Creative Technologies
Los Angeles, California, USA

ABSTRACT

The increasing presence and importance of Artificial Intelligence (AI) in our society has led to calls for its inclusion at all levels of education. However, the field is only beginning to understand what how AI learning experiences may be designed to be effective and developmentally appropriate, especially for young children. One challenge children encounter is in conceptualizing the "intelligence" of AI while they are still developing a metacognitive model of their own human intelligence. To investigate potential ways to address this, we developed a strategy, *metacognitive embodiment*, through which children are supported to (a) elicit a mental model of their own intelligent performance on a task and (b) compare that elicited model to an AI designed to accomplish the same task. From this study we found evidence suggesting that engaging children in metacognitive tasks in coordination with AI learning experiences (where the AI performs an analogous task) better positioned them for sensemaking about the AI's intelligence.

CCS CONCEPTS

• **Applied computing → Interactive learning environments**; • **Social and professional topics → Informal education**; • **Human-centered computing → Empirical studies in HCI**; **Empirical studies in HCI**.

KEYWORDS

AI literacy, informal learning, embodied interaction, K-5 education

ACM Reference Format:
Eric Greenwald, Ari Krakowski, Timothy Hurt, Kelly Grindstaff, and Ning Wang. 2024. "It's like I'm the AI": Youth Sensemaking About AI through Metacognitive Embodiment. In *ACM Interaction Design and Children (IDC) Conference*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3628516.3659395

IDC '24, June 17–20, 2024, Delft University of Technology in the Netherlands

ACM ISBN 979-8-4007-0442-0/24/06
https://doi.org/10.1145/3628516.3659395

1 INTRODUCTION

As artificial intelligence profoundly reshapes our personal and professional lives [11], there are growing calls for youth to begin developing skills to engage critically and productively with AI as early as elementary school. Long and Magerko [13] define this kind of high-level understanding of AI, or AI literacy, as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (p. 2). This need for an AI-literate public is further amplified by growing concerns about bias, breaches of data privacy, misinformation, and other ethical issues related to AI systems and their deployments.

In response, a rapidly increasing number of AI learning experiences and curricula are being developed for young children. At the same time, research on how best to support younger children's interactions with technology to promote AI literacy is still nascent and primarily focused on facilitated learning environments (e.g., classrooms or afterschool programs), with far fewer studies focused on less structured "free-choice" environments like museums and libraries. Thus, given the important role of public science museums in the STEM learning ecosystem, especially for young children [3, 19], we set out to investigate how young learners interact with and make sense of AI systems in free-choice environments, and what strategies best position them to come away with a new or better understanding of these systems. In particular, we are exploring the interplay between a child's developing mental model of how an AI system operates, and the child's own metacognitive understanding of how they, as humans, perform similar cognitive tasks. Through the design of coordinated digital and non-digital youth interactions that can leverage this interplay, we developed the strategy of *metacognitive embodiment* (McE), which involves eliciting children's metacognitive ideas about their own intelligence in coordination with guided inquiry into how an artificially intelligent agent might perform a similar task. In this paper, we present an early study of McE that offers initial empirical evidence of its promise as an AI learning strategy and explores that evidence to

propose an explanatory account of how it contributes to children's sense-making of AI. The pilot study draws from cognitive interview data, observations, and youth-created artifacts for 19 children aged 6-12 as they engaged in two coordinated activities, one digital and one non-digital, in randomly varying order.

2 RELATED WORK

Metacognition and learning. Metacognition has long been recognized as a powerful learning strategy, especially in promoting conceptual understanding, problem solving success, and in catalyzing conceptual change (e.g., [22]). Despite early assumptions to the contrary, capacity for metacognition has been demonstrated to develop rapidly during a child's first 6 years of life (e.g., [15, 17, 23]).
Embodied interaction. Research suggests that experiences where youth relate their own behaviors to that of a computational model can facilitate learning about computation [8, 14, 18, 27]. Dourish [7] positions embodied interaction in relation to both physical and social aspects of reality, highlighting how these interrelated dimensions of interaction "exploit our familiarity and facility with the everyday world—whether it is a world of social interaction or physical artifacts." Thus, we see opportunity to expand the idea of embodied learning to include "metacognitive embodiment" as a form of embodiment where the learner "acts out" the thinking of an AI agent. In particular, we draw upon Dourish's [7] framing by leveraging the intimately familiar, everyday experience of making and recognizing human facial expressions to scaffold young learners' early sensemaking about AI classification technology.

AI learning among young children. Evidence from recent work [21] points to engagement with human emotional expression as a promising approach for supporting young children in AI learning, drawing on the idea that making sense of cues about human emotion is a ubiquitous and central element of their ongoing, developing socialization and emotional regulation [5, 6]. This approach dovetails with recommendations in the AI4K12 "Big Ideas" framework [1] that young children in grades K-2 (approximately ages 5-8) should investigate AI models that recognize expressions and inferred emotional states in human faces to make sense of how AI agents interact with humans (referred to as "Big Idea 4: Natural interaction" in the AI4K12 framework). Recent studies have specifically explored the promise of embodied learning as a strategy to support AI learning among young children [2, 4, 26]. Particularly salient for the present study, Dai and colleagues [4] present an experimental study of a three-phase strategy (including embodied, analogical, and disruptive phases) to support learning that includes embodied learning as well as an explicit disruptive strategy to critically compare AI reasoning and human intelligence for similar tasks. While the authors offer evidence for both metacognitive inquiry (seen in the "analogical" phase of the learning strategy) and embodied approaches to promote learning about AI systems, we see two important distinctions from the current study. First, Dai and colleagues' study [4] is focused on somewhat older youth (upper elementary) in a formal, classroom setting, rather than a minimally facilitated, free-choice environment with early (and pre-) elementary-aged youth. Second, rather than Dai's emphasis on analogies between the kinds of tasks humans and AI agents are able to perform, our work focuses specifically on the use of metacognition as a conceptual scaffold to facilitate youth construction of a mechanistic, explanatory account of *how* an AI performs a particular task. That is to say, we are positioning metacognition as a tool to support youth in constructing a working mental model of how an AI might be accomplishing "intelligence."

3 METHODS

3.1 A/B Setup

For this study, we wanted to investigate how engagement with the metacognitive task (Activity A: Unplugged Face Expression Card Activity, described above) would interact with sensemaking about the AI Facial Expression Recognition Activity (Activity B). Because we hypothesized that providing youth with a metacognitive task would position them for sensemaking related to an AI performing a similar task, we used an A/B design for the study and randomly varied the order in which participants would engage with the two tasks. Thus, the A/B design was not used to compare the order of experiences, per se, but to create a baseline condition (B only) with which to compare the McE condition (A+B). Thus, the present study offers an existence proof of McE: can we promote McE in ideal conditions, and if so, does it contribute to learning about AI?

For our study, we invited all visitors into the Learning Lab where the research was being conducted. If a group of visitors consisted of one or more children (ages 4-12) and if a research station was available, we would invite them to participate in the study. This invitation would include language about the research process (described above) and provide a high level overview of the tasks the participants would be asked to engage in. In total, 19 children participated (16 of those aged 7-9; one 6 year old and two 10-12 year olds).

3.2 Activity A: Unplugged Face Expression Card Activity

The purpose of the unplugged activity was to support youth to think metacognitively about how they—as humans—recognize different facial expressions and to promote curiosity and ideation about how an AI system would need to work in order to accomplish the same task. We designed this activity to serve as a non-digital ("unplugged") metacognitive activity for visitors to think about how they as humans make sense of the facial expression of others. In this activity, visitors are provided with a number of laminated photos of children making different facial expressions (face expression cards) and dry erase markers that enable the visitors to markup and/or annotate the cards. The photos used in this activity were selected from the Child Affective Facial Expression (CAFE) dataset, a collection of photographs of children from diverse ethnicities posing six different emotional facial expressions — happiness, sadness, anger, disgust, fear, and surprise [12]. In initial testing of this activity, we found that without instructions, visitors were not well positioned to engage with the task as designed, but with simple instructions from a museum facilitator, visitors were quick to understand the task of drawing (or often, circling) the parts of the faces that most communicate to visitors what type of facial expression the children on the cards are making.

3.3 Activity B: AI Facial Expression Recognition Activity

The purpose of this AI activity was to present a relatively straightforward implementation of AI that participants could, using their own experiences with recognizing facial expressions, engage in sensemaking about how the AI was working. We developed this learning experience (as a web application) to support young museum visitors in learning about and making sense of AI. In this experience, the interactant is presented with the screen shown in Figure 1. As depicted in the figure, the cropped video feed is shown twice, once on the left with no post-processing done to the image, and once on the right with lines superimposed on the image to highlight certain facial features relevant to the facial expression that the AI is recognizing. The bar chart on the far right of the screen shows the probability of each expression output by the AI expression model. The inclusion of emojis in this graphic has been helpful not only for younger visitors who are pre-literate, but also for older visitors who sometimes use the emojis as examples of how they should change their expression to get the model to predict that they are making said facial expression. An affordance we discovered in prior testing is that young visitors truly enjoy that they are engaging with a video of themselves. Or, put differently, that the AI system is making a prediction about *them*. This personal element has been critical to make this experience inviting and engaging for visitors.

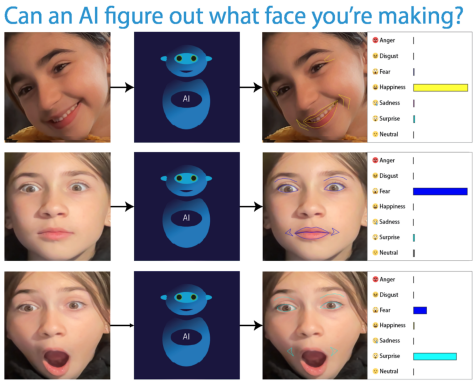

Figure 1: AI Facial Expression Recognition Activity

3.4 Data Analysis

Researchers reviewed the full set of data and identified the recordings from the AI activity (Activity B) as the primary data set to test our hypotheses about McE and to explore how youth make sense of the underlying AI. Focusing on these n=19 interviews, we then used the software to segment the mixed media file into excerpts that could be coded and analyzed for variation. Researchers then reviewed the 19 excerpts as a set and came to consensus around four codes that could be used to categorize the type of response given. The codes were designed such that only one code is assigned to each excerpt; that is, the codes were mutually exclusive. The four codes were: No response (the participant did not respond to the question, even after multiple prompts); Holistic response (the participant focuses on the face as a whole); Individual Features (the participant identifies an individual facial feature, such as the mouth, in describing how the AI recognizes expression); and Feature Set (the participant identifies multiple facial features).

4 FINDINGS

4.1 Evidence for metacognitive embodiment

4.1.1 Metacognitive priming supports youth sense-making around AI.. The A/B setup of this study enabled us to directly compare sensemaking about the AI system between those who engaged in a metacognitive task prior to encountering the AI, versus those who encountered the AI without first engaging in the metacognitive task. Our hypothesis was that if metacognitive inspection of one's own performance is supportive to sensemaking about AI, then those who have engaged in metacognition prior to encountering the AI system will be better positioned to make sense of it. As summarized in the figure 3, our analysis suggests that this was indeed the case: where youth were provided opportunities to engage metacognitively with their own interpretation of facial expressions, they were more successful in sensemaking related to an AI performing a similar task compared to youth who did not have that experience.

Specifically, we found that 62% (8 of 13) of the children who first engaged with the metacognitive card task offered an explanation of the AI system that included attention to multiple facial features (e.g., eyes and mouth), compared with only 33%, (2 of 6) children who did not have that explicit prior experience. In contrast, half of the participants (3 of 6) who were asked to make sense of the AI without first engaging in the metacognitive task provided either no response after multiple prompts (2 of 6), or offered an explanation in more holistic terms (1 of 6), such as "it can tell by my face" or "it sees my expressions." In contrast, all 13 participants in the "Cards → AI" group provided at least some response to the prompt for an explanation of the AI, and 85% (11/13) of the "cards first" group offered some evidence of decomposition (i.e., their responses were coded as either Feature Set or Individual Feature), whereas only half of the "AI first" group met this criterion. Taken together, this suggests that McE, which promotes youth reflection on their own performance of a task in coordination with inquiry into that of an AI, is a promising way to support young children's sense-making of AI.

4.2 Insight into how metacognitive embodiment operates

To better understand how McE operates in young children's sensemaking of AI, we present cases that illustrate how children leveraged the metacognitive task to step inside the "mind" of an AI. As will be demonstrated in the cases below, we see evidence that: 1) McE offers a productive working model of intelligent performance that youth can leverage to make sense of AI performance; and 2) McE provides opportunities to practice the analytic processes of decomposition and abstraction in an accessible context.

4.2.1 Theme 1: Embodying the AI's performance with a working model of intelligence. Participants who began with the laminated cards task regularly brought their mental model of how they distinguished facial expressions on the cards as a starting place for

their discussion of the AI system performing a similar task. We saw evidence for this in the parallel language many participants used between tasks, and the overt references to similar specific features (mouth, eyes, etc) that they identified between the two. For example, children who zeroed in on teeth and eyes during the card task tended to focus (at least initially) on teeth and eyes in making sense of the AI. We also saw evidence for this in participants' own reflections when asked how the experience with the cards informed their thinking about the AI. Most participants referenced the similarities between the markups they made on the laminated cards and the highlighting feature of the AI task interface. One child from the "Cards → AI" group extended this to perfectly encapsulate the strategy of McE, "when I was drawing [on the faces on the cards] it's kind of like not the same thing, but it's like on a picture, it's like I'm the AI doing it."

4.2.2 Theme 2: Accessible opportunities to practice decomposition and abstraction. In order to successfully complete the laminated cards task, participants needed to decompose the face into salient features (e.g., mouth size, teeth, eyebrow angle) and abstract patterns in those features across expressions (e.g., a "surprised" expression is composed of an open mouth and raised eyebrows). What we noticed for the overwhelming majority (85%) of participants who began with the card activity, is that they were able to apply these analytic processes to their sense-making of the AI, as in the example below.

> **Researcher**: How do you think the AI knew your expression?
> **Participant 11**: [For surprise] The mouth is really big, [child uses her hands to stretch her mouth], And not like angry, with like...[child bears teeth in an exaggerated manner]. And then with sadness, I'm trying to put my mouth down [as child uses her fingers to pull down the corners of her mouth].

In this case, the child has decomposed the complexity of facial expressions into a few salient features: mouth size, visible teeth, and the relative position of the corners of the mouth, all of which had been previously identified by the child during the card activity. Further, the child appears to be abstracting patterns in the position and arrangement of these features (e.g., connecting visible teeth to anger). Moreover, in presenting exaggerated versions of these features, the child is likely recognizing how these features, once abstracted, are adequate to signal particular expressions, that the complexity of an expression can be reduced to a limited set of identifiable patterns ("with sadness, I'm trying to put my mouth down").

5 DISCUSSION

Findings reported above suggest that providing explicit opportunities for youth to metacognitively consider how they perform a cognitive task positions them to consider how an AI system might perform a similar task. This McE strategy, in which youth reflect on their own performance of a task in conjunction with inquiry into that of an AI, offers a promising approach to support young children's early sense-making about artificial intelligence. While the working theories children discussed represent considerable simplifications of the underlying AI model, their efforts suggest that the child is beginning to decompose the problem space of a facial expression into the computational components that an automated system can be trained on to recognize, and to abstract patterns in the relationship between those components and particular expressions.

Our study further suggests a theory of action for this strategy that centers on two key mechanisms. First, by eliciting youth perspectives on how a human intelligent agent performs the task, youth are bolstered with a working model that they can then apply to considering how an artificial intelligent agent performs the same task. In this sense, the metacognitive task serves to elicit relevant ideas and help youth leverage their own experiences to construct a productive starting place from which to investigate an AI system. Second, by providing opportunity for the systematic analysis of (their own) intelligent performance on a task that they are intimately familiar with, the metacognitive process provides practice with the analytic work of interrogating how an artificial agent might perform a similar task. In particular, we saw evidence this strategy supported youth in decomposing the problem space on which an AI system operates, and abstracting patterns that relate particular features to particular expressions. These analytic processes are cornerstones of computational thinking and problem solving more generally, but have also proven challenging for youth of this age [16]. In short, McE provides both a generative working model for youth and practice with the analytic work of how to use that model to make sense of an AI system. For the particular case of McE applied to youth understanding of AI expression detection, we found this strategy to be highly engaging for youth: put simply, children like to think about themselves, and (literally) see themselves in the activity. As such, we see tremendous promise for this strategy as something accessible to young children, for whom this kind of metacognition is developmentally appropriate, and nearly universal in its relevance. Put another way, McE grounds youth sensemaking about AI in something deeply human, and that thereby enables youth to leverage those experiences as assets in understanding AI.

While we see promise for McE as a strategy, we also recognize potential challenges that require additional research. Here, we are particularly concerned with the possibility of over-literal interpretations of AI as isomorphic with human cognition. For example, it's not uncommon for young children (grades 1-3) to imbue AI systems with human emotions [9, 10]. Science educators, for example, have long been concerned about anthropomorphism as a threat to more accurate understanding of scientific phenomena like animal behavior [25] and natural selection [20], yet also recognize its promise when used thoughtfully [24]. Therefore, research is needed to understand how these human-derived working models of artificial intelligence are developed and revised as children mature and engage in increasingly advanced and technical AI learning experiences. Specific questions include: To what extent is the strategy productive beyond initial sense-making experiences? How does McE operate when youth enter into a learning experience with a more developed model of AI? To what extent does McE contribute to youth development of increasingly accurate understanding of AI systems over the long term? Such research could also contribute to articulating an empirically grounded learning progression for how youth develop understanding of AI models.

ACKNOWLEDGMENTS

This research was supported by a grant from the National Science Foundation (DRL 2116109). Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] AI4K12. 2022. Big idea 4 – natural interaction. https://ai4k12.org/big-idea-4-natural-interaction/

[2] Safinah Ali, Daniella DiPaola, Irene Lee, Jenna Hong, and Cynthia Breazeal. 2021. Exploring generative models with middle school students. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[3] Sue Allen, Patricia B Campbell, Lynn D Dierking, Barbara N Flagg, Alan J Friedman, Cecilia Garibay, and David A Ucko. 2008. Framework for evaluating impacts of informal science education projects. In *Report from a National Science Foundation Workshop. The National Science Foundation, Division of Research on Learning in Formal and Informal Settings*.

[4] Yun Dai, Ziyan Lin, Ang Liu, and Wenlan Wang. 2024. An embodied, analogical and disruptive approach of AI pedagogy in upper elementary education: An experimental study. *British Journal of Educational Technology* 55, 1 (2024), 417–434.

[5] Susanne A Denham. 2006. Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early education and development* 17, 1 (2006), 57–89.

[6] Susanne A Denham. 2007. Dealing with feelings: how children negotiate the worlds of emotions and social relationships. *Cognitie, Creier, Comportament/Cognition, Brain, Behavior* 11, 1 (2007).

[7] Paul Dourish. 2004. *Where the action is: the foundations of embodied interaction*. MIT press.

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

[9] Eric Greenwald, Maxyn Leitner, and Ning Wang. 2021. Learning Artificial Intelligence: Insights into How Youth Encounter and Build Understanding of AI Concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15526–15533.

[10] Clint Heinze, Janet Haase, and Helen Higgins. 2010. An Action Research Report from a Multi-Year Approach to Teaching Artificial Intelligence at the K-6 Level. *Proceedings of the National Conference on Artificial Intelligence* 3 (07 2010), 1890–1895. https://doi.org/10.1609/aaai.v24i3.18830

[11] McKinsey Global Institute. 2017. A future that works: automation, employment, and productivity. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx

[12] Vanessa LoBue and Cat Thrasher. 2015. The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in psychology* 5 (2015), 1532.

[13] 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.

[14] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. 2021. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Creativity and Cognition*. 1–10.

[15] Lorraine McLeod. 1997. Young children and metacognition: Do we know what they know they know? And if so, what do we do about it? *Australasian Journal of Early Childhood* 22, 2 (1997), 6–11.

[16] Wouter J Rijke, Lars Bollen, Tessa HS Eysink, and Jos LJ Tolboom. 2018. Computational thinking in primary school: An examination of abstraction and decomposition in different age groups. *Informatics in education* 17, 1 (2018), 77–92.

[17] Gregory Schraw and David Moshman. 1995. Metacognitive theories. *Educational psychology review* 7, 4 (1995), 351–371.

[18] Pratim Sengupta, Amanda Dickes, and Amy Farris. 2018. Toward a phenomenology of computational thinking in STEM education. *Computational thinking in the STEM disciplines* (2018), 49–72.

[19] Karen Sullenger. 2006. Beyond School Walls: Informal Education and the Culture of Science. *Education Canada* 46, 3 (2006), 15–18.

[20] Lucy G Sullivan. 1995. Myth, metaphor and hypothesis: how anthropomorphism defeats science. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 349, 1328 (1995), 215–218.

[21] 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.

[22] Marcel VJ Veenman, Bernadette HAM Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and learning* 1 (2006), 3–14.

[23] David Whitebread, Penny Coltman, Deborah Pino Pasternak, Claire Sangster, Valeska Grau, Sue Bingham, Qais Almeqdad, and Demetra Demetriou. 2009. The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and learning* 4, 1 (2009), 63–85.

[24] Lisa A Williams, Sarah F Brosnan, and Zanna Clay. 2020. Anthropomorphism in comparative affective science: Advocating a mindful approach. *Neuroscience & Biobehavioral Reviews* 115 (2020), 299–307.

[25] Clive DL Wynne. 2007. What are animals? Why anthropomorphism is still not a scientific approach to behavior. *Comparative Cognition & Behavior Reviews* 2 (2007).

[26] Weipeng Yang. 2022. Artificial Intelligence education for young children: Why, what, and how in curriculum design and implementation. *Computers and Education: Artificial Intelligence* 3 (2022), 100061.

[27] 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.