Rabinowitz *et al.*
International Journal of STEM Education (2025) 12:23
https://doi.org/10.1186/s40594-025-00543-5

International Journal of STEM Education

RESEARCH Open Access

Study of an effective machine learning-integrated science curriculum for high school youth in an informal learning setting

Gabrielle Rabinowitz[1*], Katherine S. Moore[2], Safinah Ali[3], Mark Weckel[1], Irene Lee[2], Preeti Gupta[1] and Rachel Chaffee[1]

Abstract

Purpose This study evaluates the effectiveness of a machine learning (ML) integrated science curriculum implemented within the Science Research Mentorship Program (SRMP) for high school youth at the American Museum of Natural History (AMNH) over 2 years. The 4-week curriculum focused on ML knowledge gain, skill development, and self-efficacy, particularly for under-represented youth in STEM.

Background ML is increasingly prevalent in STEM fields, making early exposure to ML methods and artificial intelligence (AI) literacy crucial for youth pursuing STEM careers. However, STEM fields, particularly those focused on AI research and development, suffer from a lack of diversity. Learning experiences that support the participation of under-represented groups in STEM and ML are essential to addressing this gap.

Results Participant learning was assessed through pre- and post-surveys measuring ML knowledge, skills, and self-efficacy. Results from the implementation of the curriculum show that participants gained understanding of ML knowledge and skills ($p < 0.001$, $d = 1.083$) and self-efficacy in learning ML concepts ($p = 0.004$, $d = 0.676$). On average, participants who identified as female and non-white showed greater learning gains than their white male peers (ML knowledge: $p < 0.001$, $d = 1.191$; self-efficacy: $p = 0.006$, $d = 0.631$), decreasing gaps in ML knowledge, skills, and self-efficacy identified in pre-survey scores.

Conclusions The ML-integrated curriculum effectively enhances students' understanding and confidence in ML concepts, especially for under-represented groups in STEM, and provides a model for future ML education initiatives in informal science settings. We suggest that policy makers and school leaders take into account that high school age youth can learn ML concepts through integrated curricula while maintaining an awareness that curriculum effectiveness varies across demographic groups.

Keywords Artificial Intelligence (AI), Machine learning, High school, Science-integrated curriculum, AI literacy

Introduction

As the use of Artificial Intelligence (AI) continues to expand across professional, personal, and academic domains, there is a mounting need for students at the pre-college level to develop AI literacy (Casal-Otero et al., 2023). AI literacy is a set of competencies that enables the critical evaluation of AI technologies, effective collaboration with AI, and the use of AI as a tool (Long & Magerko, 2020). In addition to these competencies,

*Correspondence:
Gabrielle Rabinowitz
grabinowitz@amnh.org
[1] American Museum of Natural History, 200 Central Park W, New York, NY 10024, USA
[2] MIT STEP Lab, 600 Technology Square, NE49 - 3021, Cambridge, MA 02139, USA
[3] Massachusetts Institute of Technology, MIT Media Lab, 75 Amherst St, Cambridge, MA 02143, USA

self-efficacy has been identified as a precursor to AI tool use in undergraduate students (Falebita & Kok, 2024), and in engagement among pre-service teachers (Ayanwale et al., 2024). The successful development of AI literacy is an interdisciplinary endeavor, incorporating data science, computational thinking and disciplinary knowledge (Ng et al., 2021). One essential sub-domain of AI is machine learning (ML). ML involves the use of algorithms to"learn"or the use of models to draw inferences from large data sets and make accurate predictions without requiring explicit instructions. A diverse set of ML methods have been developed to facilitate this work (Mahesh, 2020). The centrality of ML in AI innovations makes it a reasonable topic for K- 12 AI instruction, especially at the high school level (Lao, 2020; Sanusi et al., 2023; Touretzky et al., 2023). This paper reports the outcomes of an ML-integrated science curriculum for high school youth delivered in an informal learning environment.

A recent review of K- 12 AI education found that the integration of AI methods into existing curriculum is a prerequisite for success (Lee & Kwon, 2024). While ML can be introduced to youth with a variety of framings, scientific inquiry (including such practices as posing questions, designing and conducting investigations, and analyzing and interpreting data) is a particularly good fit (Zhou et al., 2021). ML knowledge and skills are increasingly used by non-computer science (CS) experts in a variety of scientific fields, including the Natural Sciences, that involve the organization and processing of large and complex data sets (e.g., Desjardins-Proulx et al., 2019; Longo et al., 2019). ML methods can handle complex non-linear relationships among hundreds of variables and are especially well-suited for making accurate predictions from complex data without explicit programming (Kashinath et al., 2021; Mahesh, 2020). Data sets with such complexity are commonplace in STEM fields, i.e., astrophysics (Longo et al., 2019) and genomics (Eraslan et al., 2019). Preliminary approaches to integrating ML in high school science curricula have framed ML methods as a set of tools for extracting insights from large data sets (Lee & Zhang, 2022), suggesting that high school students can learn ML in the context of STEM learning, and there are several existing efforts in advancing K- 12 ML literacy, including designing curriculum (Lee et al., 2021), standards (Touretzky et al., 2023), and tools (Zhang et al., 2024a, 2024b).

Yet few of these literacy efforts focus on integrating ML and scientific research, despite the increasing prevalence of AI in scientific discovery. To prepare high school youth for conducting scientific research in the natural sciences, it is essential for them to gain the knowledge and self-efficacy around using ML methods in the context of natural science, including forming meaningful scientific questions, analyzing natural science data sets, making scientific predictions, and understanding bias in ML.

One issue that plagues ML and STEM fields is the lack of diversity, with women and members of ethnic minority groups at risk of exclusion (West et al., 2019; Zhang & Barnett, 2015). In the case of ML, additional problems include the creation of biased algorithms and biased data sets (Leavy, 2018) that automate discriminatory practices in science and society (Shrestha & Yang, 2019). The lack of representation introduces barriers for youth with historically marginalized identities to envision a future for themselves within STEM or ML careers (Zhang & Barnett, 2015). We posit that informal learning environments—such as museums and science centers—provide a unique and strategic opportunity for access to and engagement with STEM learning (Adams & Gupta, 2013), including ML and AI concepts. Informal science learning environments are incubators for advancing our understanding of how to learn with emerging technologies free from state standards that can be applied to the formalized K- 12 system (Chaffee et al., 2021). Equally important, in informal learning settings students are able to engage in science as a part of their social interactions, increasing their comfort with STEM subjects while fostering community (Adams et al., 2012).

For 15 years, the Science Research Mentorship Program (SRMP)—a flagship STEM workforce development program of the American Museum of Natural History (AMNH)—has been increasing access to science fields and careers for New York City high school students from historically marginalized communities by providing authentic science research opportunities and meaningful mentorship in a museum setting. SRMP consists of a Summer Institute, where youth are introduced to research skills and begin participating in a community of practice (Wenger, 1999), consisting of scientists, peers, alumni, and program staff. The youth participating in SRMP then work in small groups to conduct a year of mentored scientific research on a particular natural science topic under the guidance of a scientist mentor.

We developed a new science-integrated ML curriculum for the SRMP Summer Institute, which introduces high school students to key ML skills and concepts in the context of scientific inquiry, called SRMPmachine. This new SRMP Summer Institute (hereafter referred to as the Institute) provides opportunities for high school age youth to develop AI literacy as they (1) evaluate and apply ML methods in a data science and scientific context, (2) use, modify, create, and evaluate ML models using both code and interactive tools, and (3) identify and propose mitigation strategies for both data bias and societal bias in the ML pipeline.

While several K- 12 AI curricula have been developed to introduce technical, ethical, and career aspects of AI and ML to middle school (Zhang et al., 2023) and high school students (Kaspersen et al., 2022; Sanusi et al., 2023), none of the existing ML curricula emphasize scientific inquiry with the benefits of an informal learning environment. In this paper, we present the theoretical framework that informed the design of the Institute and identify key characteristics of the curriculum that best supported youth learning. We share results of quantitative analysis measuring learning outcomes as well as qualitative analysis highlighting aspects of the curriculum that influenced students' ML learning.

Literature review

K- 12 AI/ML education and curriculum

Expanding applications of AI and ML (AI/ML) in professional fields have made AI literacy increasingly important for youth preparing to enter the future workforce (Touretzky et al., 2019). In response to this need, several K- 12 AI/ML curricula have emerged in recent years. For example, Lee et al. (2021) developed a middle school AI curriculum focusing on students' AI knowledge, attitudes, and career interests around AI. They primarily engaged youth that are under-represented in STEM, including non-male BIPOC youth. Through interactive activities, games, and ethical inquiries, students gained a deeper understanding of AI knowledge, applications and skills. Similarly, several other K- 12 AI curricula focus on developing students' personal competencies, including their competence, attitude (Su & Zhong, 2022), motivation towards AI learning (Chiu et al., 2023), creativity (Ali et al., 2019), career interests (Zhang et al., 2023), and ethical implications of AI (Williams et al., 2023). Most of the AI/ML curricula for high school youth emphasize ML basics and neural networks (Marques et al., 2020). Our work seeks to contribute to this burgeoning field an AI/ML curriculum that integrates principles of equity/ethics and STEM education for high school youth.

Equity in K- 12 AI/ML education

Many K- 12 AI/ML curricula have also specifically focused on the ethical implications and justice frameworks around AI. Some have argued that AI education cannot happen without consideration of ethics and equity (Walsh et al., 2022). While the field agrees on the importance of ethics and equity in K- 12 AI education, the understanding of how to design AI curriculum for equity and social justice has yet to settle on a particular method. In recent research in AI education, ethics and equity have been considered *(a)* a topic of instruction, *(b)* a pedagogical method, *(c)* a co-development method, or all of the above. For example, Lee and Zhang (2022) integrate ethic related topics into their AI curriculum to motivate learning and enable learners to see the relevance of AI in their everyday lives (Saltz et al., 2019). Yet, they also design their curriculum to *be* equitable by increasing accessibility through hands-on experimentation and participatory simulation (Squire & Klopfer, 2007). Other AI curriculum developers have designed curricula to be equitable through project-based learning (e.g., Aliabadi et al., 2022; Walsh et al., 2022). Centering algorithmic justice, Walker et al. (2022) leveraged project-based learning to enable students to mitigate systemic oppression through data activism. Their data science program taught African American students how to leverage their technical skills in service of projects that are more authentic and relevant to them. Centering students as co-developers of content is a practice used by Lin et al. (2022) who co-constructed their AI curriculum with traditionally marginalized students. Similarly, Long and Magerko (2022) conducted codesign activities with family groups to ensure their definitions of AI literacy were useful to learners in their everyday lives. These approaches to equity in AI education are interwoven in Ali et al. (2021), who used interactive simulation and experiential learning to help a diverse group of students understand topics such as the ethical and societal implications of generative AI that may affect them, such as the spread of misinformation through Deepfakes and relate these topics to their own lives. These learning interventions not only led to significant shifts in students' knowledge and applications of AI, but significant shifts in their attitudes towards AI (Lee et al., 2021; Williams et al., 2023; Zhang et al., 2023). This work, which frames equity and ethics as a topic, pedagogy, and method, informed the development of our curriculum.

AI in K- 12 STEM education

In spite of the value of AI/ML in STEM, emphasis has primarily been placed on providing professional development for K- 12 STEM teachers in AI/ML and supporting the integration of AI tools into STEM classes. For example, Zhou et al. (2021) introduced elementary, middle, and high school math and science teachers to a K-means clustering learning tool and supported the teachers in developing lesson plans on topics, such as flood resistance, evolution of biological characteristics, and heart disease risk factors. The ML4STEM program expanded the reach of this tool and involved K- 12 STEM educators in the co-development of additional ML-integrated STEM instruction (Tang et al., 2023). In a similar vein, Lee and Perret (2022) developed an AI and Data Science professional development program to provide high school math and science teachers with both AI content knowledge and an understanding of bias in AI. This program also aimed to support STEM teachers in the

integration of AI methods into their classrooms. Most recently, Park et al. (2023) provided professional development to three science teachers in secondary schools in Singapore to introduce a new lesson package that introduced their students to ML methods. In the informal STEM learning space, the Florida Museum recently launched Shark AI, a program which prepares Florida middle school teachers to use ML to identify fossil shark teeth. Teachers work with museum scientists and engineers to create their own lesson plans using these resources (Waisome et al., 2023). The greatest distinction between these efforts and our work is the target audience. We designed and delivered a science-integrated curriculum for high school youth themselves rather than in the form of professional development for K- 12 STEM teachers. The emphasis on research methods to prepare youth for a yearlong mentored research experience is also a unique aspect of our curriculum.

Theoretical framework

We designed the Institute curriculum in alignment with several theoretical frameworks, including the AI literacy competencies developed by Long and Magerko (2020), and the ML education Framework developed by Natalie Lao (2020). Lao draws from four established theories of learning and incorporates them into the Use–Modify–Create learning progression (Lee et al., 2011) to assemble a scaffold for transforming learners from passive users to an active role as tinkerers and makers of ML tools, emphasizing knowledge, skills, and attitudes towards ML. As an additional layer to the Use–Modify–Create progression, learners must know how to use ML models ethically, as well as how to critically evaluate these models and the algorithms used to generate them (Long & Magerko, 2020). Finally, the addition of "Choose" after the Use–Modify–Create progression, where students must select suitable machine learning models, has been shown to help students deepen their understanding of ML concepts (Martin et al., 2020).

To support the implementation of the ML Education Framework in a high school ML curriculum, Lao introduces several important considerations for teaching ML to high school students, many of whom will be introduced to the topic for the first time. In particular, enactive mastery, where learners successfully perform tasks related to the learning objectives, is considered necessary for the development of *self-efficacy* (Lao, 2020). The strategies found to effectively support high school students in reaching a state of enactive mastery in ML as detailed in Lao (2020) correspond to the key design considerations (DCs) for teaching AI Literacy identified by Long and Magerko (2020). As shown below, we aligned the Long and Magerko (2020) AI Competencies with the Lao ML Education Framework (2020) to form learning objectives for the Institute curriculum (Table 1).

Participation in informal science education (ISE), which entails voluntary participation in science learning outside of school, is associated with increased interest in science and STEM professions (Bell et al., 2009). However, not all people experience these benefits equally and institutions such as science museums have the responsibility of explicitly supporting the participation and inclusion of non-white communities (Dawson, 2014). To support the development and implementation of a curriculum for the Institute that addresses these concerns, we applied the lenses of *Shifting Narratives* and *Authority Sharing* from the YESTEM Core Equitable Practices (Archer et al., 2022; YESTEM Project Team, 2021). This equity-informed *identities-in-practice* framework centers students' developing identities in the context of science communities of practice (Barton et al., 2008). The *Authority Sharing* lens in particular dovetails with the Community Cultural Wealth Theory consisting of aspirational, linguistic, familial, social, navigational, and resistant capital (Yosso, 2005). The *Shifting Narratives* lens reframes learning experiences from perspectives other than those dominant in the field of ML.

Curriculum overview

Youth enrolled in the Institute attend five 5-h sessions for the first, second, and fourth weeks (75 total contact hours) with the third week devoted to a 4-day excursion to a biological field station. In this section, we present a summary of the overall curriculum, detailing each of 4 weeks of instruction, and highlighting aspects of the curriculum which participants identified as particularly impactful in helping them understand ML concepts and apply ML knowledge and skills.

Week 1: In the first week, youth are introduced to the program and begin developing a community of practice. The instruction in this week is informed by the YESTEM Core Equitable Practice of *Shifting Narratives*. For example, a discussion about the nature of science is framed using the indigenous metaphor of the three sisters' garden. The ML pipeline (Fig. 1) is introduced alongside Teachable Machine, serving as a physically enacted learning experience of the process of AI/ML. In addition, youth explore statistical, human, and societal biases in data, algorithms, and predictions. During this week, youth also perform data cleaning, review basic statistics skills, and apply linear regression models to make predictions using data sets of New York City street trees and dragonfly wing morphology. These data sets exemplify bias in data collection (e.g., more street trees were measured in Manhattan than outer boroughs and more dragonflies were observed in North America than other

Table 1 Curriculum learning objectives, target curricular concepts, and example activities

Curriculum learning objectives	Curriculum concepts	Example activities
1. Understand the basics of what ML is and how ML applications are created	*ML General Concepts (K1):* Knowledge of the foundational concepts in ML and the pipeline involved in creating ML tools, i.e., the purpose of ML, how testing and training data are used, what distinguishes supervised ML from unsupervised ML	"ML or not?" quiz where students vote on whether a given technology uses ML Experiment with Teachable Machine to see if an algorithm can "learn to dance" Introduction to the "ML Pipeline", showing the steps of creating and using an ML model
2. Determine which problems can be solved using ML from a technical perspective and which problems should not be solved using ML from technical, ethical, and cultural perspectives	*ML Methods (K2):* Knowledge of the core technical concepts of ML methods commonly used in the Natural Sciences: decision trees, unsupervised clustering techniques (i.e., k-means clustering, principal component analysis (PCA)), and artificial neural networks	Use a linear regression model to predict unknown characteristics of NYC street trees Explore Python code which creates and applies a decision tree Compare two dragonfly habitats using PCA Simulate a neural network, acting as input, hidden, and output layers Identify the elements of the ML pipeline in Teachable Machine
	ML Societal Implications (K4): Knowledge of the benefits and harms that can occur due to bias in ML systems and the societal implications of the harms and benefits of ML	Use a flowchart to decide whether ML should be used to answer a given research question Case study: ML and policing with COMPAS Watch Coded Bias film and discuss implications
3. Describe how ML systems may come to be unpredictably biased throughout each step of the creation process and understand the harms that can occur due to bias	*ML Bias (K3)*: Knowledge of how ML systems may come to be biased at each step of the process involved in building, testing, and deploying ML tools	Identify potential sources of bias at each step of the ML Pipeline Discuss potential sources of bias in the Decision Tree colab Examine bias at each stage of analysis using Wallace Art selfie app exploration: sources of bias
4. Understand the core technical concepts of ML methods commonly used in the Natural Sciences—decision trees, unsupervised clustering techniques, deep neural networks—as well as discern when to use these methods Implement the range of ML methods (listed above)	*ML Methods (K2): See ML Methods definition above*	Create a decision tree to classify mushrooms as edible or poisonous Generate habitability plots for a dragonfly species using Wallace Read a scientific paper about field methods using deep learning Guest speaker presentation about the use of PCA in biological anthropology
	ML Planning (S1): Skills in the scoping and planning of ML projects, including the ability to determine which problems can be solved using ML from a technical perspective and which problems should and should not be solved using ML from technical, ethical, and cultural perspectives	Discussion: When should and shouldn't we use ML? With the aid of a flowchart, decide which ML method would be most appropriate to use (and why)
	ML Analysis (S2): Skills in the analysis of results from the range of ML methods, including the interpretation of ML results, comparing output to evaluation performance, and discerning how to best apply ML algorithms given their performance	Discussion about results of dragonfly wing decision tree Compare and evaluate different models of species distribution generated by Wallace
	ML Application (S3): Skills in the application of the range of ML methods in the context of scientific research, including the ability to use, modify, and create new ML algorithms	Data cleaning and summary statistics for a dragonfly wing data set in Google Sheets Data thinning of occurrences using Wallace Students share the results of their dragonfly species distribution model

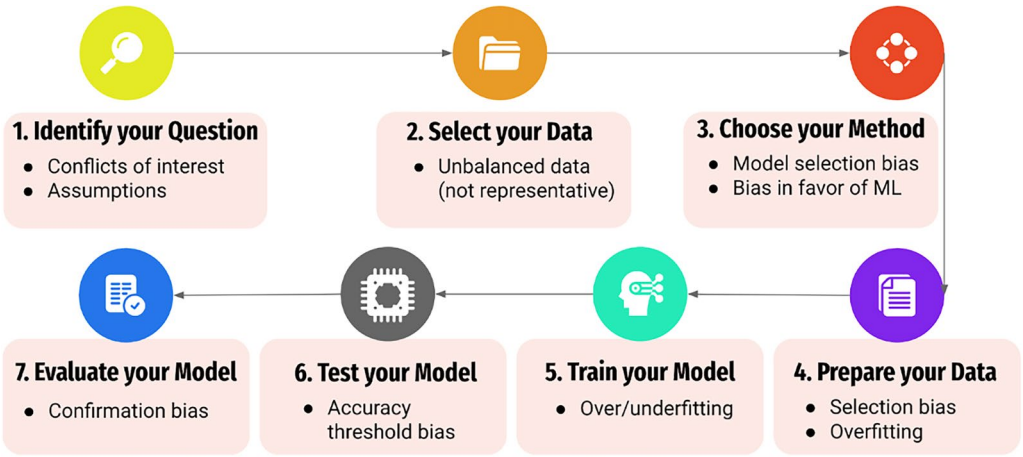

Fig. 1 Machine learning pipeline used in the institute. This framework was developed to help students integrate AI concepts and AI ethics. Examples of possible statistical and societal bias are included for each step. Participants are encouraged to identify the steps from the pipeline for each ML method introduced in the curriculum.

regions). The data is also tangible and interpretable, e.g., dragonfly wings can be held and measured through hands-on activities.

Week 2: Youth learn about two additional ML methods in the second week of the institute: decision trees and Principal component analysis (PCA). This week involves a much heavier lift in terms of assimilating and applying ML knowledge and skills. Youth progress from the "Use" stage to "Modify" and "Create" in the Use–Modify–Create progression. They have the opportunity to experiment with different algorithmic parameters for decision trees using Google Colab (Fig. 2) while choosing from among several different learning scaffolds and strategies.

Later, in the second week, youth generate their own species distribution model for a given dragonfly species using the Wallace platform, a GUI for an open source R library that gives science researchers access to large public biodiversity databases (Kass et al., 2023) (Fig. 3). Students also critically evaluate ML methods by assessing the performance of their models. This week also features tangible data sets, including animal skull morphology (PCA) and dragonfly wing measurements (decision trees).

Week 3: The third week of the Institute consists of a 4-day excursion to a biological field station. Youth have opportunities to socialize and build community in

Now it's time to split our data into a training and a testing dataset.

Bias Alert: It's important to create a random split to eliminate any clustering or sorting of the data. Run the code below to do so:

Split the data 50/50 into a training and testing set by typing the number 50 after "training_percentage =" below. Then run the code cell.

```
[ ]  #Get the features and labels from the data
     x = df.drop(['North_America'], axis=1)
     y = df['North_America']

     #Specify a 50% split: TYPE 50 AFTER THE EQUAL SIGN BELOW
     training_percentage =

     #Create the training and testing datasets
     X_train, X_test, Y_train, Y_test = train_test_split(x, y, train_size=training_percentage/100)
```

Fig. 2 Excerpt from the decision tree Colab notebook. Youth were invited to self-sort into one of three stations: self-paced, instructor-guided small group learning, and pair-programming to complete the Colab, giving them choice and agency in how they chose to encounter new material.

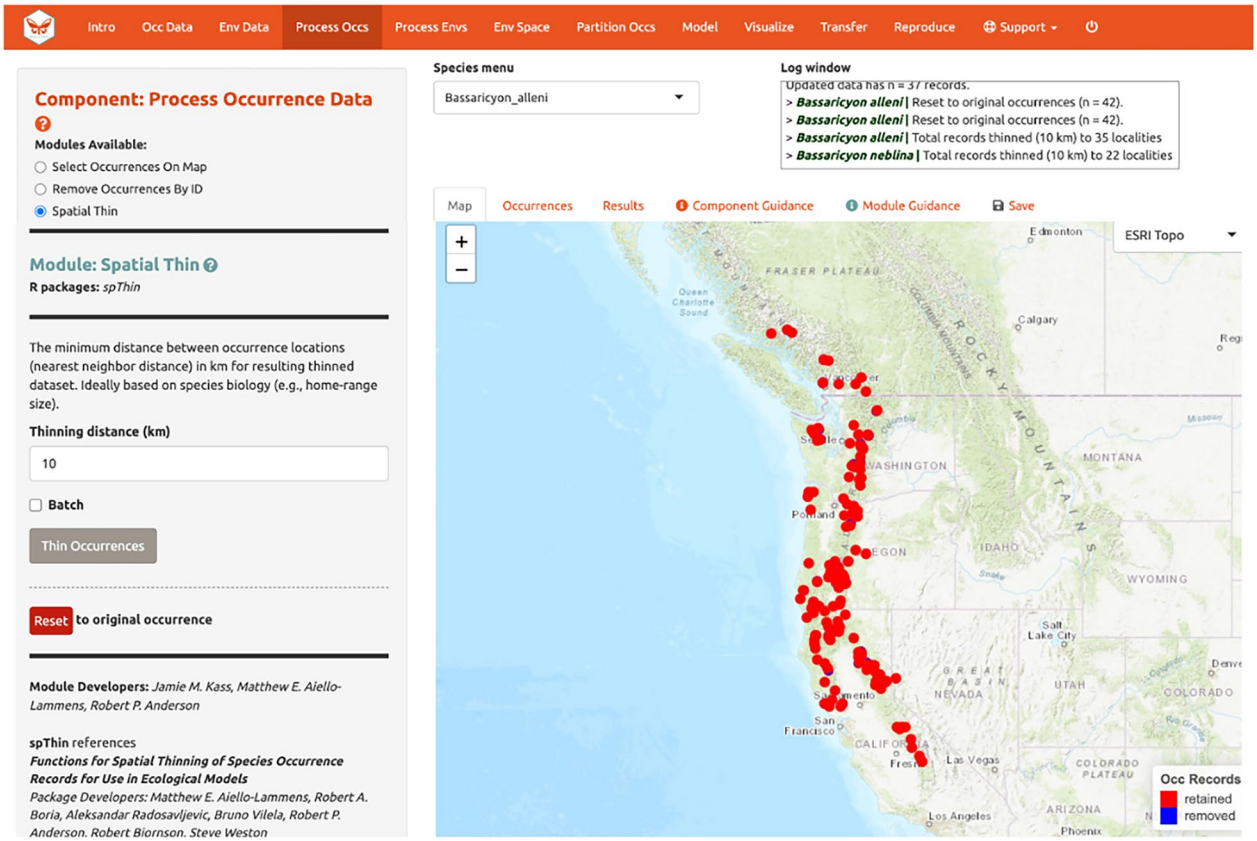

Fig. 3 Screenshot of the Wallace platform showing the data thinning process. Youth followed a modified version of the machine learning pipeline, with steps including obtaining data, processing data, creating and evaluating a set of possible models using maximum entropy analysis, and visualizing the final predictions.

addition to gaining hands-on field research experience. All youth join two field science activities: (1) setting up camera traps to collect photographs of wildlife and trapping and releasing turtles to record their size and (2) additional activities according to their interests and preferred level of physical intensity on topics including geology, insect identification, and archaeology field methods. All youth attend a short presentation about AI and scientific fieldwork.

Week 4: The fourth and final weeks of the institute introduce the last ML method of the curriculum: artificial neural networks (ANNs). Embodied interactions are again featured during the hands-on simulation of an ANN (adapted from the Artificial Neural Network Game reported in Zhang et al., 2023). The activity involves simulating the process of training an ANN involving feed-forward, evaluation, and back-propagation with youth acting in the roles of input, hidden, and output layers. Later in the week, youth compare and contrast the ML methods they learn about over the course of the program and create a flowchart to decide when to use each method to answer a scientific question (Fig. 4). This

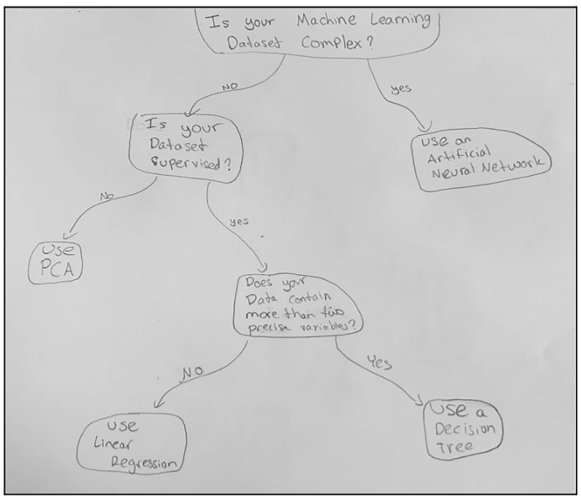

Fig. 4 Sample flowchart

week also features the YESTEM Core Equitable Practice of *Authority Sharing*, allowing students to choose their own data set to work with based on their subject matter

interests. The key theme of Bias in ML Systems is a primary focus this week. For example, artifact exploration (including apps and readings) shifts the responsibility of teaching about how AI is biased from students' personal experiences to more objective representations of the bias and potential for harm.

Methods

Participants

For this study, high school age youth were recruited to the Institute through the same application and enrollment procedures used by SRMP in years prior. Youth who had previously participated in AMNH programming or were students from partner schools or community based organizations were eligible. Participants (n =42) were majority 11 th grade (72%, n= 30) females (60%, n= 26), whose self-reported ethnicity was 28% white (12), 20% South Asian (9), 19% Black or African American (8), and 16% Other (7) including American Indian or Alaska Native, Middle Eastern or Persian, Native Hawaiian or other Pacific Islander. Sixteen percent identified as multiracial (7). Fourteen percent preferred not to share their race/ethnicity (6). Fifty-seven percent (24) reported being multilingual. Thirty-three percent identified as Hispanic (14).

A third of all participants (33%, n= 14) reported living in households with an annual income of $50,000 or less, which was below the poverty index for the region. Other participants reported living in households with a range of incomes: 25% (11) reported annual incomes ranging $50,000–$99,999; 4% (2) reported annual incomes ranging $200,000 and up. Twenty-eight percent (12) preferred not to share their household income.

Participants tended to report living in households in which at least one member of the household held either a 4-year degree (i.e., BA or BS) (19%, 8), a Master's degree (19%, 8) or had completed "some college" with no degree earned (17%, 7). However, a relatively large minority, of 14% (6) reported the highest degree in their household was a High School Diploma or Equivalent (e.g. GED). See Supplementary File 1 for detailed demographics.

For the purposes of evaluating the curricular impact on target outcomes among participants from underrepresented groups (URGs) in STEM, AI, and ML, we define URGs as any non-white or non-Asian male. URG is inclusive of participants who report a low socioeconomic status (i.e., below the poverty index for the region), multi-racial identity (i.e., two or more reported ethnic identities listed), or Hispanic cultural background, see Table 2 for details on URG participant demographics. Group membership overlaps allowing for intersectionality, e.g., a non-white, multiracial individual with low socio-economic status would be counted in three categories, but only once in the total population of URGs in STEM.

Table 2 Proportion of participant population from underrepresented groups (URGs) in STEM

Demographics	*n*	%
Non-White Male, Non-Asian Males	30	71.43
Low Socio-Economic Status	14	33.83
"Two or more" ethnic identities listed	8	19.052
Hispanic selected	14	33.38
Total Population of URGs in STEM	33	78.57

We also examine the curricular impacts on target learning outcomes on two URG sub-categories: gender and ethnicity. We define our gender categories as three self-identified groups: (1) female, (2) male, and (3) non-binary. We define our ethnic categories as comprising two groups: (1) white or Asian; (2) non-white and non-Asian, a category that is inclusive of multi-racial ethnicities and people who identify as Hispanic. We refer to this second ethnic category hereafter as non-Asian racial minority. We make this distinction to acknowledge that people of Asian descent, as well as people who identify as white, are over-represented in AI/ML, CS and STEM fields relative to the other ethnic populations (National Science Foundation, 2023). While people who identify as white make-up a near majority of the AI field, the representation of people of Asian descent in this field is likely to increase given recent patterns in the award of CS degrees and hiring in new AI jobs (Maslej et al., 2024).

Research questions

Analysis of curricular outcomes involved a mixed-methods approach (i.e., using survey, exit ticket, and interview data) to determine whether the curriculum effectively achieved its goals as a ML educational curricular intervention for high school age youth. To this end, evaluation focused on measuring: participant learning *outcomes* specific to the cognitive (i.e., knowledge and skills) and affective learning outcomes (i.e., self-efficacy in learning ML and interest in ML careers) that the curriculum was designed to impact the *alignment* of learning objectives with the curriculum's theoretical framework (see Table 1) via participant learning outcomes (RQ 1, see below); *content quality*, which is defined as the inclusivity and accessibility of the curricular content and activities such that all learners, regardless of their gender, ethnicity, or prior knowledge, are able to engage and learn from the materials (RQ 2); *aspects* of the curriculum that most impacted youth cognitive and affective learning outcomes (RQ 3).

From these goals, the following research questions were developed,

> RQ 1a. To what extent does participants' (i) ML knowledge and skills, (ii) self-efficacy in learning ML, and (iii) interest in ML careers change after experiencing the Institute curriculum?
> RQ 1b. How does participation in the Institute impact participants' understanding of target curriculum concepts?
> RQ 2. To what extent do participant responses vary across participant demographics, i.e., gender or ethnicity?
> RQ 3. What aspects of the curriculum do participants report as most impactful on their ML knowledge, skills, interest, and self-efficacy?

Instrumentation

At the beginning and conclusion of the Institute, the research team administered a pre- and post-survey. The survey comprised items designed to measure both cognitive and affective learning outcomes.

Cognitive outcomes were measured using the ML Concept Inventory (MLCI), which underwent preliminary validation over the course of this study. The MLCI (35 items) was found to have good internal consistency (Cronbach's alpha $=0.763$). Content validity was established through a review by 9 ML education researchers, who confirmed that the items were relevant to important concepts in ML education and that item responses would provide evidence of ML knowledge and skills. Agreement was established using quantitative content validity analysis methods established by Zamanzadeh et al. (2015). This analysis was followed by a series of semi-structured interviews in which each panelist was individually interviewed about their item ratings as a form of quality control. Criterion validity was assessed by correlating scores on the MLCI scores with scores on the validated *AI Concept Inventory* (Zhang et al., 2024a, 2024b). The correlation coefficient was significant ($p < 0.01$), indicating that the MLCI provides a valid measure of ML knowledge and skills. However, it is important to note that further analysis with a larger sample population is needed to establish the instruments' construct validity. Furthermore, additional work is needed to assess the validity of the MLCI in different populations.

Affective learning outcomes were measured, at the same time as the administration of the MLCI, using a survey of participants' Attitudes and Perceptions of AI (APAI). This instrument comprised a collection of scales from previously validated instruments, including the *AI/ML Career Interest* scale (17 items) (Cronbach $\alpha = 0.94$) (Zhang et al., 2024a, 2024b). The APAI also an adapted version of the Self-efficacy in Learning Medical AI (Cronbach α values ranged from 0.85 to 0.98) (Li et al., 2022) was used as our scale of *Self-Efficacy in Learning ML* (6 items). See Supplementary File 2 for details on each of the APAI items.

Exit tickets

In addition to pre- and post-surveys, exit tickets were administered as a formative assessment at the end of each day of instruction. Exit tickets used both closed and open items (see Supplementary File 3 for items). They were administered as part of a daily routine, through a link provided with the curricular materials. The Institute schedule allotted 10 min at the end of each day exit ticket completion. Anyone who finished early was asked to wait for the full time as a deterrence from rushing to finish. Only 2 of the 15 total exit tickets were examined for this paper. They were selected for analysis, because they included sets of items that prompted participants to apply their ML knowledge and skills using open-ended responses, which afforded qualitative triangulation of quantitative results.

Interviews

Immediately following the Institute (1–2 weeks afterwards), the research team conducted semi-structured retrospective interviews with 18 randomly sampled participants. Interviewees were demographically representative of youth who participated in the Institute. The protocol prompted interviewees to compare their understandings, perceptions, and attitudes towards ML before the Institute to those they had after the Institute. If interviewees described a change (positive or negative), they were asked to identify aspects of the Institute (i.e., activities, materials, interactions with people) that may have impacted that change (full protocol available in Supplementary File 4). Qualitative analysis of interviews used a reflexive thematic analysis approach with a deductive orientation (Braun & Clarke, 2022), which focused the analysis on impactful aspects of the curriculum.

Planned analysis

Analyses used quantitative methods to compare differences between MLCI and APAI pre- and post-survey responses from all participants. This was followed by an analysis using participant self-reported gender and ethnicity to determine whether there were differential impacts of the curriculum on cognitive and affective learning outcomes across demographic subgroups. Analysis then triangulated quantitative results with qualitative data from post-Institute interviews and open ended responses to select exit ticket items. Qualitative analysis

was used to identify aspects of the curriculum that were particularly impactful on target cognitive and affective learning outcomes.

Results

RQ 1a. To what extent does participant (i) ML knowledge and skills, (ii) self-efficacy in learning ML, and (iii) interest in ML careers change after experiencing the Summer Institute curriculum?

RQ 1a (i): ML knowledge and skills

Differences in MLCI responses showed large positive gains in participants' ML knowledge and skills after participating in the Institute as compared to pre-Institute responses. A paired t-survey indicated that post-survey responses were significantly higher than responses on the pre-survey, $t(41) = 8.869$, $p < 0.001$, $d = 1.083$). Figure 5 displays these results.

All interviewees shared that they had some degree of prior knowledge of ML before the Institute; yet, after the Institute all felt they had a greater understanding of the various ways that ML methods could be used in scientific research and in everyday life. Interviewee descriptions of their prior knowledge revealed a wide range of understanding: from awareness that AI existed to some knowledge of ML methods used in everyday technologies. For example, Elizabeth (all interviewee names are pseudonyms), explained that before the Institute, she didn't know that AI used ML. She understood that social media apps, i.e., TikTok, Instagram, and YouTube, use algorithms to track usage and recommend content based on a user's history; yet, she didn't know, "*how the database works or how the algorithms themselves work.*" After the Institute, she described her knowledge of ML saying,

> *Now I know about neural networks and PCA and decision trees and all these different things, which now I feel like if I'm looking at a certain platform or social media platform, I can identify which machine learning method they were using.*

Other interviewees shared that they had some prior knowledge of ML and specific ML techniques before the Institute, yet the Institute helped them better understand these terms and processes. For example, Daniel shared that he had heard about linear regression and neural networks before the Institute, but "didn't really know how they could be used for machine learning." Learning these techniques during the Institute, Daniel said, "really widened my view of AI and machine learning and how it can be used."

Another interviewee, Sophie, shared that, before the Institute, she was aware that ANNs were used to identify complex things, but she thought neural networks could only be used for one purpose. After the Institute, she realized that ML algorithms, including ANNs, could be used for different purposes in research. She said,

> *I was familiar with the artificial networking rhythms like using features to be able to identify complex things. But I didn't know that there was other types of algorithms that are correlated and were one with machine learning and that we can use different types of algorithms to be able to research different types of things. I just thought it was all in one.*

These examples show youth at different levels of prior ML knowledge, yet all three gained broader and deeper insights into ML concepts, especially ML methods and applications, after participating in the Institute.

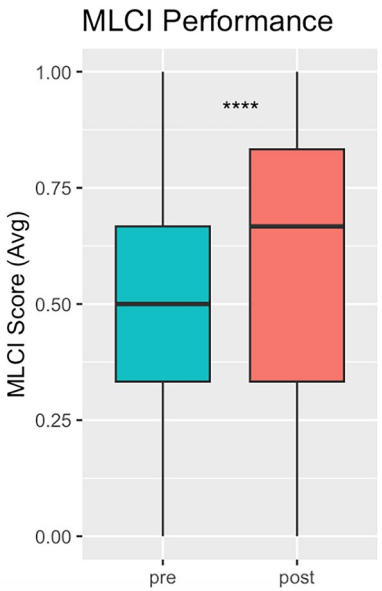

Fig. 5 Box-plot distribution of MLCI responses before the Institute (shown as "pre" on the left) and after the Institute (shown as "post" on the right)

RQ 1a (ii). Self-efficacy in learning ML

Along with the MLCI assessment, participants also completed the APAI. On average, results suggest small to moderate positive shifts in participants' perceptions and attitudes towards AI after experiencing the Institute, particularly in participants' self-efficacy towards learning ML (see Tables 7 and 8 for full details on APAI scales). A Wilcoxon signed-rank test, used to control for the non-normal distributions in the data, indicated that participants' self-efficacy was significantly higher after experiencing the Institute than before, $z = -2.88$, $p = 0.004$, $d = 0.676$ (before: $M = 3.785$, $SD = 0.637$; after: $M = 4.190$, $SD = 0.558$ (see Fig. 6).

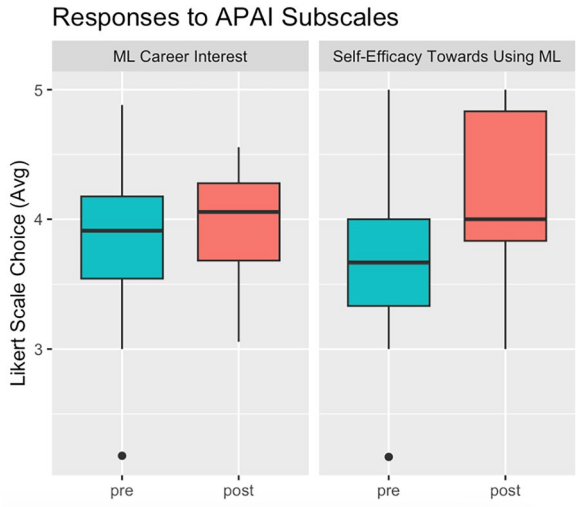

Fig. 6 Box-plot distribution of participant responses to the APAI ML Career Interest scale (LEFT) and Self-Efficacy scale (RIGHT) before the Institute (shown as "pre") and after the Institute (shown as "post").

In retrospective post-Institute interviews, 7 of the 18 youth interviewed shared that realizing how much they had learned about ML over the course of the Institute helped them feel more confident in their knowledge of ML. For example, Alexis explained that she had felt "insecure" about her ML knowledge and "super-lost" at the beginning of the Institute, but that "in the end, I think I saw a little bit improvement was when I did the Kahoot! in the end about machine learning, like I was understanding it slowly."

Another youth interviewed, Daniel, shared, "There were a lot of times where I was slightly confused and then something would happen, and it would be eye-opening when I realized what I was—I had come a long way, and I was actually learning a lot." Like Daniel, interviewees who described their confidence and self-efficacy in learning ML felt buoyed by the fact that they had learned about a complicated topic, yet these youth also felt that they had a lot more to learn. For example, Michael said, "I'm not super confident about it. I mean, I can probably summarize a decent amount, but I wouldn't be able to really teach someone exactly so well the differences and how you use different ones."

Worthy of note is that all interviewees, who spoke about their knowledge impacting their confidence and self-efficacy, had relatively low pre-survey scores and a relatively small change in their learning according to the MLCI. These interviewees scored lower than the median on the pre-survey (by 1 or 2 standard deviations). Their scores also showed relatively small change in their learning (1 standard deviation of change) between the pre- and the post-survey. Yet, in interviews they described a big change in their self-efficacy indicating a stronger positive shift in self-efficacy than our instruments were able to detect.

RQ 1a (iii): interest in ML careers

Results from the APAI also showed, on average, no significant change in participants' interests in ML careers after participating in the Institute when controlling for the non-normal distribution using a Wilcoxon signed-rank test, $z = -1.860$, $p = 0.063$. Yet, differences were approaching significance and mean scores across the seventeen 5-point Likert scale items suggest a small gain in interest (before: $M = 3.839$, $SD = 0.512$; after: $M = 4.057$, $SD = 0.466$) (see Fig. 6 for distributions).

Although our instrument did not detect a shift in interest in ML careers, interviews suggest an interesting pattern of change. Interviewees who described their level of interest in ML careers as unchanged between the beginning and the end of the Institute, shared that their *reasons* for their level of interest in ML careers after the Institute were different than before. For example, Morgan described a new interest in ML careers, because they wanted to use AI to help people,

> *Although I do I feel like it stayed the same, I feel like my reasons behind it now are different because at the beginning I was interested simply because I felt like I didn't know a lot about it. Now, I feel like I'm a bit well versed, I could say. But I still find myself interested in it because I understand now how AI is impacting our society and communities and it's something that I feel like I would like to be a part of the impact, hopefully, good impacts. I just want to honestly use or try to use AI to help people and I would like to learn more about it and about how it translates more to different work forces.*

Another interviewee, Alyssa, explained that before the Institute she had been interested in ML careers, because she was concerned about how ML/AI would impact the workforce. After the Institute, she was interested in ML careers, because she wanted to learn more about opportunities to use ML to advance her career,

> *I feel like 'interest' is such an ambiguous word where you can have positive interests and negative interests. I was kind of like, oh, is AI taking away people's-- I feel like in the media, you hear AI is taking people's jobs. So, I was interested in the jobs that AI is taking away. But now, I think I'm more interested in kind of how do jobs use AI on a daily basis? How do they use AI to benefit their work? How do they decide to rely on AI and stuff like that. Obviously, they use AI for a good reason, and it's helped them*

throughout their career. So, I'm very interested in those careers. I know lots of fields across the board that AI-- a lot of careers do use AI. So, yeah, I'm still very interested.

While our survey showed little change in interest, these vignettes show that the nature of participants' interest is evolving. These findings are further explored in the Discussion.

RQ 1b. How does participation in the Institute impact participants'understanding of target curriculum concepts?

MLCI scores suggest positive learning gains in 5 of the 7 target ML curriculum concepts (detailed in Table 1). On average, participant responses to the MLCI show moderate to large learning gains in knowledge of ML methods and societal implications of ML as well as skills in planning ML projects and analyzing ML results. Differences in responses to pre- and post-surveys of knowledge of ML bias and the skill of applying ML in scientific research were not statistically significant. These results are explored in the Discussion. Figure 7 displays the distributions of participants' average MLCI scores by curricular concept. Table 3 shows the results from Wilcoxon signed-rank tests comparing pre- and post-survey responses by curriculum concept. In the subsequent section, we triangulate the non-significant findings with qualitative data from exit ticket responses and interviews.

Knowledge of ML bias

While the MLCI did not detect a shift in participants' knowledge of ML bias, participants' written responses to exit tickets show that, by the end of the Institute, participants were correctly applying knowledge of ML bias in their decisions on whether ML could and should be used for scientific research and in the real world. For example, on the final exit ticket, a majority of the respondents who mentioned the term "bias" (79% of the 24 respondents, $n = 19$), used detail to support and justify their argument to a degree that made it possible to detect an emerging

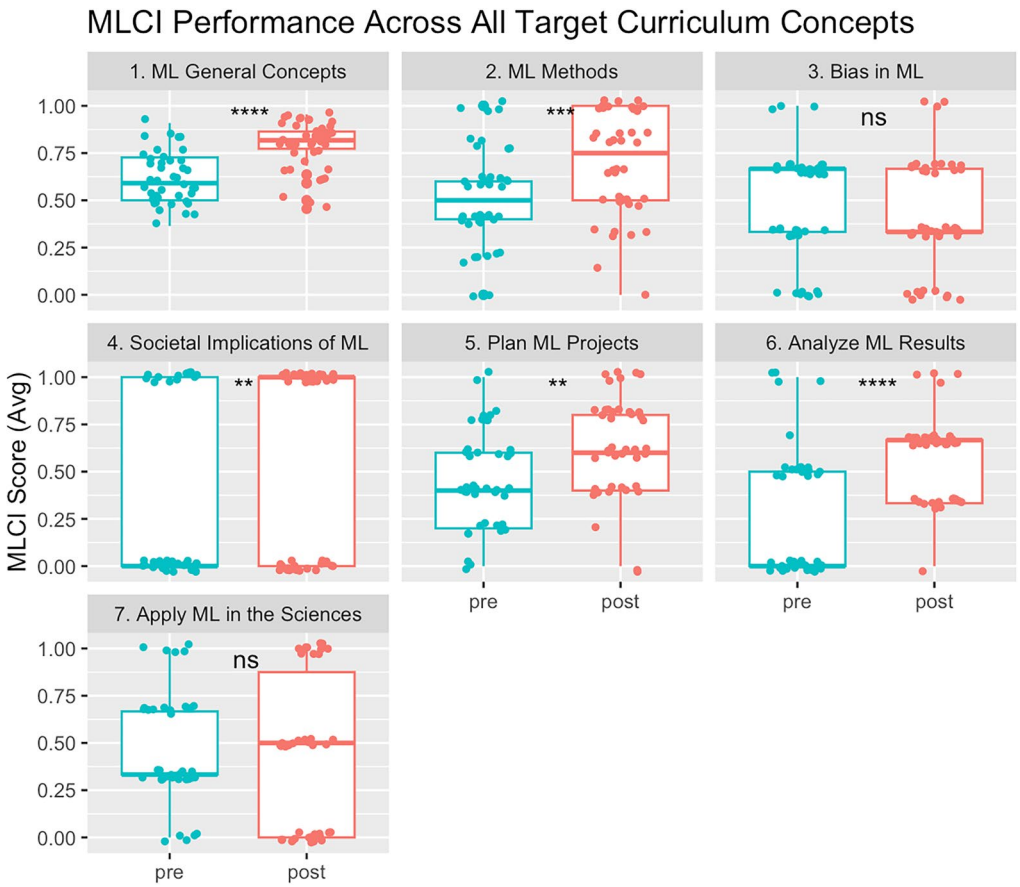

Fig. 7 Box-plots showing the distributions of participant responses to all but one of the MLCI scales. Responses after the Institute are shown as "post" on the right of each scale. Differences found to be significant are marked with asterisks. Non-significant differences are marked with ns. Dots represent individual data points and have been subjected to minimal random jitter along the x- and y-axis to improve visualization.

Table 3 Results from Wilcoxon signed-rank tests of pre- and post-survey responses by curriculum concept

Curriculum concepts	# Items	Test statistic	*p*	*d*
1. Knowledge of ML General Concepts	11	− 5.34	< 0.001*	1.289 *large*
2. Knowledge of ML Methods	12	− 3.6	< 0.001*	0.678 *moderate*
3. Knowledge of ML Bias	3	1.56	*ns*	*ns negligible*
4. Knowledge of Social Implications of ML	1	− 2.56	0.011*	0.596 *moderate*
5. Skill of Planning ML Projects	7	− 2.68	0.007*	0.663 *moderate*
6. Skill of Analyzing ML Results	3	− 4.57	< 0.001*	1.086 *large*
7. Skill of Applying ML in Scientific Research	3	0.607	*ns*	*ns negligible*

*Significant difference, *ns* no significant difference

understanding of bias in ML. For example, one respondent, Logan, argued that an ML tool should be deployed, because its health benefits outweighed the ethical risks of biased data. They wrote,

> *it should, because it could help identify people with Alzheimers in that community. because the data is randomized, there most likely will not be a sampling bias and it could eventually be scaled up and applied to ppl all over the world.*

Another respondent, Dylan, wrote, "There is place for bias to creep into this when you separate patients based on neighborhood." Another explained, "The training data may be too small, and the fact that they are considering demographic information—which includes race—has potential to introduce heavy bias."

In interviews, participants were asked to share their understanding of the "limitations" of ML (with no mention of the term "bias"). Twelve of the total 18 interviewees replied by describing their understanding of bias in ML. Several described the various types of biases they'd learned about in the Institute. For example, Katherine explained, "In the Institute, they taught me about some biases that can occur like human biases, societal biases, things like statistical biases." Similarly, Emma listed the different types of bias introduced during the Institute,

> *There's societal, human, computational, and they're all so different. And I thought bias was just like one category, but there's so many umbrellas underneath it that the types of bias can get sorted into. There's bias everywhere.*

Three interviewees not only correctly described bias in ML, but also mentioned the importance of minimizing bias in ML systems. For example, Dylan explained, "It's just not possible to create a bias-less system," concluding that "perfection," might not be attainable, but, "we can be more careful to create like strive towards that perfection." Another interviewee, Michael, shared that at the Institute, they learned some strategies that people can use "to balance it out and fix them," referring to biased ML models. Ashley described a specific strategy for minimizing bias that she experienced during the Institute, "We had to clean our dataset to make it more accurate."

Two interviewees explained that the Institute helped them understand both the limitations and affordances of various ML techniques and that these limitations and affordances can impact people differently. For example, Morgan explained that, "just because something is causing benefit to something, to another thing, it doesn't mean that it's causing the same benefit to a separate thing," later adding that with ML, "there are a lot of limitations, honestly, and we just need to think about which ones would be the ones that harm people the most."

Together, the exit tickets and interviews offer evidence of the emergence of a rich understanding of ML bias as a limitation, as well as its potential impact on society and its ethical implications. What's more, responses offer evidence of participant efforts to apply their knowledge of ML bias to make decisions as to whether ML can be used and whether it should be used from a technical and ethical perspective. Further research is needed to develop instrumentation that can detect this type of ML knowledge and skill.

Skill of applying ML in scientific research

The MLCI did not detect a shift in participants' skill of applying ML in scientific research, yet participants' exit ticket responses show that, by the end of the Institute, a majority of participants were correctly applying knowledge of ML concepts as they made decisions as to whether ML could or should be used in the context of scientific research. Exit tickets prompted all respondents to apply their ML knowledge in scenario-based contexts specific to scientific research. Questions prompted a 2 part response: 1) a binary (yes or no) response and 2) an open-ended response to justify the first binary selection (e.g., "Why or why not?"). With this design, binary responses could be scored for accuracy, while open ended responses could be qualitatively evaluated. Table 4

Table 4 Proportion of correct exit ticket responses by decision-making-based skill (a, b, c) of applied ML knowledge and skills in the sciences

Types of Exit Ticket Questions	# Items	Correct responses		Total responses
		n	%	*n*
(a) Could ML be used?	3	58	51	114
(b) Should ML be used?	4	106	70	152
(c) Which ML would be best?	4	111	73	152

displays the three types of questions used to prompt participants to apply their knowledge of ML in scientific research.

On average, exit ticket responses suggest a greater aptitude for deciding (b) and (c) than (a). Fifty-one percent of the total responses to (a) items were correct. Whereas 70% and 73% of the total responses to (b) and (c) items were correct. Correct responses to (a) type questions—*could* ML be used—tended to use examples of common limitations to ML to justify the selected answer. These responses drew from respondents' technical knowledge of ML to justify answers with facts using knowledge of ML limitations and of the ML pipeline. For example, several responses explained that ML could not be used, because the data set in the given scenario was too small or too homogeneous to train a generalizable model (see Table 5, for example, quotes marked i). Others explained that ML could not be used, because there was missing data (ii).

Correct responses to (b) type questions—should ML be used—tended to compare the benefits and harms of ML from an ethical and technical perspective. These responses seemed to draw from respondents' knowledge of the societal impacts on ML to justify answers through reasoning. For example, responses reasoned that an ML tool should not be used, because misclassification would have too great a harm (iii). Others reasoned that even correct predictions of a given tool, would not be useful given the cultural context in which the tool would be employed; thus it should not be used (iv).

Correct responses to (c) type questions—which ML method would be best—tended to describe characteristics or capabilities of ML methods that suited the given research project (v–vii). These responses drew from respondents' knowledge of the four different ML methods introduced in the curriculum.

Exit tickets responses show how participants applied their emerging ML knowledge in a series of scenario-based contexts specific to scientific research. These examples shed light on how participants use their knowledge of ML to make technical and ethical decisions. More research is needed to develop curricular activities and assessments to further develop these skills.

RQ 2. To what extent do participant responses vary across participant demographics?

Participant responses to both the MLCI and APAI varied by URG status (see section titled *Participants* above for our definition of URG and criteria for membership). Paired t tests show that participants from URGs experienced large gains in ML knowledge and skills, $t(32) = -7.800$, $p < 0.001$, $d = 1.191$; moderate gains in self-efficacy in learning ML, $t(32) = -2.934$, $p = 0.006$, $d = 0.631$; and small gains in their interest in ML careers, $t(32) = -3.376$, $p = 0.002$, $d = 0.483$ (see Table 6 for details). Their non-URG counterparts' responses also showed large gains in ML knowledge and skills, $t(6) = -3.710$, $p < 0.010$, $d = 1.215$; yet, no detectable changes in interest in ML careers or self-efficacy. It is important to note that the non-URG population is relatively small (n = 7),

Table 5 Sample statements from exit ticket open-ended responses to justify application of ML in scientific research

Type of Exit Ticket Question	Example justifications of correct responses
(a) Could ML be used?	(i) The data set is too small to train the model to identify the flowers. In addition the flowers that she wants to use for the data set are pressed, not in the natural form that they would be found in, so it would not make very good training data either (ii) this would not work because the data set she has only has the banana features without labels, and the other data set only has the labels and no Emily banana features
(b) Should ML be used?	(iii) Using the criminal activity and the demographic of a small population high school students can lead to the misidentifying of offenders. It can lead to innocent people being convicted just because they match the description of a high school student (iv) The data that is collected is biased, because it is only taken from the Met Gala events. The Met Gala fashion might not be what everyone wears in Germany. First of all, Met Gala clothing pieces are very expensive; therefore, not everyone will be wearing it
(c) Which ML would be best?	(v) PCA is the best option, because it is an unlabeled data set (vi) You would use an ANN, because you are dealing with more complex data, such as images and videos (vii) A decision tree would be able to find the biggest characteristic to split the data and further split the data into categories

Table 6 Performance on MLCI and APAI pre- and post-surveys by URG and non-URG demographic groups

Demographic groups	*n*	Pre-survey		Post-survey		*p*	*d*
		M	sd	*M*	sd		
ML Knowledge (MLCI)							
URG	33	0.512	0.127	0.671	0.140	< 0.001	1.191 *large*
Non-URG	7	0.684	0.128)	0.804	0.055	0.010	1.215 *large*
Self-efficacy (APAI)							
URG	33	3.818	0.666	4.207	0.562	0.006	0.630 *moderate*
Non-URG	7	3.762	0.568	4.048	0.583	*ns*	
Interest in ML Careers (APAI)							
URG	33	3.797	0.512	4.019	0.400	0.002	0.483 *small*
Non-URG	7	3.991	0.591	3.841	0.489	*ns*	

ns = not significant

making differences difficult to detect, see Fig. 8a for visualization of URG MLCI distributions.

Differences in MLCI pre- and post-survey scores between URG and non-URG members remained significantly different (pre: $p = 0.010$, $d = 1.358$; post: $p < 0.001$, $d = 1.022$), which suggests a gap in ML knowledge seen on the pre-survey persisted to the post-survey. Results from the APAI show no evidence of a gap between URG and non-URG members' interest in ML careers or self-efficacy in learning ML. These findings suggest the program significantly improved URG participant ML knowledge and skills, yet it did not close the knowledge gap observed between URG and non-URG members.

Further analysis identified several other patterns worthy of note. In the subsequent sections, we explore how participant responses varied across URG subgroups by reporting results from analysis of MLCI and APAI responses by gender (female, non-binary, and male) and ethnic/racial subgroups (white and Asian as compared to non-Asian minorities).

Learning outcomes

Results from paired *t* tests show moderate to large gains in ML knowledge and skills among females and participants who identified as a member of a non-Asian racial minority group: females, $t(21) = -4.709$, $p < 0.001$, $d = 0.731$; non-Asian racial minorities, $t(11) = -6.823$, $p < 0.001$, $d = 1.292$. Their male counterparts showed no significant learning gains; however, their white and Asian peers did, $t(23) = -5.324$, $p < 0.001$, $d = 0.925$ (see Table 6 for details).

On average, males scored significantly higher on the pre-survey and post-survey than their female counterparts: pre, $t(15.834) = -3.627$, $p = 0.002$, $d = 1.475$; post, $t(24.322) = -2.827$, $p = 0.009$, $d = 0.971$. In other words, on average, males started and ended the Institute with higher levels of ML knowledge than females. Yet, on average, male participant MLCI scores showed no significant difference between the pre- and the post-surveys (perhaps because of their initial highscores). This suggests that the observed gender gap in ML knowledge was decreased, but not completely closed, see Table 7 for details and Fig. 8b for a visualization of MLCI results by gender.

White and Asian participants started the Institute with relatively higher MLCI scores than their counterparts who are non-Asian racial minorities; however, this gap closed over the course of the Institute. While pre-survey scores of participants who identified as white or Asian were not significantly higher than participants who identified as non-Asian racial minorities, differences were approaching significance, $t(29.488) = -2.075$, $p = 0.090$, $d = 0.760$. What's more, differences between white/Asian and non-Asian racial minority post-survey scores were not significant, $t(19.057) = -0.646$, $p = 0.526$. This suggests that there may have been a knowledge gap between ethnic groups that closed, see Table 7 for details and Fig. 8c for a visualization of MLCI results by ethnicity.

Differences in participants' MLCI pre- and post-survey scores showed learning gains across a majority of ML concepts targeted by the curriculum. On average, participants who identified as female or as a member of a non-Asian racial minority group showed greater gains in these target concepts than their white-male and Asian-male counterparts. For example, female and non-Asian racial minority group members showed gains in their knowledge of *general ML concepts* (K1), *ML methods* (K2), and *societal implications of ML* (K4). They also demonstrated skill in *analyzing ML results* (S2) (see Supplementary File 5 for details).

While their white-male and Asian-male peers also showed gains in their *general ML knowledge* (K1) and

Table 7 Performance on MLCI by gender and ethnicity

Demographic groups	*n*	Pre-survey		Post-survey		*p*	*d*
		M	*(sd)*	*M*	*(sd)*		
Gender							
Female	26	0.517	(0.134)	0.671	(0.128)	< 0.001	0.731 *moderate*
Male	13	0.594	(0.147)	0.733	(0.139)	*ns*	
Non-binary	2	0.512	(0.165)	0.721	(0.263)	*ns*	
No answer	1	0.521		0.814		*ns*	
Ethnicity							
Non-Asian racial minority	12	0.500	(0.111)	0.678	(0.161)	< 0.001	1.292 *large*
White and Asian	24	0.579	(0.155)	0.713	(0.136)	< 0.001	0.925 *large*
No answer	6	0.504	(0.121)	0.663	(0.089)	*ns*	

non-Asian racial minority = non-white and non-Asian, a category that is inclusive of multi-racial ethnicities and people who identify as Hispanic

ns not significant

ability to *analyze ML results* (S2), they did not show significant gains in their knowledge of the *societal implications of ML* (K4). What's more, on average, males did not show significant gains in their knowledge of *ML methods* (K2), yet they showed gains in their ability to *plan ML projects* (S1) (as did white and Asian participants) to a greater degree than their peers who are female and non-Asian racial minorities (see Supplementary File 5 for details). These results are explored in "Discussion" section.

Affective learning outcomes

Changes in participant self-efficacy in learning ML also varied by gender and ethnicity. Differences in male participants' self-efficacy showed large gains after the Institute as compared to before the Institute, $t(12) = -2.30$, $p = 0.040$, $d = 1.046$. While differences in female participants' scores were non-significant, $t(25) = -1.946$, $p = 0.063$, yet differences were approaching significance, see Table 7 for details and Fig. 9a for distributions by gender.

Self-efficacy scores of participants who identified as non-Asian racial minorities showed moderate gains after the Institute as compared to before the Institute, $t(11) = -2.215$, $p = 0.049$, $d = 0.748$; as did their white and Asian counterparts, $t(11) = -2.215$, $p = 0.049$, $d = 0.748$. On average non-Asian racial minorities started and ended the Institute with levels of self-efficacy that were not significantly different from their white or Asian counterparts, see Table 8 for details and Fig. 9b for distributions by ethnicity.

Changes in participant interest in ML careers also varied by gender, but not significantly by ethnicity. On average, responses from participants, who identified as female or non-binary, to items on their interest in ML careers showed moderate positive gains after the Institute as compared to before the Institute: female, $t(25) = -3.4152$, $p = 0.002$, $d = 0.559$; non-binary, $t(1) = -1.8.143$, $p = 0.035$, $d = 0.784$. Whereas their male counterparts showed no significant changes in their interest. Differences in pre-survey scores by gender were not significant, which suggests that all participants' shared similar levels of interest in ML careers at the beginning of the Institute, see Table 9 for details and Fig. 9a for distributions by gender.

RQ 3. What aspects of the curriculum do participants report as most impactful on their ML knowledge, skills, interest, and self-efficacy?

Analysis of interviews led to the identification of 9 curricular activities that interviewees described as positively impacting their ML knowledge and skills, or their perceptions and attitudes towards ML/AI. Further analysis of these activities revealed patterns across activities that touched on similar themes. From these, 4 aspects of the curriculum emerged as most impactful on participant ML knowledge, perceptions, and attitudes towards AI. The interview protocol defines curricular "aspects" as, "parts of SRMP" such as "specific lessons and activities; hearing from Scientists, professionals and alumni; mentor meetings; and the Friday Advisories." See Supplementary File 4 for full protocol. Table 10 offers a summary of the 4 aspects.

Curriculum aspect 1: ML methods

Learning about various ML methods was impactful on participants' ML knowledge. In half of the interviews (9 of the 18), interviewees explained that they had known about AI and ML before the Institute, but only peripherally. During the Institute, interviewees gained a deeper understanding of ML methods and realized the wide

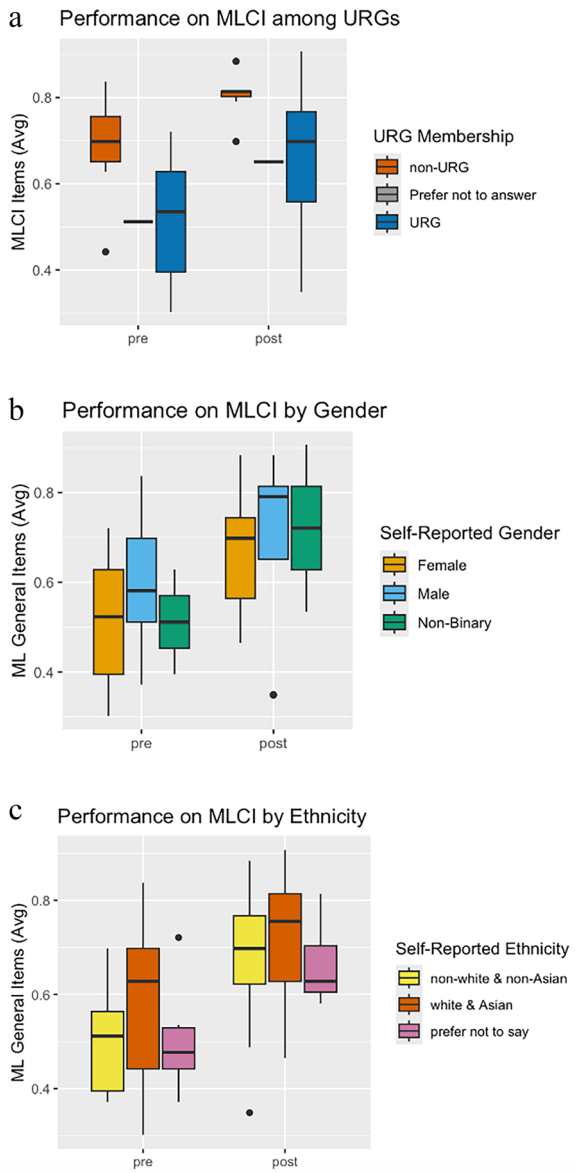

Fig. 8 **a**–**c** Boxplots display differences between pre-Institute and post-Institute MLCI scores by participant (n = 42) URG membership (TOP, **a**), self-reported gender (MIDDLE, **b**) and ethnicity (BOTTOM, **c**). A relatively large number of participants chose to not report their ethnicity; thus, these were grouped as a third category in (**a**) and (**c**).

variety of ways ML can be used in the sciences and in everyday life (see RQ 1a (i) for excerpts detailing these revelations). Interviewees mentioned several aspects of the Institute that helped them reach this level of understanding.

The participatory simulations of ML methods were repeatedly mentioned by interviewees. One interviewee, Emma, shared that it was the *Slice of ML* activity that helped her better understand ML methods,

> *It was just such an activity that really impacted what I thought of AI. Because the data – like the data biases, like the data biases, that really help me understand that AI isn't perfect, and it very much depends on data that we give it*

Several other interviewees described the un-plugged, hands-on ANN simulation as an activity that helped them understand what ML methods and how they work. Morgan shared, "I feel like that activity in and of itself helped me a lot to understand the ANN because not only did it tell us exactly what it did, but it also showed us like how to do it or how the process works." They appreciated that it was an interactive activity that helped them understand the differences between the ML methods.

Another frequently mentioned activity was the creation of the ML Flowchart, which interviewees described as an opportunity for them to practice using their knowledge of ML methods to make decisions. Alyssa shared that they felt creating their own ML flowchart helped them understand ML at a "better level." They explained,

> *The flowchart really helped. And then when we all made our own flowcharts that kind of really helped because it kind of-- I know when we were doing the exoplanet, when we split into groups in the last week when we were doing the exoplanets, and we were discussing which machine learning techniques can be used, I kind of always just referred to the flowchart I had made in my head and be like, oh, okay, this is what the question is asking for.*

For these interviewees, using the flowchart helped them apply their knowledge and make decisions about which ML method would be appropriate to use under a series of given circumstances.

Curriculum aspect 2: Wallace and dragonflies

Using ML in *Wallace* coupled with introductory activities examining a data set of dragonfly wing features had a positive impact on students' ML knowledge. Just under half of the interviewees (8 out of 18) mentioned that applying their ML knowledge in Wallace had a positive impact on both their ML knowledge. For example, one interviewee, Jacob, explained that his understanding of ML as a tool for prediction emerged from his work with Wallace to make predictions, using the dragonfly data set, as to where dragonfly habitats might be located. He explained, "the prediction aspect really came in when we did the dragonflies on Wallace, predict their habitat."

Several interviewees explained that it was the process of using the dragonfly data set along with Wallace as part of a larger project that was memorable and impactful for them. For example, Katherine explained

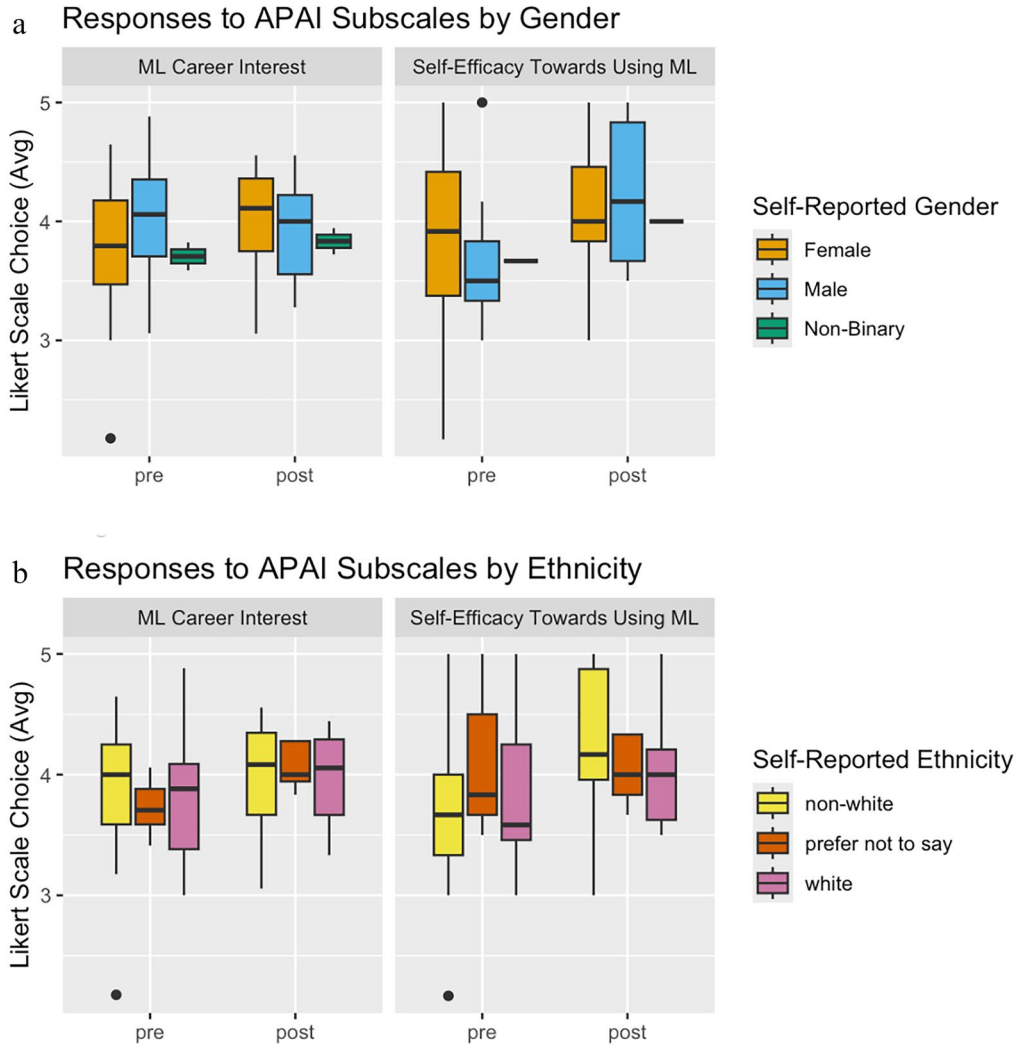

Fig. 9 **a, b** Boxplots display distributions of participant (n = 42) responses to the APAI ML Career Interest and Self-efficacy subscales by self-reported gender (TOP, **a**) and ethnicity (BOTTOM, **b**). A relatively large number of participants chose to not report their ethnicity; thus, these were grouped as a third category when reporting distributions based on ethnicity.

that her ML knowledge grew through the process of cleaning the dragonfly data set, creating graphs to visualize the data, and then using Wallace to make predictions, "like predicting if a species would be in an area in a couple of years or so based off the data." She added, "It was good to use an example that helped me understand what the machine learning is and what it does and how it functions." For these interviewees Wallace and the dragonfly data set together seemed a touchstone or exemplar case that they frequently returned to as they described their understanding of ML.

Curriculum aspect 3: data set exploration

Processing data sets to prepare them for ML analysis helped interviewees better understand concepts related to bias in ML. In several of the interviews (5 of 18), interviewees touched on how important interacting directly with the data was to the development of their understanding of ML and bias. For example, one interviewee, Michael, described the work of cleaning a data set to minimize bias, "doing data cleaning and that sort of stuff. Actually working with datasets gave some sort of idea of how data bias can be dealt with."

Table 8 Self-efficacy scores on the pre and post-survey for participants separated by gender and ethnicity

Demographic groups	*n*	Pre-survey		Post-survey		*p*	*d*
		M	(sd)	*M*	(sd)		
Gender							
Female	26	3.891	(0.70)	4.154	(0.55)	*ns*	
Male	13	3.641	(0.53)	4.244	(0.62)	0.040	1.046 *large*
Non-binary	2	3.667	(0.00)	4.000	(0.00)	*ns*	
No answer	1	3.170		4.830			
Ethnicity							
Non-Asian racial minorities	12	3.778	(0.760)	4.306	(0.647)	0.0488	0.748 *moderate*
White or Asian	24	3.750	(0.582)	4.139	(0.553)	0.020	0.685 *moderate*
No answer	6	3.945	(0.680)	4.167	(0.421)	*ns*	

non-Asian racial minority non-white and non-Asian, a category that is inclusive of multi-racial ethnicities and people who identify as Hispanic
ns not significant

Table 9 ML career interest scores on the pre and post-survey for participants separated by gender and ethnicity

Demographic groups	*n*	Pre-survey		Post-survey		*p*	*d*
		M	(sd)	*M*	(sd)		
Gender							
Female	26	3.760	(0.533)	4.028	(0.418)	0.002	0.559 (moderate)
Male	13	4.001	(0.503)	3.919	(0.429)	*ns*	
Non-binary	2	3.706	(0.167)	3.833	(0.157)	0.035	0.784 (moderate)
No answer	1	3.940		3.390			
Ethnicity							
Non-white and non-Asian	12	3.778	(0.692)	3.908	(0.460)	*ns*	
White or Asian	24	3.887	(0.469)	4.000	(0.426)	*ns*	
No answer	6	3.765	(0.241)	3.954	(0.333)	*ns*	

non-Asian racial minority non-white and non-Asian, a category that is inclusive of multi-racial ethnicities and people who identify as Hispanic
ns not significant

Another interviewee, Ashley, explained how examining real data helped her understand how data sets can introduce bias. One was the dragonfly data set, which helped her see that an unbalanced data set can introduce bias. She explained that because there was an over-representation of dragonflies from the Northern Hemisphere, "There are more northern dragonflies compared to anywhere else and that caused a bias. When the machine was using that as a decision tree, that created a bias."

Another example Ashley provided was of the tree data set, which had several issues with biased data collection methods. One issue was due to an oversampling of tree data from one region of a city. She explained, "We also had more people collecting data about trees in Manhattan compared to other boroughs." Together, these excerpts suggest that Ashley, like several other participants interviewed, was able to recall specific and clear examples of how bias can creep into the ML pipeline through issues in measurement and data collection after learning to explore data sets.

Curriculum aspect 4: real-world stories and anecdotes

Anecdotal stories about real-world cases in which ML was used correctly (and incorrectly) were impactful on participants'ML knowledge, particularly their understanding of ML limitations. A few interviewees (4 of 18) shared that these impactful stories about ML in the real world came from stories they heard during the Institute from their instructor. For example, one interviewee, Riley, shared that an aspect of the Institute that impacted their understanding of the limitations of ML, particularly issues related to the un-explainability of black-box systems, was an anecdotal story shared by their instructor about a NASA brown dwarf classifier that had become biased by an over-representation of brown dwarf training

Table 10 Aspects of the curriculum that interviewees reported as impactful on their ML knowledge

Curriculum aspect 1: ML methods
Learning about various ML methods positively impacted participants' ML knowledge. In half of the interviews (9 of the 18), interviewees explained that they had some prior knowledge of AI and ML; yet, after the Institute, they felt they understood a wider variety of ways that ML could be used in general and also in scientific research, and that they had a "deeper understanding" of these methods Impactful curricular activities: Slice of ML, Neural Networks, and making the ML Flowchart
Curriculum aspect 2: Wallace and dragonflies
Using a dragonfly data set with Wallace, a no-code online platform, to apply ML methods (i.e., linear regression and PCA) had a positive impact on interviewees' ML knowledge. Under half of the interviewees (8 out of 18) mentioned dragonflies and Wallace when describing their various types of ML knowledge, e.g., the purpose of ML, the ML pipeline. While no specific activities were mentioned, the interpretable data set and no-code platform seemed to be a touchstone, something interviewees referenced to help them explain ML concepts and processes
Curriculum aspect 3: Data sets
Working with data sets in the context of ML projects had a positive impact on interviewees knowledge of bias in ML. Several interviewees (5 of 18), touched on how important interacting directly with the data was to developing their understanding. Some shared that this helped them understand how bias could be identified in data, others described how bias could be minimized during data collection and data cleaning. These data sets used interpretable features (i.e., length and width) and described phenomena that participants could interact with during the Institute Impactful data sets: *Local tree features, dragonfly features*
Curriculum aspect 4: real-world stories and anecdotes
While mentioned by only a few interviewees (4 of 18) anecdotal stories about real-world ML use-cases seemed memorable and possibly impactful on ML knowledge of ML limitations. Interviewees described these stories as example cases from which they learned how ML tools lack human-like intelligence and insight

images (with the NASA logos printed in the corner of the picture). They explained,

> *Someone was trying to make a neural network to pick out brown dwarfs, I think. It had a very high accuracy in the training data, but then it did terribly in the testing data. That was just because all the brown dwarves in the training set had a NASA logo in the bottom left corner or something, and then because it's a black box, they didn't know that was what it was. They thought it was really accurate at producing the brown dwarves, but instead they were just looking up the NASA logo and that kind of thing.*

Other anecdotal stories from instructors emerged in the interviews. For example, interviewees recalled a story about an AI tool that had learned to identify types of fish based on whether they were being held by a person. Yet, the story about the NASA logo was most popular. Sophia explained that this story, "*really opened my eyes to the realization that okay, we have to be specific with machine learning because if not, it's going to focus on things we don't want it to focus on.*" Interviewees seemed to draw from anecdotal stories an understanding that AI/ML systems do not have human-like intelligence. This understanding helped them make sense of ML's limitations.

Discussion

In this study we used a mixed methods approach to triangulate survey results with interviews and exit tickets from 42 youth participants. Qualitative analysis offered several insights into the survey results. First, we found the curriculum had a large positive impact on participants' ML knowledge and self-efficacy. Participants' descriptions of their knowledge of ML before the Institute suggest they arrived with a range of prior knowledge; some were aware that AI/ML is involved in functions they use daily in social media and on their phones, others had some knowledge of ML *methods* used in everyday technologies. Participants explained that over the course of the Institute they developed a new understanding of the ways that a variety of ML methods can be used in scientific research as well as in everyday life. Participants shared that they enjoyed the participatory simulations in the Institute, they appreciated the creation of flowcharts to reinforce their ML knowledge and skills, and they valued the opportunity to practice decision making about whether ML could and should be used.

Emerging evidence of youth understanding of bias in ML from exit tickets

Results also showed large positive gains across the majority of the target learning outcomes. Interestingly, surveys found no evidence of gains across two key targets: (1) knowledge of bias in ML and (2) the skill of applying ML in science. This finding is noteworthy given that one of the curricular goals was to prepare high school students to use ML methods in the context of natural science, with an understanding of how ML systems can come to be unpredictably biased. There are several plausible explanations for these outcomes, such as (a) imperfect alignment of the instruments with the curriculum, (b) low fidelity of curriculum implementation (e.g., pedagogy was not aligned with the learning objectives), or (c) low

participant engagement. Exit ticket responses suggest that (a) is most likely and there may have been an issue of alignment between the instruments' sensitivity to the curriculum's impact.

A majority of exit ticket respondents correctly described examples and types of bias in ML, as well as other ML limitations. They wrote justifications for arguments for and against using ML in various scientific and everyday contexts. Answers drew from technical, ethical, and cultural perspectives, suggesting an emerging understanding of "bias" grounded in the ethical issues of bias (i.e., impacts on communities) as much as in technical issues (i.e., unbalanced data sets). Respondents tended to correctly answer exit ticket questions about which ML methods *should* be used in a given context with greater frequency than they correctly answered questions about which ML methods *can* be used. Future research into youth conceptualization of bias in ML systems may offer insights into how youth make-sense of ethical and technical issues in ML systems.

Interviews suggest a shift in youth interest in ML careers

Survey responses showed little change in another construct of interest, youth *interest in ML careers*, yet interviews show that the nature of participants' interest in ML careers evolved over the course of the Institute. At first, interviewees were interested in ML careers out of concern as to how ML would impact the workforce. After the Institute, interviewees expressed interest in ML careers, because they wanted to learn more about opportunities to use ML to (a) advance their career or (b) help others. These are youth of a generation of generative AI (Chan & Lee, 2023), thus interest in ML/AI is natural and perhaps explains the initial high interest in ML careers. However, that interest became more nuanced as participants' gained knowledge of ML methods, bias, and applications in scientific research during the Institute. Further research is needed to investigate the subtleties of these changes in participants' interests and other affective learning outcomes.

Differential impacts of curriculum on youth by URG membership

A major focus of this work was to design the AI curricular content for inclusivity and accessibility such that all learners, regardless of their gender, ethnicity, or prior knowledge, would be able to engage and learn from the materials. The participant group primarily included youth from URGs in STEM. Analysis of concept inventory results revealed that participants from URGs and non-URGs *both* show significant gains in ML knowledge and skills according to MLCI scores. While both groups experience moderate to large effects, the effect size for youth from URGs is highest. Aligning with prior STEM education literature (Whitcomb & Singh, 2021), we found that participants from URG and non-URGs had significantly different prior knowledge, where participants from URG's pre-survey scores were lower than those from non-URG's pre-survey scores. This difference persisted in post-survey results as well, which suggests that while our program significantly improved URG participant ML knowledge and skills, it did not close the gap observed between scores from URG and non-URG members. While white and Asian males began and ended the program with higher scores than their counterparts from URGs, results show that the curriculum decreased the gender and race gap in ML knowledge.

Further analysis across demographic groups by ethnicity and gender revealed interesting differences. For example, women and non-Asian participants showed significant gains in knowledge of societal implications of ML, while their white-male counterparts did not. These differences were higher between ethnic groups than between gender groups. This distinction aligns with the curricular content, which focuses more on societal implications of AI for under-represented ethnicities than for under-represented gender identities. Furthermore, this finding agrees with prior research suggesting that youth may be drawing from their own lived social experiences, such as encounters with societal biases, as they make and making connections to their newly gained ML knowledge (Solyst et al., 2023). On the other hand, changes in participant interest in ML careers varied by gender, but not significantly by ethnicity. Responses from participants who identified as female or non-binary demonstrated a positive shift in interest in ML careers, while responses from their male counterparts did not. This finding is largely explained by pre-survey differences, where female and non-binary participants' expressed interest in ML careers was significantly lower than that of their male counterparts upon beginning the Institute. This pattern in our data mirrors the findings from prior research, which show similar gender differences in STEM educational and career choices among high school age youth (Chen et al., 2024; Sadler et al., 2012; Wang & Degol, 2013). Our work shows that while similar patterns exist in ML education, there is potential for these gender gaps to close. While youth participants demonstrated overall shifts in ML knowledge and attitudes, differences across groups warrant for further research on curricular elements that may interest or benefit specific groups.

Implications for policy and practice

- When considering recommendations for high school science curricula or grade-level standards for AI edu-

cation, policy makers should take into account that high school age youth can learn and develop self-efficacy in learning foundational ML knowledge and skills through engagement with a science-integrated ML curriculum.

- When evaluating curricula, policy makers and school leaders should consider the demographic groups of participants. Our findings suggest that ML curricula might have differential effects on cognitive and affective learning outcomes (i.e., ML knowledge and skills, self-efficacy in learning ML, and interest in ML careers) of youth belonging to different demographic groups.
- When determining whether a curriculum effectively impacts youth knowledge of ML bias, policy makers should be aware that measuring this construct may be particularly challenging. Our findings suggest the evaluation of a curricular impact on youth knowledge of bias in ML may require triangulation across multiple measures, including open ended response questions and retrospective, individual interviews.

Conclusions

In this paper, we present a 4-week science-integrated high school ML curriculum that aims to prepare youth for a scientific research mentorship program in a museum setting. The curriculum leverages unplugged resources, interactive tools, scientific data sets, embodied learning, active learning and project-based learning methods to make advanced concepts accessible to high school youth. Participant responses to pre-/post-surveys, exit tickets, and retrospective interviews offer evidence that the curriculum achieves three key goals: (1) increases knowledge of ML methods and societal implications of ML as well as skills in planning and analyzing ML projects and results; (2) decreases the URG gap in ML knowledge; and (3) positively impacts students' self-efficacy in learning ML. This work is a unique contribution to ML learning in informal spaces, offering findings from an evaluation of the impacts of a novel curriculum on ML in the Natural Sciences on ML knowledge, skills and attitudes for diverse high school learners.

Limitations and future work

Triangulation of non-significant survey results with exit tickets and interviews suggests that, at the time of this publication, the current version of the MLCI may not be aligned with two of the ML concepts targeted by the curriculum: (1) knowledge of ML bias and (2) application of ML in the sciences. Alternatively, it may be that advancing knowledge of bias in ML and applications of ML in scientific research among high school age youth is a particular challenge for ML educational interventions.

We are continuing our analysis to explore these outcomes. For example, after participants complete an academic year of mentored research, we will again measure their understanding of bias in ML and their skill in applying ML to scientific research. It may be that after a year of studying how to apply ML in the context of scientific research, participants' knowledge and skills in these areas may improve. In future work, we will also further hone the MLCI, with particular attention to items on bias when administered to larger and more diverse audiences.

In the long term, we aim to make our learning materials more accessible to science educators in both formal and informal settings. This would involve scaffolding for various levels of expertise, documentation and distribution efforts, and modifications to the curriculum for different subject domains. Continual effort is needed to keep learning materials relevant as algorithms for scientific research and analysis evolve rapidly. Finally, there is a need for analyzing long-term influence of the curriculum on youths' ML knowledge, skills, attitudes and career interests.

Abbreviations

ML	Machine learning
AI	Artificial Intelligence
CS	Computer science
STEM	Science, Technology, Engineering, and Mathematics
PCA	Principal component analysis
ANNs	Artificial neural networks
URGs	Under-represented groups (URGs)
RQ	Research question
MLCI	ML Concept Inventory
APAI	Attitudes and Perceptions of AI

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40594-025-00543-5.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Acknowledgements

This work was made possible by the generous support of the National Science Foundation (ITEST Award # 2049022), as well as by the gracious contributions of our alumni advisory board -- Michelle Cao, Swathi Chakrapani, Hunter Dillard-Jakubowicz, Ariana Gamarra, Wilson Hernandez, Desiree Pante, Michelle Li, Douglas Reigel -- and AMNH research scientists: Ruth Angus, Jessica Ware, and Ward Wheeler. We thank the anonymous reviewers for their comments and suggestions, which significantly improved the quality of this manuscript.

Author contributions

All authors contributed to the study conception and design. Material preparation was led by Gabrielle Rabinowitz in collaboration with Katherine Moore. Data collection and analysis were performed by Katherine Moore. The

manuscript was written by Gabrielle Rabinowitz, Safinah Ali, and Katherine Moore and all other authors provided comments and suggested edits. All authors read and approved the final manuscript.

Funding
This work was supported by the National Science Foundation (AWARD# 2049022).

Data availability
The figures are available in this repository: https://bit.ly/3CqSQkm. All data and materials, as well as software applications, support published claims and comply with field standards. The datasets analyzed during the current study are not publicly available in order to protect participant privacy, but are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
This manuscript reports results from the analysis of data collected from humans. Research protocols were approved by the AMNH Institutional Review Board (IRB #: IRB00004118).

Informed consent
Informed consent was obtained from all participants and their legal guardians to report anonymized individual data.

Competing interests
The authors declare that they have no competing interests.

Received: 13 November 2024 Accepted: 24 March 2025
Published online: 19 April 2025

References

Adams, J. D., & Gupta, P. (2013). 'I learn more here than I do in school honestly, I wouldn't lie about that': Creating a space for agency and identity around science. *International Journal of Critical Pedagogy, 4*(2), 87–104.

Adams, J. D., Gupta, P., & DeFelice, A. (2012). Schools and informal science settings: Collaborate, co-exist, or assimilate? *Cultural Studies of Science Education, 7*(2), 409–416. https://doi.org/10.1007/s11422-012-9399-x

Ali, S., Dipaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education: Artificial Intelligence.* https://doi.org/10.1016/j.caeai.2021.100040

Ali, S., Payne, B. H., Williams, R., Park, H. W., & Breazeal, C. (2019). Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International workshop on education in artificial intelligence k-12 (eduai'19)* (Vol. 2, pp. 1–4). Palo Alto, California.

Aliabadi, R., Carter, J., & Wilson, J. (2022). *Community-powered problem solving with AI: A case study of boys and girls clubs.* In G. Stump (chair), *Aspiring for equity: Perspectives from design of AI education [Symposium]*. International Society of the Learning Sciences.

Archer, A., Godec, S., Patel, U., Dawson, E., & Calabrese Barton, A. (2022). 'It really has made me think': Exploring how informal STEM learning practitioners developed critical reflective practice for social justice using the Equity Compass tool. *Pedagogy, Culture & Society.* https://doi.org/10.1080/14681366.2022.2159504

Ayanwale, M. A., Frimpong, E. K., Opesemowo, O. A. G., & Sanusi, I. T. (2024). Exploring factors that support pre-service teachers' engagement in learning artificial intelligence. *Journal for STEM Education Research*. https://doi.org/10.1007/s41979-024-00121-4

Barton, A. C., Tan, E., & Rivet, A. (2008). Creating hybrid spaces for engaging school science among urban middle school girls. *American Education Research Journal, 45*(1), 68–103. https://doi.org/10.3102/0002831207308641

Bell, P., Lewenstein, B., Shouse, A., & Feder, M. (2009). *Learning science in informal environments: People, places, and pursuits*. National Academies Press.

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology, 9*(1), 3. https://doi.org/10.1037/qup0000196

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*. https://doi.org/10.1186/s40594-023-00418-7

Chaffee, R., Gupta, P., Hammerness, K., & Jackson, T. (2021). Centering equity and access: An examination of a museum's mentored research youth program. In B. Bevan & B. Ramon (Eds.), *Making museums more equitable: Structural constraints and enduring challenges surfaced through research and practice perspectives.* Taylor and Francis/Routledge. https://doi.org/10.4324/9780367823191

Chan, C. K. Y., & Lee, K. K. W. (2023). The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial generation teachers? *Smart Learning Environments*. https://doi.org/10.1186/s40561-023-00269-3

Chen, C., Doyle, J., Sonnert, G., & Sadler, P. M. (2024). Shrinking gender gaps in STEM persistence: A ten-year comparison of the stability and volatility of STEM career interest in high school by gender. *International Journal of Science Education*. https://doi.org/10.1080/09500693.2024.2388880

Chiu, T. K., Moorhouse, B. L., Chai, C. S., & Ismailov, M. (2023). Teacher support and student motivation to learn with Artificial Intelligence (AI) based ChatBot. *Interactive Learning Environments, 32*(7), 3240–3256. https://doi.org/10.1080/10494820.2023.2172044

Dawson, E. (2014). 'Not designed for us': How science museums and science centers socially exclude low-income, minority ethnic groups. *Science Education, 98*(6), 981–1008. https://doi.org/10.1002/sce.21133

Desjardins-Proulx, P., Poisot, T., & Gravel, D. (2019). Artificial intelligence for ecological and evolutionary synthesis. *Frontiers in Ecology and Evolution*. https://doi.org/10.3389/fevo.2019.00402

Eraslan, G., Avsec, Ž, Gagneur, J., & Theis, F. J. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics, 20*(7), 389–403. https://doi.org/10.1038/s41576-019-0122-6

Falebita, O. S., & Kok, P. J. (2024). Artificial intelligence tools usage: A structural equation modeling of undergraduates' technological readiness, self-efficacy and attitudes. *Journal for STEM Education Research*. https://doi.org/10.1007/s41979-024-00132-1

Kashinath, K., Mustafa, M., Albert, A., Wu, J-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H. A., Marcus, P., Anandkumar, A., … Prabhat, N. (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*. https://doi.org/10.1098/rsta.2020.0093

Kaspersen, M. H., Bilstrup, K. K., Van Mechelen, M., Hjort, A., Bouvin, N. O., & Petersen, M. G. (2022). High school students exploring machine learning and its societal implications: Opportunities and challenges. *International Journal of Child-Computer Interaction*. https://doi.org/10.1016/j.ijcci.2022.100539

Kass, J. M., Pinilla-Buitrago, G. E., Paz, A., Johnson, B. A., Grisales-Betancur, V., Meenan, S. I., Attali, D., Broennimann, O., Galante, P. J., Maitner, B. S., Owens, H. L., Varela, S., Aiello-Lammens, M. E., Merow, C., Blair, M. E., & Anderson, R. P. (2023). Wallace 2: A shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. *Ecography, 3*(e06547), 1–9. https://doi.org/10.1111/ecog.06547

Lao, N. (2020). *Reorienting machine learning education towards tinkerers and ML-engaged citizens* [Doctoral dissertation, Massachusetts Institute of Technology].

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering* (pp. 14–16). https://doi.org/10.1145/3195570.3195580

Lee, I., Ali, S., Zhang, H., DiPaola, D., & Breazeal, C. (2021, March). Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM technical symposium on computer science education* (pp. 191–197).

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., & Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads, 2*(1), 32–37. https://doi.org/10.1145/1929887.1929902

Lee, I. & Perret, B. (2022). Preparing high school teachers to integrate AI methods into STEM classrooms. In *Proceedings of the AAAI conference*

on artificial intelligence (Vol. 36, pp. 12783–12791). https://doi.org/10.1609/aaai.v36i11.21557

Lee, I., & Zhang, H. (2022). Addressing equity in AI and AI education in the Developing AI Literacy project. In G. Stump (chair), *Aspiring for equity: Perspectives from design of AI education [Symposium]*. International Society of the Learning Sciences.

Lee, S., & Kwon, K. (2024). A systematic review of AI education in K-12 classrooms from 2018 to 2023: Topics, strategies, and learning outcomes. *Computers and Education: Artificial Intelligence, 6*(1), 100211. https://doi.org/10.1016/j.caeai.2024.100211

Li, X., Jiang, M. Y. C., Jong, M. S. Y., Zhang, X., & Chai, C. S. (2022). Understanding medical students' perceptions of and behavioral intentions toward learning artificial intelligence: A survey study. *International Journal of Environmental Research and Public Health, 19*(14), 8733. https://doi.org/10.3390/ijerph19148733

Lin, G., Kim, Y. J., Stump, G. S., Stoiber, A., Altuwaiyan, A., Abelson, H., Klopfer, E., & Breazeal, C., (2022). Responsible AI for computational action: Fostering AI literacy in middle school students. In G. Stump (chair), *Aspiring for equity: Perspectives from design of AI education [Symposium]*. International Society of the Learning Sciences.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Long, D., & Magerko, B. (2022). The role of equity in designing co-creative, embodied AI literacy activities for informal learning spaces. In G. Stump (chair), *Aspiring for equity: Perspectives from design of AI education [Symposium]*. International Society of the Learning Sciences.

Longo, G., Merényi, E., & Tiňo, P. (2019). Foreword to the focus issue on machine intelligence in astronomy and astrophysics. *Publications of the Astronomical Society of the Pacific, 131*(1004), 1–6. https://doi.org/10.1088/1538-3873/ab2743

Mahesh, B. (2020). Machine learning algorithms—a review. *International Journal of Science and Research (IJSR), 9*(1), 381–386. https://doi.org/10.21275/ART20203995

Marques, L. S., Gresse von Wangenheim, C., & Hauck, J. C. R. (2020). Teaching machine learning in school: A systematic mapping of the state of the art. *Informatics in Education, 19*(2), 283–321. https://doi.org/10.15388/infedu.2020.14

Martin, F., Lee, I., Lytle, N., Sentance, S., & Lao, N. (2020). Extending and evaluating the use-modify-create progression for engaging youth in computational thinking. In *Proceedings of the 51st ACM technical symposium on computer science education (SIGCSE '20)* (pp. 807–808). Association for Computing Machinery. https://doi.org/10.1145/3328778.336697

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J.C., Shoham, Y., Wald, R., & Clark, J. (2024). *Artificial intelligence index report 2024*. https://doi.org/10.48550/arXiv.2405.19522

National Science Foundation. (2023). *Diversity and STEM: Women, minorities, and persons with disabilities.* Retrieved March 26, 2024, from https://ncses.nsf.gov/pubs/nsf23315/

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Park, J., Teo, T. W., Teo, A., Chang, J., Huang, J., & Koo, S. (2023). Integrating artificial intelligence into science lessons: Teachers' experiences and views. *International Journal of STEM Education, 10*(61), 1–22. https://doi.org/10.1186/s40594-023-00454-3

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education, 96*(3), 411–427.

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE), 19*(4), 1–26. https://doi.org/10.1145/3341164

Sanusi, I. T., Oyelere, S. S., Vartiainen, H., Suhonen, J., & Tukiainen, M. (2023). A systematic review of teaching and learning machine learning in K-12 education. *Education and Information Technologies, 28*, 5967–5997. https://doi.org/10.1007/s10639-022-11416-7

Shrestha, Y. R., & Yang, Y. (2019). Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms, 12*(9), 199. https://doi.org/10.3390/a12090199

Solyst, J., Yang, E., Xie, S., Ogan, A., Hammer, J., & Eslami, M. (2023). The potential of diverse youth as stakeholders in identifying and mitigating algorithmic bias for a future of fairer AI. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), 1–27.

Squire, K., & Klopfer, E. (2007). Augmented reality simulations on handheld computers. *The Journal of the Learning Sciences, 16*(3), 371–413. https://doi.org/10.1080/10508400701413435

Su, J., & Zhong, Y. (2022). Artificial Intelligence (AI) in early childhood education: Curriculum design and future directions. *Computers and Education: Artificial Intelligence, 3*, 100072. https://doi.org/10.1016/j.caeai.2022.100072

Tang, J., Zhou, X., Wan, X., Daley, M., & Bai, Z. (2023). ML4STEM professional development program: Enriching K-12 STEM teaching with machine learning. *International Journal of Artificial Intelligence in Education, 33*, 185–224. https://doi.org/10.1007/s40593-022-00292-4

Touretzky, D., Gardner-McCune, C., & Seehorn, D. (2023). Machine learning and the five big ideas in AI. *International Journal of Artificial Intelligence in Education, 33*, 233–266. https://doi.org/10.1007/s40593-022-00314-1

Touretzky, D. S., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019) Envisioning AI for K-12: What should every child know about AI? In *Proceedings of AAAI-19*. https://doi.org/10.1609/aaai.v33i01.33019795

Waisome, J., Parnell, D., Antonenko, P., Abramowitz, B., & Perez, V. (2023). Board 385: Shark AI: Teaching middle school students AI fundamentals using Fossil Shark Teeth. In *2023 ASEE annual conference & exposition*. https://doi.org/10.18260/1-2--43089

Walker, R., Sherif, E., & Breazeal, C. (2022, May). Liberatory computing education for African American students. In *2022 Conference on RESEARCH in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)* (pp. 95–99). IEEE.

Walsh, B., Dalton, B., Forsyth, S., Haberl, E., Smilack, J., Yeh, T., Zhang, H., Lee, I., Lin, G. C., Kim, Y. J., Stump, G. S., Stoiber, A., Altuwaiyan, A., Abelson, H., Klopfer, E., Breazeal, C., Wilson, E., Aliabadi, R., Tian, J., Carter, J., Long, D., Magerko, B., & Sengupta-Irving, T. (2022). Aspiring for equity: Perspectives from design of AI education. In Chinn, C., Tan, E., Chan, C., & Kali, Y. (Eds.), *Proceedings of the 16th international conference of the learning sciences - ICLS 2022* (pp. 1771–1778). International Society of the Learning Sciences.

Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy–value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340.

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

West, S. M., Whittaker, M. & Crawford, K. (2019). *Discriminating systems: Gender, race and power in AI.* AI Now Institute. Retrieved March 26, 2024, from https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2

Whitcomb, K. M., & Singh, C. (2021). Underrepresented minority students receive lower grades and have higher rates of attrition across STEM disciplines: A sign of inequity? *International Journal of Science Education, 43*(7), 1054–1089.

Williams, R., Ali, S., Devasia, N., DiPaola, D., Hong, J., Kaputsos, S. P., Jordan, B., & Breazeal, C. (2023). AI + Ethics curricula for middle school youth: Lessons learned from three project-based curricula. *International Journal of Artificial Intelligence in Education, 33*(2), 325–383. https://doi.org/10.1007/s40593-022-00298-y

YESTEM Project Team. (2021). *YESTEM Insight 2: What are core equitable practices in informal STEM learning?* Retrieved March 26, 2024, from https://yestem.org/tools/core-equitable-practices/

Yosso, T. J. (2005). Whose culture has capital? A critical race theory discussion of community cultural wealth. *Race Ethnicity and Education, 8*(1), 69–91. https://doi.org/10.1080/1361332052000341006

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences, 4*(2), 165.

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2023). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of*

Artificial Intelligence in Education, 33, 290–324. https://doi.org/10.1007/s40593-022-00293-3

Zhang, H., Lee, I., & Moore, K. (2024, March). An effectiveness study of teacher-led AI literacy Curriculum in K-12 classrooms. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38(21), pp. 23318–23325). https://doi.org/10.1609/aaai.v38i21.30380

Zhang, H., Perry, A., & Lee, I. (2024b). Developing and validating the artificial intelligence literacy concept inventory: An instrument to assess artificial intelligence literacy among middle school students. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00398-x

Zhang, L., & Barnett, M. (2015). How high school students envision their STEM career pathways. *Cultural Studies of Science Education, 10*, 637–656. https://doi.org/10.1007/s11422-013-9557-9

Zhou, X., Tang, J., Daley, M., Ahmad, S., spsampsps Bai, Z. (2021). "Now, I want to teach it for real!": Introducing machine learning as a scientific discovery tool for K-12 teachers. In: Roll, I., McNamara, D., Sosnovsky, S., Luckin, R., Dimitrova, V. (Eds.), *Artificial intelligence in education. AIED 2021*. Lecture notes in computer science (Vol. 12748). Springer. https://doi.org/10.1007/978-3-030-78292-4_39

Publisher's Note