Family Learning Talk in AI Literacy Learning Activities

Duri Long
Georgia Institute of Technology
Atlanta, GA, USA
duri@gatech.edu

Anthony Teachey
Georgia Institute of Technology
Atlanta, GA, USA
ateachey3@gatech.edu

Brian Magerko
Georgia Institute of Technology
Atlanta, GA, USA
magerko@gatech.edu

ABSTRACT

The unique role that AI plays in making decisions that affect humans creates a need for public understanding of AI. Informal learning spaces are important contexts for fostering AI literacy, as they can reach a broader audience and provide spaces for children and parents to learn together. This paper explores 1) what types of dialogue familes engage in when learning about AI in an at-home learning environment to inform our understanding of 2) how to design AI literacy activities for informal learning contexts. We present an analysis of family dialogue surrounding three AI education activities and use our findings to update existing principles for designing AI literacy educational interventions. Our findings indicate that embodied interaction, collaboration, and lowering barriers to entry were effective at fostering learning talk. Our results also reveal emergent areas for future research on how to support parents and design visualizations and datasets for AI learning.

CCS CONCEPTS

• **Social and professional topics** → **Informal education**; • **Human-centered computing** → *Empirical studies in interaction design*; • **Computing methodologies** → *Artificial intelligence*.

KEYWORDS

AI literacy, collaborative, embodied, learning talk, informal learning, family learning, AI education

ACM Reference Format:
Duri Long, Anthony Teachey, and Brian Magerko. 2022. Family Learning Talk in AI Literacy Learning Activities. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3491102.3502091

1 INTRODUCTION

Artificial intelligence (AI) technologies are currently making decisions for and with humans in a variety of consequential contexts, including recidivism [56], hiring [90], newsfeed and search result curation [21], entertainment and shopping recommendations [71], and military and law enforcement [64, 77]. The integration of AI into our day-to-day decision-making processes will only increase as new AI technologies are developed for use in our homes, cars, governments, social lives, and workplaces. We have already seen how serious the consequences of misunderstanding or failing to question/regulate AI's decisions can be–leading to issues like viral misinformation [5, 62], biased systems that disproportionately impact marginalized communities [9, 12, 14, 60, 88], and serious concerns about data privacy [73].

The unique role that AI plays in making decisions that affect human lives creates a need to foster better public understanding of AI systems. We assert that it is critical for people interacting with and using AI tools to have "AI literacy"–or 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 [51]. Fostering public AI literacy could lead to more productive human-AI communication, better human-in-the-loop AI systems that can take advantage of AI's strengths while mitigating its weaknesses, skills that enable people to better advocate for themselves on AI-related issues, and more informed public debate about the role AI should play in our society.

Informal learning spaces are particularly important contexts for fostering AI literacy. Research has shown that much of science and technology learning and interest development happens outside of formal classroom settings [23], and interventions in informal spaces can often reach a broader audience of learners than would self-select to attend a formal course [70]. Despite this, most existing research on AI education for individuals without computer/data science backgrounds has focused on K-12 classroom environments.

Informal learning contexts have the additional benefit of providing spaces for family groups to learn together. Children and adults both are making sense of AI and its role in their lives, and there are many open questions about the personal, societal, and ethical implications of AI technologies. Collaborative dialogue and multi-generational perspectives are important in helping families to make sense of these issues. In addition, providing adults with opportunities to learn about AI alongside their children is important, as AI education for adults without computing backgrounds is an underexplored area.

This paper explores how to foster learning about AI with family groups in informal learning environments. We investigate our first research question– *What types of dialogue do family groups engage in when making sense of and learning about AI?*–in order to inform our understanding of our second research question–*How can we design activities to facilitate family group learning about AI literacy competencies in informal learning environments?*. We created three different activities intended to foster learning about AI, informed by a set of existing design considerations for AI literacy learning interventions [51]. We conducted an analysis of family group dialogue surrounding the three activities in order to determine whether and

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502091

how they fostered group learning, and we reflect on how our findings contribute to a more refined understanding of the AI literacy design considerations [51].

We found that certain design features–especially using tangible and full-body interfaces, supporting collaboration, and lowering barriers to entry by designing activities that require no prerequisite computing or programming knowledge–were effective at fostering family group engagement and dialogue relevant to AI literacy competencies. Our findings also reveal areas in need of further investigation, including how to scaffold novice engagement with explanatory visualizations, how to curate datasets used in AI learning activities, and how to best support parents and young learners who are engaging as part of a group with AI-related activities. In summary, this paper contributes a better understanding of how family groups engage in dialogue and learn about AI and uses these findings to reflect on, update, and add to existing principles for designing AI literacy educational interventions for informal learning environments.

2 RELATED WORK

2.1 AI Education

There has been a recent push in the CHI community and beyond to expand opportunities to learn about AI beyond expert audiences. In particular, there has been a significant effort to develop technologies and curricula to teach K-12 audiences about AI. Several groups are working to develop curricula and standards for incorporating AI in K-12 education [4, 28, 40, 43, 79, 80]. Researchers and designers have created a wide variety of new tools to support learning about AI, including plugins for popular coding platforms like Scratch and MIT App Inventor [2, 10, 17, 82], new tools for teaching specific AI/ML concepts through interactive activities (e.g. [69, 84, 94]), and "unplugged" or no-technology-needed paper-based AI education activities [45].

Much of the current research on how people learn about AI is synthesized in a CHI 2020 literature review paper written by the first and third author (Long and Magerko) that outlines a set of competencies and design considerations for learning interventions intended to foster AI literacy. In this 2020 review, we derive a set of competencies as well as design principles for designing AI literacy learning interventions based on a review of prior literature related to AI education, computing education, human-centered AI, and more. The competencies we present are high-level ideas that are intended to aid in public understanding of AI–such as recognizing when technologies use AI, understanding that computers learn from data, or understanding some strategies AI uses for decision-making. The design principles we present include suggestions such as designing learning experiences that foster collaboration or embodied interaction, providing support and scaffolding for parents, and unveiling system components gradually so as not to overwhelm learners. In this paper, we use our previously presented competencies and design principles as guidelines for our designs and we reflect on each of the design principles in the Discussion section (section 6).

Since the AI education field is in a stage of rapid exploration and development, more recent updates to our literature review and additional literature reviews on related topics have already been published [24, 47, 54, 59, 83, 92]. We refer the interested reader to these works for a more thorough review of recent work related to AI education. Several papers have expanded on the design considerations in [51] or suggested some additional design considerations for AI literacy activities. Zhou et al. suggest incorporating opportunities for learners to "learn by teaching," using gamification to teach about AI concepts, supporting iteration with immediate feedback, promoting reflection, providing opportunities for teachers and parents to learn alongside children, and integrating AI across disciplines in other K-12 course curricula [92]. Van Brummelen et al. suggest adding several AI-related concepts to Brennan and Resnick's computational thinking framework [11]–including classification, prediction, generation, training/validating/testing, and evaluation [81]. Touretzky and Gardner McCune suggest drawing on the growing number of browser-based tools (e.g. Cognimates, eCraft2Learn, TensorFlow, Teachable Machine) as well as incorporating paper-based activities like interacting with maps and decision trees or doing hand simulations of simplified ML algorithms [78].

Not a lot of existing work has focused on designing AI-related activities specifically for informal learning contexts. Several papers have explored how adults develop and revise "folk theories" about opaque algorithms online [21, 22] and how adults make sense of algorithms when provided with explanations [20]. A recent poster paper explores how to use a kiosk exhibit to teach about AI history in public libraries [85]. Some museums have curated AI-related installations–including artifacts, artwork, and interactive demonstrations–into overarching exhibits that explore more holistic representations of AI. The Barbican developed *AI: More than Human* [8], which presented the history of AI via artifacts, interactive timelines, artwork, and demonstrations. Ars Electronica also recently curated *Understanding AI* [25], which featured several installations that aimed to demystify machine learning algorithms. Visitors could interactively explore how ML technologies such as image recognition, unsupervised learning, and neural networks worked. Both of these exhibits have taken a step towards facilitating AI-related interactive learning experiences. However, there are still many aspects of AI that have yet to be explored in an informal learning context and there are few existing projects that explicitly draw on research or theory on AI education and museum exhibit design.

Our paper departs from a growing body of research in the field that explores how to teach AI/ML through programming platforms (e.g. [2, 10, 17, 81, 82]). Most of these projects have been targeted at integrating with K-12 classroom contexts or structured summer programs. Given our focus on informal learning contexts where groups contain visitors of all ages, individuals may have little or no prior knowledge of coding, and participants have limited time to learn, we focus on high-level AI-related competencies that can be quickly communicated to individuals who may not know how to program.

The activities presented in this paper build instead on prior work that suggests that embodied interaction can be a particularly effective way of concretizing abstract AI concepts for learners [17, 76] and engaging learners with diverse interests in learning about AI [39, 94]. We draw inspiration from "unplugged" CS and AI activities that are often well-suited for learners of all ages with a range of technical literacy, due to their hands-on, no-technology-needed

Table 1: Summary of competencies and design principles for each activity

Activity Name	Competencies	Design Principles
Knowledge Net	ways of representing knowledge, strengths and weaknesses of AI, role of humans in programming AI, how agents make decisions	embodied interaction, social interaction, explainability, opportunities to program or teach AI, leveraging learners' interests, facilitating a low barrier of entry
Creature Features	steps and practices of machine learning, ways of representing knowledge, how agents make decisions, computers learn from data	encourage learners to contextualize data, social interaction, opportunities to program or teach AI, explainable algorithms
LuminAI	ways of representing knowledge, how agents make decisions, computers learn from data, steps and practices of machine learning	embodied interaction, social interaction, explainability, opportunities to program or teach AI, incorporating learner interests, engaging with lesser-known forms of AI

Table 2: Summary of differences between different iterations of the system

Activity Name	Iteration 1	Iteration 2
Knowledge Net	laminated paper gameboard, wooden tiles, photo taken with smart phone, some issues with image recognition	wooden game board, paper tiles, photo taken with Osmo device and iPad to improve image recognition and usability
Creature Features	laminated paper gameboard, front/back card design, photo taken with smart phone, only a positive dataset	wooden gameboard, front-only card design, photo taken with Osmo device and iPad to improve image recognition, positive and negative datasets
LuminAI	-	some minor bugs in the system fixed, otherwise the same as Iteration 1

nature [45, 78, 89]. Our work also draws on the body of research that explores tangible interfaces as an effective way of engaging families in learning together about computer science and other topics [31–33].

2.2 Assessing Learning at AI Education Activities

Numerous groups have developed AI-related learning interventions and assessed them using a variety of means, including collecting students' daily reflections and conducting interviews with students [41], asking learners to take pre/post tests or complete questionnaires [3, 80] or asking teachers for their impressions of student learning [3]. Register and Ko assessed learning about AI concepts by asking learners to write self-advocacy letters in response to AI-related scenarios [65]. Others have conducted studies to assess learner preconceptions of AI [35] and worked with teachers to understand pedagogical content knowledge necessary for teaching machine learning to non-computing majors [76]. All of these studies have led to insights into how to best design learning interventions to support learning about AI literacy competencies.

However, traditional assessments can be challenging to use in informal learning contexts like museums and at-home learning, due to limits on visitor time and low learner motivation to take a "test" during a leisure activity [33]. Observational assessment techniques such as analysis of "learning talk"–or group dialogue that relates to the learning goals of the exhibit or activity–is often a more practical and insightful way of assessing learning in informal spaces [6, 67]. Insight into how learners engage in dialogue surrounding AI literacy activities can also further deepen our understanding of social learning about AI.

There is not a significant amount of existing work analyzing participant learning talk surrounding AI education activities for non-expert audiences. The most relevant studies are papers that examine the types of questions and conversations that young learners and families ask when interacting with AI voice assistants [17, 18, 63]. Our prior work has also examined family conversation in a co-design workshop related to AI education [48]. Our paper adds to this body of work by assessing the activities we develop using a learning talk analysis and in doing so contributing a study of family group learning talk surrounding AI literacy learning activities.

2.3 Family Learning and Engagement in Museums

Research on family learning in informal learning spaces is also relevant to our work. Family group and peer collaboration both encourage learning and can facilitate constructive dialogue, parental and peer scaffolding, and sharing of perspectives [16, 87]. Collaboration is particularly important to design for in museums since visitors most often come in groups [30].

Prior work has explored how family groups engage with tangible and embodied interfaces in museums [32]. Embodied experiences in museums can aid in understanding concepts like scale and size [57], encourage learners to empathize with others via the exertion of physical effort [53, 72], and help to connect abstract concepts with children's existing bodily experiences [1, 13, 44, 76]. Tangible interfaces are able to evoke *cultural forms*, or recognized conventions and social patterns of activity [31]. For example, a tangible interface for computer science education that is puzzle-like will evoke social dynamics and interaction patterns that families are already familiar with from completing puzzles together. We build off of this prior

work to create interfaces that feel familiar to families, even as they broach unfamiliar topics.

Other work has studied family dialogue to explore how to foster group inquiry [29], curiosity [66], and sense-making practices [93] surrounding museum exhibits. Aspects of our activity design and the design of the associated instructional materials build on this prior work to foster group inquiry and discussion.

3 ACTIVITIES

We analyzed family group dialogue with three different activities designed to foster learning about AI. These activities were originally prototyped as exhibits for installation in a science museum, and were later adapted for use in at-home learning environments due to the COVID-19 pandemic. The three activities focus on communicating a variety of AI literacy competencies, ranging from the steps and practices of machine learning to understanding knowledge representations. We drew on an array of design principles when creating the activities, but especially focused on incorporating opportunities for tangible or full-body interaction and group collaboration (see Section 2.3).

This section provides a brief description of each activity we developed and the AI literacy competencies and design considerations that they incorporate (Table 1 summarizes the competencies/design principles for each activity). The core contribution of this paper is the learning talk analysis and reflection on AI literacy design principles, so we do not go into detail on the design process here, but we refer the interested reader to [52] for more information.

Each of these exhibits was developed iteratively. We conducted user studies with family groups in two different sessions–the families participating in the first session of studies interacted with an earlier prototype of the exhibits than the families participating in the second study session. The overarching exhibit design is summarized below, and Table 2 outlines the differences between the two iterations of the exhibit design to contextualize the learning talk analysis.

3.1 Knowledge Net

Knowledge Net (Figure 1) is an activity in which learners can use a tangible interface of wooden tiles and arrows to collaboratively build semantic networks (a type of AI knowledge representation that contains concepts and relationships between them) about topics of interest to them (e.g. family, animals, music, etc.). Once learners build their network, they can take a photo of it, upload it to a website, and ask questions to an AI chatbot that uses their network as its knowledge base (e.g. Learner: What is a cat?, Computer: A cat is a mammal). This prototype aims to communicate AI competencies such as ways of representing knowledge, how agents learn and make decisions, understanding strengths/weaknesses of AI, and recognizing the role that humans play in programming and teaching AI [51]. It incorporates AI literacy design principles such as embodied interaction, collaboration, explainable algorithms, opportunities for individuals to program or teach AI (in this case, the focus was not on programming but on 'teaching' the AI by creating a network of information), and facilitating a low barrier of entry [51]. We encouraged learners engage in a discussion after their interaction about the types of relationships they were able to capture, what would happen if they put false information into the network, and whether the computer understands concepts in the same way a human does.

We had a number of technology related challenges with the *Knowledge Net* activity, particularly in the first round of studies. Learners were supposed to photograph the playmat using their phones and upload the information to the AI chatbot, but often the participants took photos of the playmat in poor lighting conditions or with part of the board obscured, which caused issues with the image recognition algorithm. These issues were mitigated in the second round of studies by our use of an Osmo device [1], which had a fixed location. Due to the challenges with image recognition, not all groups in the first session were able to engage with the chatbot. Groups that were unable to engage with the chatbot were prompted to engage in a "unplugged" simulation/role-play exercise, where one group member acted as the computer and used the network to answer questions that other group members asked.

3.2 Creature Features

Creature Features (Figure 2) is an activity in which learners can use a card deck and "weight tokens" to build a training dataset for a feature-based machine learning algorithm that classifies birds. Each card depicts a creature (e.g. bluebird, bat) and includes a list of descriptive features (e.g. color, habitat, size). Learners are encouraged to look at the features and consider how to place their weights to create an algorithm that can correctly recognize many different types of birds. The more weight tokens that are placed on a card, the more examples of that creature are going to be added to the training dataset for the algorithm. In the second iteration of the design, learners were asked to build both positive (examples of birds) and negative (examples of non-birds) datasets. Learners can take a picture of their playmat and upload it to a website, which will tell them how well their algorithm classifies birds. This prototype aims to communicate AI-related competencies such as the steps and practices of machine learning, ways of representing knowledge, how agents make decisions, and data curation and interpretation [51]. It incorporates AI literacy design principles such as embodied interaction and metaphors, collaborative discussion, providing opportunities to program or teach AI, and creating explainable algorithms [51]. We had learners engage in discussion after their interaction about what birds were hard to get the AI to recognize, why certain creatures were misclassified, whether they were surprised by the role humans play in programming the AI, and whether they could foresee any issues with using this technology to recognize other things, like faces or objects.

3.3 LuminAI

LuminAI (Figures 3, 4) is an activity in which learners can improvise movement together with an AI dance partner that is projected onto a screen. *LuminAI* was an existing AI research project [37, 49] that we expanded into a learning experience. In the expanded version of *LuminAI*, learners can engage with an interactive visual interface to explore different aspects of the dancer's decision-making processes and memory, such as manipulating the dancer's response modes (i.e. mimicry, transforming a gesture, performing a gesture

[1] https://www.playosmo.com

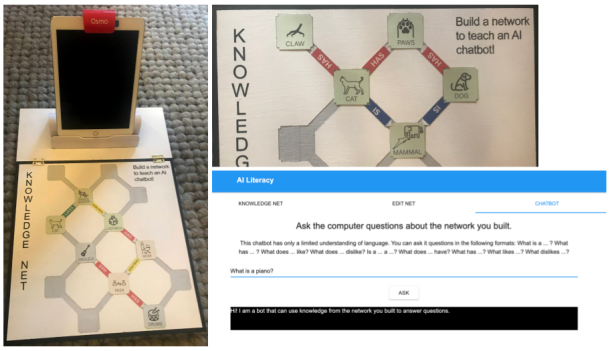

Figure 1: *Knowledge Net* Prototype

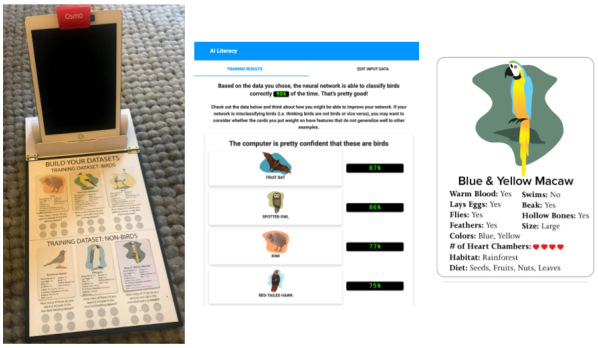

Figure 2: *Creature Features* Prototype

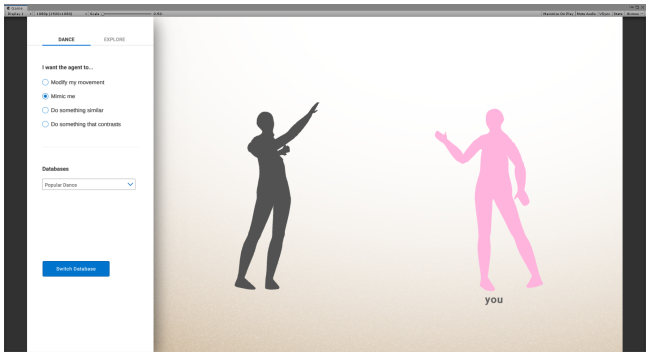

Figure 3: *LuminAI* dance interface

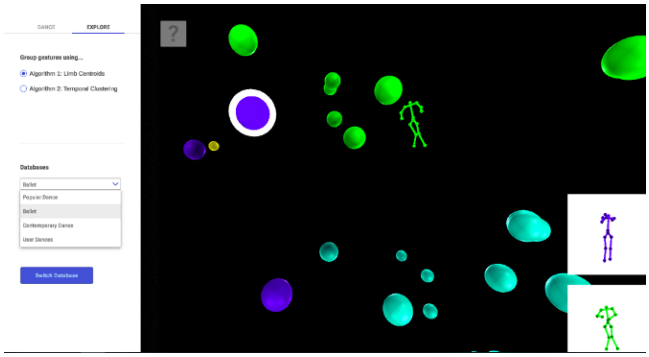

Figure 4: *MoViz* interface

from memory that is similar or contrasting to the observed gesture), switching between different databases of dance gestures (e.g. ballet, popular dance), and exploring a 3D visual representation of the way the dancer uses unsupervised machine learning to cluster gestures in memory (called *MoViz* [46]). This activity aims to communicate AI-related competencies such as ways of representing knowledge, how agents learn and make decisions, and aspects of machine learning [51]. It utilizes AI literacy design principles such as embodied interaction, varying degrees of collaboration depending on the installation setup [50], creating explainable algorithms providing opportunities to program or teach AI, incorporating learner interests, and engaging with lesser-known forms of AI [51]. After the interaction, we prompted learners to discuss whether they thought the AI was creative, whether the AI "thought" about dance in a different way than they did, how their actions affected the AI, and whether the agent clustered/grouped gestures well.

4 METHODS

We conducted a user study session for each iteration of the prototypes to better understand learner engagement with the activities. We had originally planned to test each of the prototypes by installing them as pop-up exhibits in a science and technology museum, but due to COVID-19 we had to pivot and instead conducted remote user studies with families who engaged with the modified activities from their homes. This section outlines the methods we used for participant recruitment, data collection, and analysis.

4.1 Participant Recruitment

We recruited family groups to participate using a variety of different methods, including posting on NextDoor (and having friends and colleagues share the post in their neighborhoods), social media, coordinating with our university's education outreach program and reaching out to local organizations including the public library, the YMCA, Boys and Girls Clubs, and local Girls Who Code (GWC) club leaders. We received most of our responses from NextDoor, Girls Who Code, and "parents groups" on social media.

Interested study participants contacted us with the number/age of family members that would be participating. We followed up with an email that asked them about several requirements for study completion (i.e. access to a stable internet connection, access to a 4x4' open space to dance for *LuminAI*, access to a smartphone that could upload a photo to a website) and asked them for scheduling details regarding a good time for delivery/pickup and study completion. Activities were designed to fit in labeled boxes that we delivered to participants at a scheduled time. One box contained data collection materials and a written instruction packet; the other two boxes each contained a prototype. All boxes and activity components were sanitized according to our university's Environmental Health and Safety standards between user groups to prevent the spread of COVID-19.

The enclosed written instructions were intended to provide approximately the same amount of data as a participant would find when walking up to an exhibit in a museum. We provided detailed instructions on how to set up and use the exhibit (something that visitors might gather visually from observing others in a museum environment), but kept explanatory content-related text to a minimum (i.e. the amount that you might find on a sign next to the exhibit). The instructional packets provided with the activities are included in the supplemental materials.

4.2 Consent and Study Procedure

We asked adults participating in the study to complete a consent form online prior to delivering the boxes. This was done to ensure that the parents had a clear idea of what was involved in the study prior to going through the process of delivery, and also to simplify the in-home consent process. Children were asked for verbal assent during the study (see next paragraph). All participants were given a copy of their signed consent forms to keep. Family groups were compensated with $40 (either in cash or an Amazon gift card) for their time.

We dropped off the boxes on the participants' doorstep at the scheduled time, then called via either video or phone call (participant preference) at the designated study time. We briefly explained the study, pointing out key materials, and asked participating children for assent (verbal for children under 11; written for children 11 and up). Participants were then given the option to have the researcher stay on the call to answer questions or for the researcher to hang up and be readily available should the participants need to ask a question. We gave participants the option of having a researcher present during the activity since this is a stressful time for families everywhere with regards to jobs, childcare, and general health and safety concerns. We did not want our research to add to these burdens, so we tried to make the activities fun and engaging, rather than feeling like an additional task or Zoom meeting to complete. For calls that the researcher stayed on, they took on the role of an observer, watching quietly and only answering questions when asked so as not to unduly influence the interaction.

4.3 Data Collection

We collected audio and video recordings of participant interactions as well as survey data from all participants ages seven and up. The data from the surveys is not discussed in this paper (except for a few notes on participant feedback and demographic data), and we refer the interested reader to [52] for this information.

Since we were not present in the participant's homes, we asked participants to record their own data. We drew some inspiration from cultural probes [27] at this stage, since cultural probe kits typically include tools/instruments to aid participants in collecting their own data. We included an audio and video recorder in the kit along with detailed instructions on how to record the data. We anticipated potentially running into some technical difficulties when asking participants to record their own data, so we sought to introduce redundancy. We had participants record their interactions using both a video and an audio recorder, so that if one failed, we would at least have audio of the group's interactions from the other device. We offered groups the option of participating in a video call during the study, and for groups that agreed, we additionally recorded the video call using the video conferencing software.

5 ANALYSIS

We conducted a learning talk analysis of participant dialogue with the three activities. Conversation analysis is used in many contexts to understand the ways in which learners engage with content knowledge (e.g. [6, 55]). Understanding learning talk is particularly important when designing informal learning experiences like museum exhibits because participant dialogue is the "most reliable and accurate" indicator of learning at museum exhibits [68]. Even though we ended up testing the activities out in an at-home environment due to COVID-19, understanding learning talk contributes both to the ability to ultimately introduce these activities in a museum down the road and to a broader understanding of how novice learners engage with AI-related content knowledge.

We used Roberts and Lyons's framework for analyzing learning talk at museum exhibits to guide our analysis, since this is a framework that was designed for understanding family group learning at museum exhibits involving embodied interaction. We chose to use Roberts and Lyons's framework for several reasons. First, we wanted our results to be able to easily transfer to a museum setting, as our intent is to scale up the activities for museums down the road. Second, our research questions were focused on assessing whether and how the activities supported learning and discussion of AI literacy competencies in order to inform understanding of how to design AI literacy activities. Roberts and Lyons's framework afforded this focus on learning talk surrounding specific competencies. Finally, it is a quantifiable framework that affords comparisons across exhibits and participant groups. Roberts and Lyons set their framework apart from other quantifiable conversation analysis methods in the computer-supported collaborative learning community by highlighting that their framework supports the ability to characterize intersubjective meaning-making in a way that other frameworks do not [68]. They define intersubjective talk as "learners echoing and reiterating ideas as well as introducing ideas of their own, which depending on the coding categories and counting procedures can either over- or underrepresent the learning evidenced by the talk" [68].

Roberts and Lyons define five different types of learning talk that occurs at museum exhibits–*management*, *instantiations*, *evaluations*, *integrations*, and *generations* [68]. *Management* is "talk related to the establishment of joint attention, negotiation of action, or scaffolding exhibit use" including "explaining, asking and answering questions, and suggesting actions"; *instantiations* "indicate when a user says aloud a piece of information, providing opportunities for other visitors to internalize that information (i.e. learn from the exhibit)"; *evaluations* "make a judgment or assessment about a piece of information by assigning some kind of value, whether qualitative or quantitative"; *integrating* is "the act of pulling together multiple pieces of information presented in an exhibit...mak[ing] explicit connections or comparisons between multiple pieces of information"; and *generate* statements "combine information from the exhibit with visitors' own prior knowledge and experiences" [68]. We focused on *instantiate*, *evaluate*, *integrate*, and *generate* statements in our analysis and did not include *management* codes

at this time as they were less relevant to the learning outcomes for each exhibit and we anticipated they would be the most likely to change in a museum vs. at-home setting.

Audio recordings of participant interactions with the exhibits were transcribed for analysis by an external transcription service (Rev.com). Following the procedure outlined by Roberts and Lyons, we broke down transcripts of the audio recordings into *idea units*, where an idea unit is "marked by a distinct shift in focus or change in topic or purpose...Idea units can range in length from a single word, e.g., reading aloud a category name, to a multi sentence utterance" [68]. For example, the following chunk of dialogue would constitute a single idea unit:

> P1: Which of these on here do you really strongly feel is a bird?
> P2: Definitely hummingbird.
> P1: Okay.
> P2: Because that literally has the name, "Bird" in its name.

We then developed a set of codes for each exhibit to assess *instantiate*, *evaluate*, *integrate* and *generate* learning talk. Codes were developed using a combination of a top-down and bottom-up approach. There were certain AI literacy competencies we had designed for that we wanted to make sure to code (top-down), but we also familiarized ourselves with the data by listening through interaction sessions and reading transcripts to identify emergent codes (bottom-up). Each code was given a score on a scale of 1-3, where 3 indicated the most relevance to the learning goals of the exhibit (i.e. the AI literacy competencies it aimed to communicate, see Table 1). Multiple codes could be applied to a single idea unit. Aligning the codes with AI literacy competencies means that a higher learning talk score indicates increased discussion of relevant competencies.

A subset of the data (at least one transcript per activity) was coded by two analysts who then compared their results. Conflicts were resolved via discussion until the analysts came to a mutual agreement, and some code definitions were iteratively revised during this calibration period. The final codes we defined for each activity are summarised in Table 3, including the learning talk score we assigned to each code. A full coding scheme is also provided in the supplemental materials. All additional transcripts were coded by one of the two analysts who engaged in the initial calibration process. All transcripts were broken down into idea units and analyzed while also listening to the dialogue in order to ensure that contextual information such as pacing and inflection were taken into consideration [68].

5.1 Results

We recruited a total of 14 family groups (38 participants; 21 age 6-17 and 17 age 18+) to interact with the exhibit prototypes. One group (G2.4) was excluded from analysis due to a failure with audio and video data collection (see Table 5). Eight groups (22 participants) interacted with Iteration 1 prototypes, and six groups (16 participants) interacted with Iteration 2. Table 4 shows a breakdown of how many groups/participants interacted with each prototype. Among the 14 adults who answered the demographic questions, nine identified as White/Caucasian, four as African American, two as Asian American, and 1 as other Latin American (two participants were biracial). Most participants reported having a 4-year degree or graduate education (79%). Among the children, 10% were 6 years old, 30% were 7-9 years old, 50% were 10-14, and 10% were 15. 60% of children identified as female and 40% as male. We also asked families about their prior experience with computing and their children's prior experience with computing. Most adults considered their children to have "some" prior experience with computers (70%) and AI (60%). Most adults also reported that they worked with computers a lot or sometimes (79%) and had some prior experience interacting with AI technologies (79%) but did not write code (93%). Table 5 describes each group that participated in the study and summarizes the data that was collected for each group.

5.2 Overarching Analysis

With the data from the videos we coded for learning talk, we calculated 1) the *amount* of learning talk that occurred for each exhibit, on average; 2) the *type* of learning talk that occurred the most at each exhibit, on average (i.e. instantiate, evaluate, integrate, or generate); and 3) the *quality* of learning talk. This information is presented in Tables 6 and 7. Table 6 summarizes the total scores and median scores for 1) the number of total learning talk codes applied, 2) the learning talk score (calculated by multiplying the number of instances of each code by its assigned relevance score and adding all scores together), and 3) the number of codes scored at a level 1 (L1), level 2 (L2), and level 3 (L3), indicating the relevance of the code to the learning goals of the exhibit (3 being the highest). Table 7 summarizes the number of codes of each type that occurred at each exhibit. We also used the individual transcript-level learning talk data to examine "ceiling" and "floor" interactions for each prototype–that is, what did the most in-depth interactions look like vs. the least in-depth. We describe this in the next section alongside qualitative descriptions of learning talk with each exhibit.

We do not present results from statistical tests of this data since the sample size of the population was not large enough to produce statistically significant results. Study size was limited due to the time-consuming COVID-19 precautions and the challenging nature of recruiting study participants during a pandemic. We compare exhibits here using median scores (also reported in Tables 6 and 7) since *LuminAI* had fewer interaction sessions than the other two exhibits (6 rather than 10), meaning that the total scores are only useful for comparing *Knowledge Net* and *Creature Features*. It should be noted that different coding schemes were used to code each exhibit due to the differing content/interactions at each exhibit (see Table 3).

Knowledge Net generated the highest median number of codes across all sessions, followed by *LuminAI*, then *Creature Features*. This same ordering was true of the learning talk median scores. However, it is important to note that while *Creature Features* had lower median scores, it actually had almost exactly the same overall learning talk score as *Knowledge Net* despite having notably fewer codes for the same number of interaction sessions. This suggests that there were a few interaction sessions with very high quality learning talk for *Creature Features*, which we discuss more below.

Comparing the median scores for each learning talk type (i.e. instantiate, evaluate, integrate, or generate) indicates that all three

Knowledge Net		Creature Features		LuminAI	
K1: INSTANTIATE read instructions aloud	1	C1: INSTANTIATE read instructions aloud	1	L1: INSTANTIATE read instructions aloud	1
K2: INSTANTIATE tile or arrow name	1	C2: INSTANTIATE name of creature	1	L2: INSTANTIATE recognize dance move	1
K3: INSTANTIATE read chatbot response aloud	2	C3: INSTANTIATE features from card	2	L3: INSTANTIATE response modes	2
K4: EVALUATE strengths and weaknesses of network as a representation of knowledge	3	C4: INSTANTIATE read results of algorithm out loud	2	L4: INSTANTIATE databases	2
K5: EVALUATE chatbot answers	2	C5: EVALUATE whether creatures are "birdlike"	2	L5: EVALUATE quality of agent's moves	1
K6: EVALUATE quality of network structure	2	C6: EVALUATE consider where to place weights	2	L6: EVALUATE compare datasets/dance types	3
K7: INTEGRATE make connection between tiles	1	C7: INTEGRATE justify decision	3	L7: EVALUATE compare clusters	3
K8: INTEGRATE make connection between network and chatbot	3	C8: GENERATE features from prior knowledge	1	L8: EVALUATE compare clustering algorithms	3
K9: GENERATE compare/contrast network to human intelligence	3	C9: GENERATE make a plan for how to modify the dataset	3	L9: EVALUATE notice what is lost in representation	3
K10: GENERATE connect network to personal interests	1	C10: GENERATE make prediction about algorithm results	3	L10: INTEGRATE recognize that agent is responding to you	2
K11: GENERATE discuss nuances of network concepts and relationships, drawing on prior knowledge	2	C11: GENERATE notice surprising result	2	L11: GENERATE hypothesis	2
K12: GENERATE explanation for chatbot response	3	C12: GENERATE explanation of algorithm results	3	L12: GENERATE recognize hardware or software from prior experience	1
K13: GENERATE question to ask chatbot	2	C13: GENERATE connect activity to other application domains	3	L13: GENERATE teach agent dance move	2
K14: GENERATE make prediction about chatbot response	3			L14: GENERATE discuss whether agent is creative	3

Table 3: Summary of learning talk codes for each activity. Learning talk scores for each code are included in the table. Scores range from 1 to 3 with 3 being the most relevant to the learning goals of the activity. Full definitions for each code are provided in the supplemental materials.

Activity	Iteration 1		Iteration 2		Total	
	Groups	Participants	Groups	Participants	Groups	Participants
LuminAI	4	12	3	9	7	21
Creature Features	6	17	4	10	10	26
Knowledge Net	6	14	5	16	11	29

Table 4: Number of participants who interacted with each prototype

exhibits fostered roughly the same amount of instantiation. There was more evaluation dialogue at *Creature Features* and *LuminAI* than *Knowledge Net*. This may reflect that learners were more engaged in evaluating the AI's accuracy in *Creature Features* and *LuminAI* than in *Knowledge Net*, which makes sense because most learner groups did not engage in lengthy interactions with or discussions about the chatbot in *Knowledge Net* (see next section for more detail). *Knowledge Net*, however, fostered the most integrate and generate dialogue. This suggests that learners were making more connections between exhibit concepts and their prior experiences at *Knowledge Net* than the other two exhibits.

5.3 Knowledge Net

5.3.1 Ceiling and Floor Interactions. The session with the highest learning talk for *Knowledge Net* was G1.4 with a score of 310, 161 total codes, and a hold time of around an hour. The session with

Group Number	Group Description	Exhibits and Hold Time	Audio	Video
Group 1.1	Mom and 11 year old son	*Knowledge Net* (17:01) *Creature Features* (24:36)	Partial	Partial
Group 1.2	Mom (only engaged with KN), nanny (only engaged with CF), 9 year old girl	*Knowledge Net* (30:33) *Creature Features* (15:53)	Yes	Yes
Group 1.3	Mom, 11 and 8 year old sons	*Knowledge Net* (22:48) *Creature Features* (28:45)	Yes	Partial
Group 1.4	Mom, 8 year old daughter	*Knowledge Net* (1:01:43) *Creature Features* (20:16)	Yes	Yes
Group 1.5	Mom, 6 year old son	*LuminAI* (11:55) *Knowledge Net* (32:17)	Yes	Yes
Group 1.6	Mom, Dad, 11 and 12 year old daughters	*LuminAI* (26:36) *Creature Features* (27:10)	Yes	Yes
Group 1.7	Mom, 6 year old daughter	*LuminAI* (14:27) *Creature Features* (14:02)	Yes	Yes
Group 1.8	Mom, Dad, 11 year old son, 15 year old daughter	*LuminAI* (47:01) *Knowledge Net* (37:19)	Yes	Yes
Group 2.1	Mom, 9 year old daughter	*Knowledge Net* (20:55) *Creature Features* (21:48)	Yes	Yes
Group 2.2	Mom, 9 and 15 year old daughters	*Knowledge Net* (20:33) *Creature Features* (52:20)	Yes	No
Group 2.3	Mom, 10 year old daughter	*Knowledge Net* (54:00) *Creature Features* (36:20)	Yes	No
Group 2.4 (excluded from analysis)	Dad, 10 and 12 year old sons	*LuminAI* *Knowledge Net*	No	No
Group 2.5	Mom, Dad, 9 year old son	*LuminAI* (31:09) *Creature Features* (36:00)	Yes	Yes
Group 2.6	12 and 14 year old daughters, mom (took on observer role, did not engage)	*LuminAI* (14:14) *Knowledge Net* (22:55)	Yes	Yes

Table 5: Group descriptions and summary of data collected during at-home user studies

Exhibit	Num. Codes	LT Score	L1 Codes	L2 Codes	L3 Codes
Knowledge Net (n=10)	938, *Mdn*=100	1476, *Mdn*=144	510, *Mdn*=122	318, *Mdn*=59	110, *Mdn*=21
Creature Features (n=10)	736, *Mdn*=48	1475, *Mdn*=86	182, *Mdn*=49	369, *Mdn*=74	185, *Mdn*=23
LuminAI (n=6)	371, *Mdn*=59	702, *Mdn*=106	121, *Mdn*=30	169, *Mdn*=45	81, *Mdn*=14

Table 6: Learning talk score summary. N values indicate number of groups, not number of participants.

Exhibit	Instantiate	Evaluate	Integrate	Generate
Knowledge Net (n=10)	244, *Mdn*=21	77, *Mdn*=7	248, *Mdn*=30	310, *Mdn*=30
Creature Features (n=10)	327, *Mdn*=24	189, *Mdn*=14	62, *Mdn*=3	199, *Mdn*=17
LuminAI (n=6)	152, *Mdn*=25	99, *Mdn*=15	46, *Mdn*=7	74, *Mdn*=13

Table 7: Learning talk types summary. N values indicate number of groups, not number of participants.

the lowest score was G1.1 with a score of 19, but due to issues with the recording devices, this was only a partial recording of G1.1's interaction. The sessions with the next lowest score were G2.1 and G2.6, both with a score of 65 (hold times around 15-20 min, total codes around 40). The median score was 144. This indicates that learning talk scores for *Knowledge Net* overall were quite high (in comparison to the shortest interaction).

Groups with lower learning talk scores for *Knowledge Net* tended to move through the interaction quickly and in a linear, not iterative fashion. G2.1 and G2.6 both created relatively simple networks without a lot of discussion, tested them out with the chatbot, and moved onto the next activity. G1.1 actually engaged for longer and built a more complex network, but their video was cut short and they did not engage in as much verbal dialogue as the other groups did. Groups with higher learning talk scores generally spent a long

time constructing the network, considering many different tiles and nuanced connections between them.

Only a couple of groups (mostly G1.4) engaged for an extended period with the chatbot, asking it many questions and reflecting on their network's strengths/limitations. This led to more L1 codes and less L2 and L3 codes for the activity. Despite having the highest overall learning talk score, this was the one activity where learners could engage for a long period of time with the exhibit without really discussing topics that were highly relevant to the learning goals of the exhibit. Some groups that had quite high learning talk scores still did not appear to fully understand the chatbot. For example, one of the groups that engaged the longest with *Knowledge Net* was not able to get the chatbot to work because they thought they could talk to it using Siri on the iPad (rather than typing) ("Maybe we have to use speaker"..."You're trying to use the voice?" "Yeah, that's what I was trying" (G2.3)).

5.3.2 Description of Learning Talk. All participant groups built networks and connected tiles with arrows to represent relationships. *Knowledge Net* fostered a lot of discussion across groups about how to connect different concepts and what relationships to represent on the playmat. Many learner groups connected *Knowledge Net* to their personal interests, building networks about their personal lives (e.g. "Mom has child. / Or mom has me, right? / Right. Yeah, I like mom has me" (G2.3)) or making connections between the network and their interests (e.g. "Mommy used to play clarinet" (G1.2)).

Several groups had discussions about whether the data in their network should be general or specific–in other words, whether it should be true in all cases or just in a specific case (e.g. for their family). For example, G2.3 wondered whether the relationship Father DISLIKES cat had to be true in all cases: "Papa don't mind cats. / It doesn't have to be us, it could be anybody." Others considered creative or double meanings of certain connections–"I was about to say tails dislikes insects. Because you know horses have tails, and the tail keeps swatting the insect away" (G2.1); "Snake IS sticks. / Snake is like sticks, because it's slithery and like ... snakes. / Oh, snake is like a stick. Snake IS sticks. Okay" (G1.4).

A number of learners recognized some of the limitations of the knowledge representation they were building. Some learner groups noted that certain items on the playmat could not be connected due to the layout of the map, and that there were no "verb" connector arrows (e.g. runs, eats, plays) ("Well, there's no verbs. I mean, these are the verbs, so... / So, there's no actions? / Yeah./ ... There's like state of being, but no action" (G1.8)). These features limited the type of information they were able to teach the computer. Other learners recognized nuances of certain concepts that they were not able to capture in the semantic network–for example, G1.3 built a network with the relationship "boss HAS employee" but then realized this may not hold true for all employees and brought up the question of self-employment.

Of the groups that were able to interact with the chatbot (rather than doing the simulation role play activity, see section 3.1 for more detail), many engaged in relatively minimal conversation with the chatbot, using it only to confirm that their network had been accurately captured by the computer, then moving on. However, one 8 year old girl (G1.4) spent a lot of time engaging with the chatbot (hold time for the entire interaction was around an hour) and probing to understand what the AI learned from her network. The actual interaction with the chatbot clarified the connection between the network and how the computer used it. For example, in G1.4, the 8 year old girl asked the computer "What's in the carriage?" since the picture on the tile of a mother depicted a figure with a baby carriage. When the chatbot responded that it did not know what a carriage was, the girl had a discussion with her mom about how the computer did not really know what was depicted in each image (other than the name of the concept it represented). Unfortunately, this level of questioning and reflection did not occur with other groups.

Most learner groups seemed to recognize that their network was being used to teach or program the AI chatbot (e.g. "You can tell it anything" (G2.6)). This was even true for some of the learner groups that did the simulated exercise rather than using the actual chatbot. In G1.3, the mom played the role of the chatbot while her sons asked her questions. At one point, one of the sons was displeased with the result and asked "Can I edit your programming?" However, other groups that participated in the simulation activity got confused about how to act as the "computer" and not use their own knowledge–for instance, when one girl asked their mom (who was role-playing the computer) "What is clarinet?," the mom responded "Clarinet is another type of wind instrument that has one reed. A flute doesn't have any reeds, did you know that?" (G1.2).

During the discussion portion of the activity, some learner groups recognized that if you put false information in the network, the chatbot would give incorrect results ("So what happens if you put false information in the semantic network / Well it says it *does* have it" (G2.6)). A few groups tested this out and built networks containing false information ("Nose DISLIKES eyes, that's not true. I'll see what that does." (G1.5); "Does, oh no, sister HAS sheep. No. Wait, I'm going to say sister HAS eggs" (G2.6)).

One of the discussion questions asked learners to reflect on whether the computer really understands what a cat is, for example, if the computer does not know any information outside of what was built into the semantic network. Most learners felt that the computer did not fully understand what a cat was. Several groups also engaged in fruitful discussions about whether or not human reasoning processes were similar to or different from the semantic network ("I think it's different because, well, I mean, no, not really. I mean it tells someone something, then it kind of understands it." (G2.6); "Different...Because computers don't really learn on their own. They have to be taught more. And humans, they kind of learn at their own pace while computers, they just all learned the same way." (G1.3); "...in our brains we subconsciously know everything that we know about this one thing, so every time it's like brought up...we're thinking about all of the things that we know and all of the things that have connections to that one thing, when that one thing is said. But with the bot, it only knows... / Right. / That stuff. / So it doesn't have experience from before" (G1.8)).

An unexpected result we had was that *Knowledge Net* engaged younger learners in discussions where they were able to learn about topics unrelated to AI education. For example, in a mother-daughter duo (G2.3), the mom was able to teach her daughter several new words for relationships throughout the activity (e.g. "What's a peer? / You don't know what a peer is? A peer is like someone your age

Key Takeaways for *Knowledge Net*
Consistently supported long interaction times and physical and verbal collaboration
Had a low barrier of entry for adults with lower technical literacy and younger kids
Consistently fostered a lot of learning talk during network construction, though at a lower relevance level (L1, L2)
Did not foster as much dialogue about the actual chatbot interaction, which limited *evaluation* and learning talk that was more relevant to key competencies (L3)
Learners made connections between exhibit concepts and prior experiences and interests (as evidenced by *integrate* and *generate* dialogue)
Supported dialogue with potential for cross-disciplinary learning

Table 8: Key takeways for *Knowledge Net*

/ Oh, someone who I hang out with? / Yes"). This suggests that *Knowledge Net* may be well-suited for interdisciplinary settings (e.g. learning about AI in the context of a biology course or a unit on the self/family). Existing research investigating how to integrate AI education into core curriculum areas could inform future exploration on how to adapt *Knowledge Net* for interdisciplinary learning contexts [42].

5.4 Creature Features

Key takeaways from our analysis of learning talk at *Creature Features* are summarized in Table 9. When designing *Creature Features*, we envisioned a quality interaction with the activity to look as follows. A learner group would build a dataset from the bird cards, considering the features on the cards and which cards to include. Learners would then photograph the playmat and upload it to the website to see results. If the recognition algorithm performed poorly, we expected that learner groups would iterate on their dataset, perhaps more carefully considering the features on the cards or looking at a wider/more representative variety of creatures to include in the dataset. Discussion around which cards to include in the dataset at this stage would demonstrate understanding that the computer learned from the data on the board, recognition of the role that features and variety play in determining the results, and learners would generate hypotheses and predictions about the computer's decision-making processes. During the discussion portion of the activity, learner groups may begin to comment on strengths/weaknesses of this approach and connect what they have learned to other application domains (e.g. facial recognition).

5.4.1 Ceiling and Floor Interactions. The session with the highest learning talk for *Creature Features* was G2.5 with a score of 398, 188 total codes and a hold time of around 36 minutes. The session with the lowest score was G1.1 with a score of 43, 22 total codes, and a hold time of around 24 minutes. The median score was 86. This illustrates that there was a wide variance in learning talk scores, but that most of the interactions fell on the lower end (only three groups had a score of over 100). This does not necessarily mean that the groups with lower scores had poor or low quality interactions. Our observations indicate that the groups with comparatively lower scores tended to create a dataset after considering several different creatures, test it out, and iterate on it once or twice. This includes G1.1, where there was not as much learning dialogue mostly because the 11 year old took the lead on the dataset creation/iteration and did not engage in a lot of dialogue with his mom during the activity.

The groups with higher learning talk scores (over 100) at *Creature Features* tended to engage in more in depth discussions of the features on the cards and make predictions about what the algorithm would do when they made changes to their dataset. These groups sometimes hit a ceiling in the interaction when they had iterated numerous times on their dataset and could not figure out why their score was not improving or reaching 100% accuracy.

5.4.2 Description of Learning Talk. All learner groups engaged in initial discussion of which cards to include in their dataset(s). In some cases, the initial decision-making process was quick–learners did not consider many cards and made decisions based on personal preference or prior knowledge rather than the features on the cards (e.g. "And a turkey vulture because they get rid of roadkill" (G2.1), "Which ones do you think you would see around here?" (G1.2)). Other groups made initial decisions based on which cards they thought were most "birdlike" (e.g. "So which one would identify mostly as a bird? What examples and features on each one?" (G2.1?); "Honestly, a cardinal is the most bird-like thing here." (G1.4); "Well I think hummingbirds / Are important. / I think they're a pretty good representation. They're small but...they've got wings / Wings, beaks. / And a beak. / They look like birds" (G2.5)) or based on features that they thought most birds have (e.g. "I mean, can you think of another animal that's not a bird that lays eggs?" (G1.4)).

There were many features for people to consider and so some groups narrowed their focus to a few different features rather than thinking about all of them. For example, G1.1 tried to create a dataset that represented birds of a variety of sizes ("So you're going to put one right there. Two right there. Two right there. And all of these are big birds, large birds. / We only have two left. / So these are... this is a small bird, and these two are medium sized birds. (G1.1)). The groups with higher learner talk scores tended to discuss a wider variety of features and spent more time considering how placing extra weight on certain examples would affect the outcome of the algorithm (e.g. "Because if we only give more stuff to the cardinal and the robin, then it might ignore other things that are like... It might ignore a chicken because it doesn't fly." (G1.3); "So I would think for the kiwi, I would give it less emphasis because the kiwi actually doesn't fly. It's flightless like a penguin and or an ostrich, or an emu. But it still has the beak and lays eggs" (G2.5)).

Most learner groups recognized that they were in charge of teaching the computer how to recognize a bird (e.g. "Let's see what the robot learns" (G2.5)), but we did see some confusion between computer knowledge and human knowledge during the activity. In G2.3, the girl very carefully considered the features on the cards, but was focused on including features that she thought all people would want to know (e.g. "Habitats, yes, because some people might want to go looking, not just to hunt it down, but some people might want to know what area it lives in to maybe get a good view at it"). She included cards with features that she thought humans

might not know and focused less on features she thought all people would already understand (e.g. "I don't know what basically a heart chamber is, so I'm going to say that it is necessary for them to know, because I don't know it, so I'm pretty sure a lot of kids out there wouldn't know it" (G2.3); "I don't think it's necessary that they need to know that it has a beak, because usually all birds have beaks." (G2.3)). Later, spurred by the mom, they had a discussion about the difference between the computer's knowledge and the human's knowledge–"For instance, you may know that a bird has a beak. / Yes. / But does the computer know the bird has a beak? / No."

All groups tested their initial datasets and discussed the results. Participants often expressed surprise over the results ("A platypus!" (G1.6); "How does it get crocodile?" (G1.3); "I don't get the penguin. I don't know why." (G2.5)). Many made hypotheses about why the computer wrongly categorized certain creatures ("I mean I get why they say airplane, and I also get why they– / Because of the wings? / ... and I also get why they said Superman, he could fly." (G2.1); "Maybe it's because it's the only bird that swims. So it's getting more non birds don't swim data than birds can swim" (G2.5)).

All groups except G1.4 iterated on their datasets at least once (G1.4 spent a very long time on the *Knowledge Net* activity prior to *Creature Features* and the daughter was growing tired). Groups with higher learning talk scores iterated multiple times. Sometimes the results spurred participants to pay closer attention to the features on the cards ("Make sure you're looking at the features that's listed. / Which features? / The features on here. You see some of these might have a lot in common features with non-birds." (G2.1)). Groups tried to improve their scores in a variety of ways, such as switching out cards to add more variety (e.g. including a swimming bird) or weighting cards differently (e.g. less weight on a non-bird like a chipmunk). Groups were often disappointed if their score went down on the second iteration ("That's really confusing. 59 is not good." (G2.3); "So this time it says that based upon our data, it was only able to do 77% of the time instead of 90. I'm surprised. / ... This is hard. What are we supposed to do?" (G1.6)).

During the discussion portion of the activity, some groups expressed surprise by the role the human played in programming the AI ("Were you surprised at all by the role that people played in determining whether an AI algorithm works? / I would say yes / All right. I think so because we got to input this information for it to learn and work properly / Instead of looking it up on Google" (G2.1); "Kind of surprised, because I thought you would basically just, I don't know. I just, it seems a lot harder than I thought it was, I guess, because you have to try to train it the right way. But if you like steer it towards one direction of a certain type of bird, then it might think the types of birds are not birds, or they're not sure about it." (G1.6)). Other groups were more familiar with the role humans played in programming computers ("So did you know that this is how it works, that people feed data into AI and AI learns? / Only because we did a unit in computer science" (G1.3)).

As in *Knowledge Net*, some groups recognized that they could teach the AI false information: "And I bet we could rig it, we could, trick it to get really low percentages, you know, where it's really bad at figuring out what a bird is. If humans have to make the AI smart, could humans make not smart AI? Yeah? If we didn't do a good job" (G2.5). One group (G1.3) purposefully made a bad training dataset to see what would happen ("So it put Superman as a bird? / Yes. / Because you gave it some weight, right? / Right./ Okay, well that'll do it (G1.3)).

Key Takeaways for *Creature Features*
Supported constructive controversy dialogue [16]
Supported *evaluation* dialogue related to AI literacy competencies such as the steps and practices of machine learning and how agents make decisions
Of all three activities, CF had the most evenly distributed relevance of learning talk (between L1, L2, and L3)
Wide variance of learning talk between participant groups
Most groups grasped basic idea and iterated on dataset at least once, but did not engage in in-depth discussion
Iterative cycle of testing and revision supported in-depth discussion of features and their impact on the algorithm for some groups
Most successful with families with older children (ages 10 and up)
Some groups grew frustrated when they hit a ceiling and could not improve score further
Need more scaffolding to support learner groups with less prior knowledge in transferring knowledge from the activity to other contexts

Table 9: Key takeaways for *Creature Features*

Only a few learner groups (mostly those with higher overall learning talk scores) considered how what they learned in this activity might transfer to other applications. Most of the connections made were explained by parents to their kids. Some examples are included below.

> G1.6 (parent): I can give you an example too, like with our video cameras...It's really good at identifying people during the day, but at night it goes to black and white. And so there's, I don't think, the colors and stuff, I think, make it difficult to pick out people more because it's just black, gray and white, and it's a lot harder. So I think if it had been trained on lots of normal color, daytime pictures than black and white pictures, it would have been better.
>
> G1.6 (child): I think recognizing a face would be really hard because it has to do every exact detail and stuff. And on here we found that it's kind of hard for it to recognize the details and stuff.
>
> P1 (parent): If you said I'm going to rely on this computer program to pick human faces out of pictures–
> P2 (child): It could say the wrong face, it could say someone's face is not the correct one.
> P3 (parent): Make a mistake.
> P1 (parent): Yeah. Pick something. That's not a face, but maybe looks like a face with either eyes or a mouth, but really it's like a tree with a wood pattern or something. (G2.5)

5.5 LuminAI

Key takeaways from our analysis of learning talk at *LuminAI* are summarized in Table 10. We envisioned a high quality interaction with *LuminAI* starting with family members testing out and dancing with the agent. They may notice things as they dance, like recognizing that the agent is responding to them or learning from their movements. They may also recognize movements that the agent is performing. As the family member interacts with the agent for longer, another family member might investigate how to toggle some of the dancer "settings." They could try out switching a database from hip-hop to ballet, or switch response modes to see what happens when the agent responds with moves that are different from the ones you have performed. This may spur new observations about the agent's abilities. After dancing with the agent for awhile, the family could check out the *MoViz* tab and move around in the virtual space to explore the different clusters of agents. They might switch between databases and different algorithms to see how that affects the composition of the clusters, and compare different gestures that are in each cluster. They might look around to try to find their own dance moves and see where they are clustered.

5.5.1 Ceiling and Floor Interactions. The session with the highest learning talk for *LuminAI* was G1.8 with a score of 221, 113 total codes, and a hold time of around 47 minutes. The session with the lowest learning talk was G2.6 with a score of 50, 23 total codes, and a hold time of around 11 minutes. G1.8 danced with the agent for a long period of time, putting on music in the background and taking turn teaching it dance moves and toggling between the controls. They engaged with *MoViz* for a shorter period of time (but longer than most other groups) and tried out all of the features. They then engaged in a lengthy discussion about the strengths and weaknesses of the agent as a dancer and whether or not the AI was creative.

G2.6 (two teenage girls) interacted with all exhibit components despite their low learning talk score and short interaction time. They 1) verbally recognized that the agent responded to and learned from them; 2) noticed that the agent modified their dance moves; 3) engaged with *MoViz* and noticed that different clusters had different types of movements in them. This indicates that even though there is a high ceiling for interaction (indicated by G1.8), groups that interact for shorter periods of time (for example, in a museum) can still interact with all components of the installation and touch on the key learning goals.

However, some of the other groups that had lower scores had younger kids that engaged primarily with the dancer on *mimic* mode and were not interested in interacting with *MoViz*. This a pattern we have observed in our prior installations of the *LuminAI* exhibit–young kids often enjoy seeing a "magic mirror" effect while older kids are more likely to recognize the backend reasoning capabilities of the agent [50].

MoViz engaged older kids and adults more than our youngest participants, although learners in general did not spend a very long time interacting with it. Groups with higher levels of learning talk engaged with the agent for longer, trying out multiple different response modes and datasets and spending a shorter period of time interacting with and inspecting the gestures in *MoViz*. Most learners did not engage much with the toggle feature to try out different clustering algorithms–this may need further explanation or improved UI scaffolding if we include it in a future exhibit.

5.5.2 Description of Learning Talk. In the vast majority of the interactions (even ones with lower learning talk scores), learners recognized that the AI was learning from them. This is an improvement over previous versions of *LuminAI* that did not have the tools built in to foster learning through interaction [49]. This improvement was evidenced by learners' commentary about teaching the AI ("Maybe you should try to teach it one of the dances from your dance class" (G1.8); "That's the AI. It's learning my moves" (G2.6); "How did your actions affect the AI's dance moves / Well I told them some dance moves, right? I put dancers in the memory" (G2.6); "Whoa, look at it doing [name]'s dance." (G1.5)).

Most learner groups interacted with the toggles on the menu that allowed participants to switch out the database or try different response modes (e.g. "So, you can do user dances, popular dances, ballet, contemporary dance. You should do ballet. Do your ballet from your class." (G1.8)). Some learner groups engaged in discussion about the way the response modes affected the interaction (e.g. "Well I only did mimic, but then the other one, it kind of, it was contrast with mine, so it was like we were dancing together" (G2.6); "Do something to the contrast. Okay, so do a dance. / So it'll do something different? / Yeah. All right, okay. / Okay, hold still. / So it's doing bigger movements. / And slower maybe" (G2.5)).

The youngest children in our study (age 6, G1.5 and G1.7) had a harder time recognizing that the agent was learning from them and just enjoyed dancing with the AI (without engaging with the learning content). When the parents started to look at the *MoViz* tab, the youngest kids were uninterested and wanted to return to dancing ("I don't want these ones, I want the dancing. I just want the dancing" (G1.5); "I don't really get this one" (G1.7)). The younger kids did enjoy having the databases switched to try different types of dance ("I want the ballet!" (G1.5)) and recognized when the AI was performing a common dance move ("It's dabbing" (G1.7)).

We coded for a variety of different learning interactions participants could have with the *MoViz* interface. Most groups did not engage for a long period of time with or talk much about *MoViz*. Several groups observed differences between the different clusters of dance moves ("I feel like the green ones are very hip-hoppy I think / Yeah / And the yellow ones...are the classic moves I think" (G2.6); "Yeah, doesn't it seem like the green ones are more toward centered...and then the blue ones are more out from the body?" (G1.8)).

During the discussion portion of the activity, several learner groups engaged in discussion of whether the computer was creative and what qualified as creativity (e.g. "I think it's creative because it learns and it has the mimic but it also has the do something new with it." (G1.8)). Learners also discussed the strengths and limitations of the AI agent and its representation of the human body. For example, the mom in G1.7 brought up a technique her daughter was learning in dance class–"They're pretty soft. And we've been learning about soft arms. / Mm-hmm (affirmative). / It's kind of hard for the computer to do super soft, smooth movements. Right? It's a little bit more rigid"; "It seems like the AI is not as graceful as a human. I don't know how you quantify that..." (G1.8); "It doesn't have the physical constraints that humans do" (G1.8)).

Key Takeaways for *LuminAI*
Collaboration took the form of turn-taking
Most groups with lower interaction times were still able to touch on all relevant competencies
Youngest participants (age 6) did not engage with competencies
Supported *evaluation* talk related to knowledge representations and agent decision-making
Supported a high ceiling for interaction time and discussion
Learners (especially younger kids) did not engage as much with *MoViz* as with the main system
Spurred discussion about AI and creativity

Table 10: Key takeways for *LuminAI*

Some groups also discussed how the AI agent and humans were similar or different: "Well, like I said before, it takes other dances and it can make it its own or it can learn that specific dance. That's kind of how humans do it, too, because if you have a choreographer, you get taught the exact dance but then if you are your own choreographer, you make up your own dance based off of other people's dances. So it can do both of those things" (G1.8).

6 DISCUSSION: REFLECTING ON AI LITERACY DESIGN PRINCIPLES

In this section, we return to the AI literacy design principles (DPs) we originally developed based on a review of related literature in the field [51]. Our empirical investigations seek to contribute a deeper understanding of how these principles work in practice. We review most of the principles, but skip over several that our exhibits did not explicitly address or provide significant insight into.

6.1 DP1: Explainability and DP4: Promote Transparency

DP1: Explainability calls for making interactions with AI "more explainable," making functions transparent where possible. *DP2: Transparency* calls for promoting transparency in all aspects of AI design [51]. We discuss these design principles together since they are closely related. We aimed to increase transparency in *LuminAI*, taking a previously opaque human-AI interaction and adding in user controls and visualizations of the AI's reasoning. We explicitly designed the *LuminAI* UI to be exploratory and open ended, aiming to foster creative expression and active prolonged engagement [34] in addition to making the AI algorithm "explainable."

Most learner groups successfully engaged with the part of the *LuminAI* interface that allowed learners to toggle response modes and databases. Learners recognized that the agent was able to learn from them and discussed the different ways in which the agents were able to respond, which is not behavior we have observed in prior versions of the system without the educational interface components.

However, more learner groups struggled to use the *MoViz* interface, especially those with younger kids. Even the learner groups that engaged with *MoViz* only did so for a short period of time. Some learners expressed that they were overwhelmed by the *MoViz* interface (e.g. "Whoa, limb centroids has way more green guys than the...temporal clustering guys. I don't really understand what these differences are. (G1.5); "All right. Now you can do it. So these are, this is temporal clustering. The other was limb centroids. / What's that? / I'm not sure exactly what that means" (G1.8)).

These findings suggest that although they can lead to learning, *explanatory interactive visualizations provided with AI algorithms may be intimidating for novice users and require additional scaffolding*. It is also possible that the *MoViz* UI was confusing to users and this issue is specific to our project–additional investigation into this question is needed.

The less intimidating nature of the *LuminAI* response mode interface indicates that *an effective way to expose the inner workings of AI to a novice audience may be to make components of the algorithm itself adjustable/customizable, rather than or in addition to providing explanatory visualizations*. Incorporating user controls in particular led to significantly more understanding of the agent's abilities than we have seen in prior installations of this project [36, 49]. This suggests that allowing learners to interactively explore/customize components of AI projects could be useful in other contexts (e.g. workplaces) to familiarize users with aspects of AI systems they are using.

Learners requested more "explainability" in the *Creature Features* exhibit, asking for explanations of why the machine learning algorithm classified certain creatures the way it did in order to aid in the iterative design of their datasets. The lack of explanation present in the algorithm's feedback placed a "ceiling" on the interaction, preventing learners from further iterating on their designs when they could not figure out why their accuracy scores were not improving. This emphasizes the potential of explainability to deepen learners' engagement with AI literacy learning activities.

6.2 DP2: Embodied Interactions

Our explorations of embodied interaction in exhibit design demonstrated that *tangible and full-body interfaces were engaging and lowered the barrier of entry for learners, but embodied metaphors needed to be made more explicit to minimize learner confusion. Creature Features* and *Knowledge Net* utilized tangible pieces like tiles, cards, and tokens as interfaces for training or teaching AI algorithms, drawing on Horn's theory of cultural forms [31]. *Creature Features* and *LuminAI* both used embodied metaphors to communicate abstract AI concepts. In *Creature Features*, "weight" tokens are used to place emphasis on a specific item in a dataset. In *LuminAI*, a 3D virtual space is used to visualize clusters of gestures created by an unsupervised learning algorithm. *LuminAI* was the only exhibit that involved a full-body experience [74].

The tangible interfaces were largely a success, receiving positive feedback from many participants and lowering the barrier to entry for teaching and training AI algorithms. The full-body dance interaction was also engaging for users–many learners found dancing with *LuminAI* to be particularly fun and enjoyed seeing the agent respond to their personal dance moves.

Interaction and discussion time was often skewed towards the embodied component of activities. For instance, in *Knowledge Net*, participants focused heavily on selecting tiles and relationships to construct the network, with less time spent interacting with and

discussing the chatbot. Similarly, participants spent significantly more time dancing with *LuminAI* than they did exploring the *MoViz* interactive visualization. This imbalance points to the engaging nature of the embodied interactions but raises additional research questions about how to foster AI learning experiences that span both physical and digital interfaces.

The embodied metaphors in *Creature Features* and *LuminAI* elicited more mixed results. An early paper-based prototype of *Creature Features*–called *Neural Net*–involved a very literal weight metaphor in which learners physically placed weight on different nodes of an actual net [48]. Learners expressed confusion over the exact effect of the weight on the results of the algorithm. Some of this confusion was mitigated by focusing on a feature-based learning algorithm in the next iteration of *Creature Features*, but several learners still expressed confusion over the exact functionality of the weight tokens. The final version of *Creature Features* resolved most of this confusion (indicated by lengthier hold times and no participant survey comments about confusion) by making the function of the tokens much more explicit in the text engraved on the gameboard ("How many of this creature do you want to include in your dataset?"). Numerous participants also expressed confusion over the *MoViz* interface in *LuminAI*, not fully understanding why gestures were grouped together in space or why some were closer to each other than others. These findings suggest that if designers incorporate embodied metaphors in the design of AI literacy learning experiences, the connection between the embodied metaphor and its relationship to the algorithm needs to be made very explicit. This is supported by prior work that suggests that mappings between embodied metaphors and concepts must be readily discoverable in order to be understood by learners [7].

6.3 DP3: Contextualizing Data

DP3: Contextualizing Data suggests encouraging learners to investigate contextual information about datasets such as how they were collected, who they were made by, and what the limitations of the data are. Prior work suggests that data that is relevant to learners' lives, low-dimensional, and/or "messy" (not clean or neatly categorizable) can aid in data contextualization [15].

Creature Features was the exhibit prototype that touched the most on this design principle. Learners were put in the shoes of a data scientist and asked to curate a low-dimensional dataset to define a "messy" concept–what a bird is. The low dimensionality of the dataset (six items, multiplied by the number of weight tokens on each) created a low barrier of entry for learners and allowed learners to discuss details of each item they were considering for inclusion in the dataset. The constraint of only being able to include six creatures in the dataset (or, in the later iteration, three positive examples and three negative examples) was intended to make learners carefully consider and discuss which creatures to include. However, it also imposed a "ceiling" for interaction–some of the more engaged learner groups wanted to continue improving their results, but after a certain point felt they could not be optimized further without expanding their dataset beyond the allotted number of spaces.

Overall, the "messiness" of defining what a bird is seemed to be an effective method of encouraging learners to consider the challenges involved in developing classification algorithms. One learner commented that the "nuances" of the bird data is what made the activity interesting. We chose birds as the topic of the activity both because of the "messiness" of categorizing birds and because we thought most learners would be able to draw on their prior knowledge when talking about birds. However, some learners were not particularly interested in birds as a topic. Recent research has found that learners interacting with personal data are able to better advocate for themselves in cases of wrongdoing related to machine learning [65]. Using personal data in *Creature Features* in the future might be more effective at engaging all learners and help learners make deeper connections to AI-related ethical issues.

One potentially contrasting view on using personal data was brought up by Touretzky and Gardner-McCune in a recent book chapter preprint [78]. Touretzky and Gardner-McCune hypothesize that learners may have an easier time building an understanding of machine learning when they are teaching the computer about topics they do not have prior knowledge of, since they will be forced to consider gaps in the machine's knowledge that they cannot fill on their own. This hypothesis has yet to be empirically tested but does find some support in participant interactions with *Creature Features*. Since learners had considerable prior knowledge about the topic (birds), they often based their decisions off of features drawn from their prior knowledge rather than the features listed on the cards. This meant that learners were sometimes grounding their decisions in data the AI had no conception of (e.g. whether a certain bird lived in a participant's backyard, which birds the participant liked the most). If learners had been teaching the exhibit about a totally novel topic–e.g. features of imaginary cartoon characters–they would have been forced to consider the features on the cards.

6.4 DP5: Unveil Gradually

DP5: Unveil Gradually talks about unveiling components of AI systems gradually and using scaffolding to reduce cognitive load. All three of the exhibits involved compartmentalized activities to help learners grasp one concept before moving onto the next. *Creature Features* and *Knowledge Net* had clearly delineated co-construction and testing stages, and the *LuminAI* interface was separated into two different tabs, encouraging learners to dance with the AI agent before moving on and engaging with the *MoViz* interface. This generally worked effectively but at times the gradual unveiling of system components may have caused learners to focus on one component at the expense of the other. In *Knowledge Net*, learners often spent so long constructing the network that they did not invest much time in interacting with the chatbot. In *LuminAI*, learners similarly often danced with the agent for a long time and only had a little bit of time remaining to engage with *MoViz*.

The *MoViz* interface could also have used some additional scaffolding for learners, perhaps interactively explaining the clusters of dance moves. In addition, most learners did not engage much with the toggle to try out different clustering algorithms, or expressed confusion when they did. It could be that this was one feature too many to digest in a complex user experience–it should either be unveiled gradually with scaffolding to support its introduction or excluded from the experience to allow learners to focus on the other components.

6.5 DP7: Milestones

DP7: Milestones suggests the importance of considering developmental milestones when designing AI literacy learning interventions. *Creature Features* was not as engaging for our youngest participants (6-7) but was quite engaging for just slightly older kids (10-11). Young learners similarly enjoyed dancing with *LuminAI* but were not nearly as interested in *MoViz* (although most of the kids we tested the final version of *LuminAI* with were over 10, this observation may need to be verified in a larger study). These findings suggest that in addition to age-appropriate learning outcomes (which groups like AI4K12 are currently in the process of developing [79]), researchers may need to consider developing a set of AI literacy design principles that are specific to particular age bands, investigating what types of activities are most engaging for learners of different ages.

6.6 DP10: Support for Parents

DP10: Support for Parents calls for providing support for parents when engaging families in learning about AI. We tried to keep technical terms to a minimum in the instructions/exhibit explanations, but occasionally used terms such as "algorithm," "dataset," and "semantic network" when explaining the activities (we purposefully avoided using the term *network* to describe the machine learning algorithm in later iterations of *Creature Features* both for accuracy and to reduce confusion between a neural network and a semantic network). Parents often stumbled over technical terms, unsure of how to define/explain them to their kids and sometimes stumbling on their pronunciation or skipping over them entirely when reading the instructions aloud. We did not interrogate this further during the studies, but technical terms can be intimidating and discouraging for novice learners [58]. Providing adequate support for parents guiding their kids through AI literacy activities may involve 1) reducing the use of unnecessary technical language as much as possible; 2) providing a glossary of relevant terms with explanations and pronunciations that parents can use to aid in explaining new topics to their kids; and/or 3) pausing to explain new terms in a non-condescending way as they come up during the activity.

Sometimes the learning dialogue at the exhibits seemed to indicate that learners had skipped over some of the basics (What is AI?) to a more advanced topic (What is a semantic network?). This worked fine for groups that had some prior knowledge of AI, but some groups appeared to walk away without a clear sense of what technologies used AI despite understanding how a semantic network worked. We included a page in the instruction packet that provided a brief introduction to AI, but many learners skipped over it as it was lengthy to read aloud (as many learners would probably skip over a lengthy sign-based explanation in a museum). In contrast, we have previously conducted interactive paper-based activities in co-design workshops that have been effective at engaging participants in dialogue about what AI was and what their preconceptions about it were [48]. *Museums that are considering implementing AI related installations may want to ensure they have an introductory exhibit or activity similar to the one we provided in the worksheet packet in the co-design study.*

Future research is needed that more closely investigates the roles that parents and children take on when learning about AI and explores the types of scaffolding that parents provide for their children (c.f. [91]). This work could provide valuable insight into how to best support parents in AI literacy learning activities.

6.7 DP11: Social Interaction

Both adults and kids commented how much they enjoyed getting to do the activity with their families and learn about their family members' perspectives on the topics covered. *All three exhibits supported collaboration, although the form it took looked different at each exhibit. Knowledge Net* was the most successful at consistently supporting both physical and verbal collaboration amongst group members of all ages. *Creature Features* was particularly successful in many cases at facilitating "constructive controversy" [16, 38] as learners debated which creatures to include and emphasize in their datasets. There were a few breakdowns in collaboration that occurred when, for example, two kids each developed a hypothesis and wanted to test it out at the same time. *LuminAI* facilitated more of a turn-taking style of group work, where one learner actively engaged with the system at a time while others looked on and discussed together. Our prior research suggests that *LuminAI* could easily facilitate more joint collaboration if installed in a larger space with additional interaction stations [49].

Research on designing co-creative AI for public spaces suggests that facilitating multiple levels of entry is essential for fostering collaboration [50]. This issue arose a few times in the exhibit prototypes. For example, our youngest participants (6-8 years old) often quickly lost interest in *Creature Features* (and we did not test the exhibits with even younger kids). This meant that in some cases, older children and adults were not able to engage as deeply as they wanted to in the exhibit. This issue could lead to even more rapid disengagement in a museum setting as families often move on to other exhibits when one member gets distracted. Providing alternative levels of entry or explicit activities for young learners is particularly important at AI education exhibits, which may be difficult to grasp for learners under seven who have not yet developed theory of mind (i.e. our ability to "explain and predict other people's behavior by attributing to them independent mental states, such as beliefs and desires") [26, 86]. We tried to incorporate explicit activities for younger kids in the second iteration of *Creature Features*, where we assigned the "youngest member of the group" to find bird and non-bird cards in the deck for the group–however, additional activities would likely be needed to sustain interaction in a museum space.

6.8 DP12: Leverage Learners' Interests

Allowing learners to incorporate their personal interests (*DP12: Leverage Learners' Interests*) may have made the exhibits more engaging. The instructions for *LuminAI* encouraged learners to put on their own music as they danced with the agent. Numerous families responded to the suggestion and played their favorite songs while interacting with the system. Many learners also tried to teach the agent dance moves they had learned before. Playing customizable music is something that would be more difficult to implement in a museum due to royalty fees and the presence of other visitors,

but was a nice way to allow learners to incorporate their personal interests in the at-home interaction environment.

Numerous families incorporated their interests into their creations with *Knowledge Net*. For example, several groups built networks about their own family relationships and interests. However, some learners wanted to customize their networks further, asking for expanded sets of tiles on different topics. Similarly, the topic of birds was uninteresting for several kids in particular, who may have been more engaged if the classification task related to their interests. Providing an increased number of options of tiles could lead to improved likelihood that the exhibit will intersect with learners' interests. Both *Knowledge Net* and *Creature Features* were designed so that they could be easily expanded with add-on card decks or tile sets.

Some learners expressed that they enjoyed the ability to use their prior knowledge and build networks and/or discuss creatures using information they already knew. We hypothesize that the ability to use prior knowledge during the activity made learning about AI less intimidating–one learner commented that they liked *Knowledge Net* because they could "make relationships between things they already knew." Bolstering learners' confidence by allowing them to succeed at pulling in prior knowledge and use familiar interfaces could enable them to explore activities they may otherwise feel less confident or discouraged in [19, 31]. However, certain aspects of the activities connected with broader groups than others. The tiles having to do with animals and family members were more widely relatable than the tiles focused on musical instruments (e.g. "These are musical. I don't know nothing about music" (G2.3)).

6.9 DP14: New Perspectives

DP 14: New Perspectives calls for AI literacy learning interventions that expose learners to new perspectives on AI that they may not have seen before. We aimed to do this in two ways. First, *Knowledge Net* was focused on competencies related to cognitive systems, which is a subdiscipline of AI that has not received as much attention as machine learning and robotics in recent media or the education space [51]. We also aimed to expose learners to creative AI via interaction with *LuminAI*, in an effort to spur them to consider applications of AI that are more open-ended and exploratory rather than mechanical or detached [61]. Discussion after the LuminAI activity yielded interesting family dialogue about whether or not computers can be creative. Overall, there is a significant space for future research related to introducing lesser-known types of AI research to the public. We hope that the way in which we adapted the *LuminAI* exhibit to an educational context can provide a model for other researchers looking to take similar steps to share their projects with the public.

6.10 DP15: Low barrier to entry

All three exhibits were successful at facilitating a low barrier of entry to learning about AI (DP15: Low Barrier to Entry), and numerous participants expressed that they were surprised that they were able to learn about how AI works without having much prior experience with the topic or being able to code. The embodied modes of interaction, abstracted interfaces for teaching and training AI, incorporation of topics that built on learners' prior knowledge, and minimization of necessary prerequisite technical knowledge likely all contributed to reducing the barrier of entry for participants. We did observe a few obstacles to facilitating a low barrier of entry–for example, the use of technical domain-specific vocabulary that learners may be unfamiliar with or the need to ensure that the mapping between abstracted interface and AI algorithm is authentic and explicit.

7 LIMITATIONS AND THREATS TO VALIDITY

The exhibits we designed did not provide significant insight into all of the design principles set forth in [51]. We did not focus on designing activities where learners could write code to program AI (DP6), both because we wanted to investigate what AI concepts learners could understand without any prerequisite programming knowledge and because numerous existing projects already explore how to integrate AI learning activities into programming platforms (e.g. [10, 17]). In addition, none of the three exhibits we developed focused explicitly on addressing learners' preconceptions about AI or on incorporating learners' identity, values, or backgrounds as key design principles (DP9, DP13). Future work will be needed to investigate these AI literacy design principles in more detail.

The learning talk framework we used to analyze participant dialogue provides insight into whether and how groups discussed relevant AI literacy competencies at each exhibit. However, it does not capture other aspects of social family group learning, such as the roles that parents and children take on in the interaction or the types of scaffolding that parents engage in when explaining concepts to their kids. Future work will be needed to investigate these important aspects of family group learning. Prior work studying parent-child perceptions of agent intelligence [18, 63] and identifying the types of scaffolding parents engage in and roles parents take on [75, 91] can inform future research in this space.

We made an effort to recruit a diverse population, and the demographic data we collected indicated that many groups did not have prior experience with computer programming or developing AI. However, selection bias may still influence the results of this study. The sample size was small due to the challenges of recruiting participants for a study during COVID-19, and some participants may have self-selected for the study due to an existing interest in AI or technology. Further studies will need to be conducted to determine whether our findings generalize to a larger population.

Finally, we had some challenges with missing data (see Table 5), both because participants were in charge of recording their own data and because of technical challenges with the video recorders we used. We provide some more detail on the challenges we faced here for context and for other researchers who may be conducting similar studies during the pandemic. We did not want participants to record data on their own personal devices due to the large file size of 2 hour long video recordings. We also did not want to assume a high level of technical literacy amongst participants, so we were looking for an easy-to-use camcorder where participants could simply turn on the camera, adjust its position to record their playspace, and hit record. However, due to the commercial popularity and camera quality on smartphones, most easy-to-use personal camcorder companies have gone out of business. The personal cameras

that are currently on the market are geared more towards high-quality recording for hobbyists or professionals, and are relatively expensive and complex to set up. We ended up purchasing several used Flip cameras, which are simple personal camcorders with an extremely easy-to-use interface. However, Flip went out of business in 2011, and we ran into technical challenges due to the age of the cameras–two of the cameras consistently died after 15 minutes of recording (a problem that did not reveal itself until we deployed the cameras). This caused technical difficulties for the participants (who had to keep switching out the batteries with replacements we provided) and caused us to not receive video data from some of the groups. For the second iteration of studies, we replaced the two faulty Flip cameras with GoPro-like cameras. These worked more consistently, but the lack of screen feedback meant that some groups thought their camera was recording when in fact it was not. Ultimately, we did not find an adequate solution for the camera issue. Although our policy of redundancy helped ensure that audio data was collected for almost all of the groups, two groups who did not use the audio recorder correctly still ended up with missing data.

8 CONCLUSION

In this paper, we present an analysis of participant learning talk at three different activities designed to encourage learning about AI literacy competencies. We examine the dialogue around each exhibit, then reflect on the implications for the AI literacy design principles we previously deffined based on a literature review [51]. This paper contributes an empirically grounded understanding of these design principles in light of the learning talk analysis we conducted. This paper can be useful to AI educators, designers of tools/interfaces for promoting AI literacy and explainability, and researchers seeking to better understand how people learn about AI.

ACKNOWLEDGMENTS

We would like to thank Aadarsh Padiyath, Jonathan Moon, Lucas Liu, Cassandra Naoimi, and Rhea Laroya for their contributions to prototype development; all of our study participants for participating in this study during a challenging time, and our collaborators at the Museum of Science and Industry, Chicago for their ongoing support of this research. This work was funded by the National Science Foundation (DRL #1612644).

REFERENCES

[1] Edith Ackermann. 2004. Constructing knowledge and transforming the world. *A learning zone of one's own: Sharing representations and flow in collaborative learning environments* 1 (2004), 15–37.

[2] Adam Agassi, Hadas Erel, Iddo Yehoshua Wald, and Oren Zuckerman. 2019. Scratch Nodes ML: A Playful System for Children to Create Gesture Recognition Classifiers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

[3] 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.

[4] Safinah Ali, Blakeley H Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, Ethics, and Creativity: Developing Primary and Middle School Artificial Intelligence Education. In *Proceedings of IJCAI 2019*. Palo Alto, CA.

[5] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[6] Sue Allen. 2003. Looking for learning in visitor talk: A methodological exploration. In *Learning conversations in museums*. Routledge, 265–309.

[7] Alissa N Antle, Greg Corness, and Milena Droumeva. 2009. Human-computer-intuition? Exploring the cognitive basis for intuition in embodied interaction. *International Journal of Arts and Technology* 2, 3 (2009), 235–254.

[8] The Barbican. 2019. AI: More Than Human. https://www.barbican.org.uk/whats-on/2019/event/ai-more-than-human

[9] Ruha Benjamin. 2019. Race after technology: Abolitionist tools for the new jim code. *Social Forces* (2019).

[10] Nikhil Bhatia. 2020. *Using transfer learning, spectrogram audio classification, and MIT app inventor to facilitate machine learning understanding*. PhD Thesis. Massachusetts Institute of Technology.

[11] Karen Brennan and Mitchel Resnick. 2012. New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American Educational Research Association, Vancouver, Canada*. 1–25.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[13] Patricia Charlton and Stefan Poslad. 2019. Engaging with computer science when solving tangible problems. In *Proceedings of the 3rd Conference on Computing Education Practice*. 1–4.

[14] Alexandra Chouldechova and Max G'Sell. 2017. Fairer and more accurate, but for whom?. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

[15] Catherine D'Ignazio. 2017. Creative data literacy. *Information Design Journal* 23, 1 (2017), 6–18.

[16] Willem Doise, Gabriel Mugny, and Juan-Antonio Pérez. 1998. The social construction of knowledge: Social marking and socio-cognitive conflict. *The psychology of the social* 77 (1998).

[17] Stefania Druga, Sarah T.Vu, Eesh Likhith, and Tammy Qiu. 2019. Inclusive AI literacy for kids around the world. In *Proceedings of FABlearn '19*. ACM, New York City, NY, USA.

[18] 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.

[19] Carol S Dweck. 2000. *Self-theories: Their Role in Motivation, Personality, and Development*. Psychology Press, Lillington, NC, USA.

[20] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, and Mark O Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv preprint arXiv:2107.13509* (2021).

[21] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[22] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, A Elazari, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. https://doi.org/10.1145/3290605.3300724

[23] John H Falk, Martin Storksdieck, and Lynn D Dierking. 2007. Investigating public science interest and understanding: Evidence for the importance of free-choice learning. *Public Understanding of Science* 16, 4 (2007), 455–469. Publisher: Sage Publications Sage UK: London, England.

[24] Luca Ferrari, Anita Macauda, Alessandro Soriani, and Veronica Russo. 2020. Educational robotics and artificial intelligence education: what priorities for schools? *Form at re-Open Journal per la formazione in rete* 20, 3 (2020), 68–85.

[25] Ars Electronica Futurelab. 2019. Understanding AI. https://ars.electronica.art/aeblog/en/2019/08/06/understanding-ai-futurelab-installations/

[26] Helen L Gallagher and Christopher D Frith. 2003. Functional imaging of 'theory of mind'. *Trends in cognitive sciences* 7, 2 (2003), 77–83.

[27] William W Gaver, Andrew Boucher, Sarah Pennington, and Brendan Walker. 2004. Cultural probes and the value of uncertainty. *Interactions* 11, 5 (2004), 53–56. Publisher: ACM New York, NY, USA.

[28] Christiane Gresse von Wangenheim, Lívia S Marques, and Jean C R Hauck. 2020. Machine Learning for All – Introducing Machine Learning in K-12. https://doi.org/10.31235/osf.io/wj5ne

[29] Joshua P Gutwill and Sue Allen. 2010. Facilitating family group inquiry at science museum exhibits. *Science Education* 94, 4 (2010), 710–742. Publisher: Wiley Online Library.

[30] Christian Heath and Dirk Vom Lehn. 2004. Configuring Reception: (Dis-) Regarding the 'Spectator'in Museums and Galleries. *Theory, Culture & Society* 21, 6 (2004), 43–65. Publisher: Sage London, Thousand Oaks and New Delhi.

[31] Michael S Horn. 2018. Tangible interaction and cultural forms: Supporting learning in informal environments. *Journal of the Learning Sciences* 27, 4 (2018), 632–665. Publisher: Taylor & Francis.

[32] Michael S. Horn, Amartya Banerjee, David Bar-El, and Izaiah Hakim Wallace. 2020. Engaging Families around Museum Exhibits: Comparing Tangible and Multi-Touch Interfaces. In *Proceedings of the Interaction Design and Children Conference (IDC '20)*. Association for Computing Machinery, New York, NY, USA, 556–566. https://doi.org/10.1145/3392063.3394443 event-place: London, United Kingdom.

[33] Eva Hornecker and Luigina Ciolfi. 2019. Human-computer interactions in museums. *Synthesis Lectures on Human-Centered Informatics* 12, 2 (2019), i–171.

[34] Thomas Humphrey, Joshua Gutwill, and The Exploratorium APE Team. 2005. *Fostering Active Prolonged Engagement: The Art of Creating APE Exhibits*. Routledge, Abingdon, UK.

[35] Science and Economic Development Canada Innovation. 2021. *Views of Canadians on Artificial Intelligence: Final Report*. Technical Report. https://www.competitionbureau.gc.ca/eic/site/112.nsf/eng/07662.html#S423

[36] Mikhail Jacob, Gaëtan Coisne, Akshay Gupta, Ivan Sysoev, Gaurav Verma, and Brian Magerko. 2013. Viewpoints AI. In *Proceedings of the Ninth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE '13)*. AAAI, Boston, MA, USA.

[37] Mikhail Jacob and Brian Magerko. 2015. Viewpoints AI. In *Proceedings of the Artwork Exhibition at the 10th ACM Conference on Creativity and Cognition*. Glasgow, Scotland.

[38] David W Johnson and Roger T Johnson. 1979. Conflict in the classroom: Controversy and learning. *Review of educational research* 49, 1 (1979), 51–69.

[39] Brian Jordan, Nisha Devasia, Jenna Hong, Randi Williams, and Cynthia Breazeal. 2021. PoseBlocks: A Toolkit for Creating (and Dancing) with AI. In *The 11th Symposium on Educational Advances in Artificial Intelligence*.

[40] Eunkyoung Lee. 2020. A comparative analysis of contents related to artificial intelligence in national and international K-12 curriculum. *The Journal of Korean Association of Computer Education* 23, 1 (2020), 37–44. Publisher: The Korean Association of Computer Education.

[41] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. 191–197.

[42] Phoebe Lin and Jessica Van Brummelen. 2021. Engaging Teachers to Co-Design Integrated AI Curriculum for K-12 Classrooms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[43] Phoebe Lin, Jessica Van Brummelen, Galit Lukin, Randi Williams, and Cynthia Breazeal. 2020. Zhorai: Designing a Conversational Agent for Children to Explore Machine Learning Concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13381–13388. Issue: 09.

[44] Robb Lindgren and J Michael Moshell. 2011. Supporting children's learning with body-based metaphors in a mixed reality environment. In *Proceedings of the 10th International Conference on Interaction Design and Children*. 177–180.

[45] Annabel Lindner, Stefan Seegerer, and Ralf Romeike. 2019. Unplugged Activities in the Context of AI. In *International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*. Springer, 123–135.

[46] Lucas Liu, Duri Long, and Brian Magerko. 2020. MoViz: A Visualization Tool for Comparing Motion Capture Data Clustering Algorithms. In *Proceedings of the 7th International Conference on Movement and Computing*. 1–8.

[47] Duri Long. 2021. *Designing Co-Creative, Embodied AI Literacy Interventions for Informal Learning Spaces*. Ph. D. Dissertation. Georgia Institute of Technology. http://hdl.handle.net/1853/64754

[48] Duri Long, Takeria Blunt, and Brian Magerko. 2021. Co-Designing AI Literacy Exhibits for Informal Learning Spaces. In *Accepted to Proceedings of The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*.

[49] Duri Long, Mikhail Jacob, Nicholas Davis, and Brian Magerko. 2017. Designing for Socially Interactive Systems. In *Proceedings of the 11th Conference on Creativity and Cognition*.

[50] Duri Long, Mikhail Jacob, and Brian Magerko. 2019. Designing Co-Creative AI for Public Spaces. In *Proceedings of the 12th ACM Conference on Creativity and Cognition*. ACM, San Diego, CA, USA.

[51] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 ACM Conference on Human Factors in Computing Systems (CHI 2020)*. ACM, Honolulu, Hawaii. https://doi.org/10.1145/3313831.3376727

[52] Duri Long, Aadarsh Padiyath, Anthony Teachey, and Brian Magerko. 2021. The Role of Collaboration, Creativity, and Embodiment in AI Learning Experiences. In *Proceedings of the 13th ACM Conference on Creativity and Cognition*. 1–10.

[53] Leilah Lyons, Brian Slattery, Priscilla Jimenez, Brenda Lopez, and Tom Moher. 2012. Don't forget about the sweat: effortful embodied interaction in support of learning. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*. ACM, 77–84.

[54] Lívia S Marques, Christiane Gresse von Wangenheim, and Jean CR Hauck. 2020. Teaching Machine Learning in School: A Systematic Mapping of the State of the Art. *Informatics in Education* 19, 2 (2020), 283–321. Publisher: Vilnius University Institute of Data Science and Digital Technologies.

[55] Neil Mercer. 2008. The seeds of time: Why classroom dialogue needs a temporal analysis. *The journal of the learning sciences* 17, 1 (2008), 33–59. Publisher: Taylor & Francis.

[56] Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. 2021. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law* 29, 2 (2021), 111–147. Publisher: Springer.

[57] Joan Mora-Guiard and Narcis Pares. 2014. Child as the measure of all things: the body as a referent in designing a museum exhibit to understand the nanoscale. In *Proceedings of the 2014 conference on Interaction design and children*. ACM, 27–36.

[58] IT Chan Mow. 2008. Issues and difficulties in teaching novice computer programming. In *Innovative techniques in instruction technology, e-learning, e-assessment, and education*. Springer, 199–204.

[59] Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel KW Chu, and Maggie Qiao Shen. 2021. Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* (2021), 100041. Publisher: Elsevier.

[60] Safiya Umoja Noble. 2018. *Algorithms of oppression*. New York University Press.

[61] Marina Papastergiou. 2008. Are computer science and information technology still masculine fields? High school students' perceptions and career choices. *Computers and Education* 51, 2 (2008), 594–608.

[62] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[63] 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*. 1–12.

[64] Stephan Raaijmakers. 2019. Artificial intelligence for law enforcement: challenges and opportunities. *IEEE Security & Privacy* 17, 5 (2019), 74–77. Publisher: IEEE.

[65] 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*. 67–78.

[66] Jessica Roberts, Amartya Banerjee, Annette Hong, Steven McGee, Michael Horn, and Matt Matcuk. 2018. Digital Exhibit Labels in Museums: Promoting Visitor Engagement with Cultural Artifacts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 623.

[67] Jessica Roberts and Leilah Lyons. 2017. Scoring Qualitative Informal Learning Dialogue: The SQuILD Method for Measuring Museum Learning Talk. Philadelphia, PA: International Society of the Learning Sciences.

[68] Jessica Roberts and Leilah Lyons. 2017. The value of learning talk: applying a novel dialogue scoring method to inform interaction design in an open-ended, embodied museum exhibit. *International Journal of Computer-Supported Collaborative Learning* 12, 4 (2017), 343–376. Publisher: Springer.

[69] Juan David Rodríguez-García, Jesús Moreno-León, Marcos Román-González, and Gregorio Robles. 2021. Evaluation of an Online Intervention to Teach Artificial Intelligence with LearningML to 10-16-Year-Old Students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 177–183. https://doi.org/10.1145/3408877.3432393 event-place: Virtual Event, USA.

[70] Mark Rosin, Jen Wong, Kari O'Connell, Martin Storksdieck, and Brianna Keys. 2021. Guerilla Science: Mixing science with art, music and play in unusual settings. *Leonardo* 54, 2 (2021), 191–195.

[71] Cinthia Ruiz and Manuela Quaresma. 2021. Explainable AI for Entertainment: Issues on Video on Demand Platforms. In *Congress of the International Ergonomics Association*. Springer, 699–707.

[72] Marie-Monique Schaper, Maria Santos, Laura Malinverni, and Narcis Pares. 2017. Towards the design of a virtual heritage experience based on the world-as-support interaction paradigm. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2034–2041.

[73] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele Von Voigt. 2012. Big data privacy issues in public social media. In *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*. IEEE, 1–6.

[74] Scott S Snibbe and Hayes S Raffle. 2009. Social immersive media: pursuing best practices for multi-user interactive camera/projector exhibits. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1447–1456.

[75] Reed Stevens and L Takeuchi. 2011. The new coviewing: Designing for learning through joint media engagement. (2011). Publisher: The Joan Ganz Cooney Center.

[76] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R Cooperstock. 2019. Can You Teach Me To Machine Learn?. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, 948–954.

[77] Peter Svenmarck, Linus Luotsinen, Mattias Nilsson, and Johan Schubert. 2018. Possibilities and challenges for artificial intelligence in military applications. In *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting*. Neuilly-sur-Seine France, 1–16.

[78] David Touretzky and Christina Gardner-McCune. 2022. Artificial Intelligence Thinking in K-12. In *Computational Thinking in K-12: Artificial Intelligence Literacy and Physical Computing*. MIT Press.

[79] 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 2019 Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence.

[80] Jessica Van Brummelen, Tommy Heng, and Viktoriya Tabunshchyk. 2021. Teaching Tech to Talk: K-12 Conversational Artificial Intelligence Literacy Curriculum and Development Tools. In *2021 AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*.

[81] Jessica Van Brummelen, Judy Hanwen Shen, and Evan W Patton. 2019. The Popstar, the Poet, and the Grinch: Relating Artificial Intelligence to the Computational Thinking Framework with Block-based Coding. In *Proceedings of International Conference on Computational Thinking Education*, Vol. 3. 160–161.

[82] J. Van Brummelen, C. Yeo, and K. Weng. 2020. Learning to Program Conversationally: A Conversational Agent to Further Democratize Programming. In *INTED2020 Proceedings (14th International Technology, Education and Development Conference)*. IATED, 8950–8951. https://doi.org/10.21125/inted.2020.2446 ISSN: 2340-1079 event-place: Valencia, Spain.

[83] Christiane Gresse von Wangenheim, Jean CR Hauck, Fernando S Pacheco, and Matheus F Bertonceli Bueno. 2021. Visual tools for teaching machine learning in K-12: A ten-year systematic mapping. *Education and Information Technologies* (2021), 1–46. Publisher: Springer.

[84] Xiaoyu Wan, Xiaofei Zhou, Zaiqiao Ye, Chase K Mortensen, and Zhen Bai. 2020. SmileyCluster: supporting accessible machine learning in K-12 scientific discovery. In *Proceedings of the Interaction Design and Children Conference*. 23–35.

[85] Trevor Watkins. 2020. Cosmology of artificial intelligence project: Libraries, makerspaces, community and AI literacy. *AI Matters* 5, 4 (2020), 14–17. Publisher: ACM New York, NY, USA.

[86] Henry M Wellman, David Cross, and Julanne Watson. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development* 72, 3 (2001), 655–684.

[87] Linda L. Werner, Brian Hanks, and Charlie McDowell. 2004. Pair-programming helps female computer science students. *Journal on Educational Resources in Computing (JERIC)* 4, 1 (2004), 4. http://dl.acm.org/citation.cfm?id=1060075

[88] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating systems: Gender, Race, and Power in AI*. Technical Report. AI Now.

[89] Gary KW Wong, Xiaojuan Ma, Pierre Dillenbourg, and John Huan. 2020. Broadening artificial intelligence education in K-12: where to start? *ACM Inroads* 11, 1 (2020), 20–29. Publisher: ACM New York, NY, USA.

[90] Lynette Yarger, Fay Cobb Payton, and Bikalpa Neupane. 2019. Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review* (2019). Publisher: Emerald Publishing Limited.

[91] Nicola Yelland and Jennifer Masters. 2007. Rethinking scaffolding in the information age. *Computers & Education* 48, 3 (2007), 362–382. Publisher: Elsevier.

[92] Xiaofei Zhou, Jessica Van Brummelen, and Phoebe Lin. 2020. Designing AI Learning Experiences for K-12: Emerging Works, Future Opportunities and a Design Framework. *arXiv preprint arXiv:2009.10228* (2020).

[93] Heather Toomey Zimmerman, Suzanne Reeve, and Philip Bell. 2010. Family sense-making practices in science center conversations. *Science Education* 94, 3 (2010), 478–505. Publisher: Wiley Online Library.

[94] 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*. ACM, 121–132.