Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications

Luis Morales-Navarro
luismn@upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Yasmin B. Kafai
kafai@upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Vedya Konda
vedyask@upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

Danaë Metaxa
metaxa@seas.upenn.edu
University of Pennsylvania
Philadelphia, PA, United States

ABSTRACT

As artificial intelligence/machine learning (AI/ML) applications become more pervasive in youth lives, supporting them to interact, design, and evaluate applications is crucial. This paper positions youth as auditors of their peers' ML-powered applications to better understand algorithmic systems' opaque inner workings and external impacts. In a two-week workshop, 13 youth (ages 14-15) designed and audited ML-powered applications. We analyzed pre/post clinical interviews in which youth were presented with auditing tasks. The analyses show that after the workshop all youth identified algorithmic biases and inferred dataset and model design issues. Youth also discussed algorithmic justice issues and ML model improvements. Furthermore, youth reflected that auditing provided them new perspectives on model functionality and ideas to improve their own models. This work contributes (1) a conceptualization of algorithm auditing for youth; and (2) empirical evidence of the potential benefits of auditing. We discuss potential uses of algorithm auditing in learning and child-computer interaction research.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Social and professional topics → K-12 education**; *Computing literacy*.

KEYWORDS

youth, algorithm auditing, algorithmic justice, machine learning, child-computer interaction, artificial intelligence

ACM Reference Format:
Luis Morales-Navarro, Yasmin B. Kafai, Vedya Konda, and Danaë Metaxa. 2024. Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications. In *Interaction Design and Children (IDC '24), June 17–20, 2024, Delft, Netherlands.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3628516.3655752

IDC '24, June 17–20, 2024, Delft, Netherlands

ACM ISBN 979-8-4007-0442-0/24/06
https://doi.org/10.1145/3628516.3655752

1 INTRODUCTION

Today, children and youth interact with artificial intelligence/machine learning (AI/ML)-powered applications and algorithmic systems when they socialize with friends, go to school, play games, listen to music, do homework, order food, or watch videos. Given the increasing prevalence of AI/ML in their lives, it is crucial to provide young people with the necessary support to engage with, create, and evaluate AI/ML applications. As such, child-computer interaction (CCI) research on AI/ML literacy has received increasing attention [28, 40, 62]. An obstacle in supporting young people in understanding and engaging with AI/ML ideas is the lack of transparency in ML models. Furthermore, existing research gives little attention to critical issues of computational empowerment [16] such as supporting youth in thinking about the limitations and implications of AI/ML technologies [60] or in considering algorithmic justice; that is, how algorithmic systems may be ineffective, even perpetuate harm, and disproportionately impact vulnerable people [6].

In human-computer interaction (HCI) and algorithmic justice research, an effective strategy for investigating and understanding the opaque inner workings and implications of AI/ML systems is algorithm auditing. Algorithm auditing is a method introduced about a decade ago that involves "repeatedly querying an algorithm and observing its output in order to draw conclusions about the algorithm's opaque inner workings and possible external impact" [45, p. 272]. Most of these audits are conducted with the goal of identifying problematic system behaviors in AI/ML-powered systems [4]. But to date, most research on algorithm auditing has focused on experts, with a few recent studies on how non-expert adults engage with the method.

In this paper, we investigate the role of youth as auditors of ML-powered applications by building on CCI's rich tradition of exploring the various roles young people can have in contributing to the design of computing applications [22, 39]. We conducted a two-week workshop with 13 youth (ages 14-15) in which they designed and audited each other's ML applications. We analyzed pre and post clinical interviews in which youth were presented with auditing tasks to address the following research questions:

- How did youth's identification of potential algorithmic biases and harm change from pre to post?

- How did youth's inferences about data and model design change from pre to post?

In the post interviews, we also analyzed students reflections about auditing activities during the workshop to address:

- What benefits did youth find in auditing applications and having their applications audited?

Our analysis revealed that in post, all participants identified potential algorithmic biases and made inferences about dataset and model design issues. In post, more youth talked about algorithmic justice and next steps to further improve ML models. Furthermore, participants reflected that auditing provided them with new perspectives on model functionality and ideas to improve their own models. This paper contributes (a) a conceptualization of algorithm auditing for youth, adapting methods used with adults in algorithmic justice research by grounding them in the rich history of child-computer interaction research; and (b) an empirical clinical pre/post interview study in which youth completed auditing tasks. We discuss future directions for incorporating algorithm auditing in learning activities and CCI research as a promising practice to promote computational empowerment [16].

2 BACKGROUND

Child-computer interaction (CCI) has been concerned with the different roles that young people can play in the design of computing applications since its early days [22, 39]. This has led to the development of various rich methods to involve children in design processes as informants [18, 51], design partners [20, 65], testers [37, 54], and designers [27, 29] (for a detailed review of the theories and methods driving the participation of children and youth in the design process, see [22]). In the following sections, we delve into how CCI has addressed the role of children as testers and evaluators at large and in the context of AI/ML. Following, we address how auditing differs from other testing and evaluation methods, review current research on non-expert auditing, and work on youth's perspectives on algorithmic justice to propose positioning youth as auditors of their peers' applications.

2.1 Youth as testers and evaluators

Research on testing and evaluation can be grouped into two broad categories: (1) when children test and evaluate applications created by experts; and (2) when children test and evaluate child-designed applications. Druin [20] defines the role of testers as users who also help identify "design and usability issues for revision of prototypes." Engaging children and youth as testers in the design process of technologies created by experts can be traced back to Solomon and Papert's work on LOGO in the mid-1960s, when they conducted year-long iterative test sessions with children to refine the design of the programming language [20, 54]. Since then, children have been involved as testers in the design of Smalltalk [24, 34], Scratch [41], and most child-facing applications. Traditionally, when experts lead the design, it is adult researchers that interpret how children tested the technologies and synthesize the findings of testing sessions [22]. More recent work in CCI has engaged children in testing tangible interfaces for learning mathematics [68], as well as the development of instruments to better understand children's engagement when testing applications [15]. In terms of children as testers of AI/ML systems, some work has been conducted with children testing a machine translation application [37] as well as an application for learning about reinforcement learning [14].

A second strand of research on children as testers and evaluators involves child-designed applications. This work can also be traced back to LOGO, in particular Kafai and Harel's late 1980s research on children as designers of software [29]. Positioning children as designers of instructional software for learning mathematics [25] and video games [31], they created environments in which children could test their software with their peers. Additionally, older peers could take on the role of "consultants" or outside evaluators who examined peer-created software and, by "playing doctor," assisted in identifying and diagnosing problems [32]. While being consultants, children benefited from cognitive distance and were able to provide designers with new perspectives, refining their understanding of problem behaviors. Later work [30] looked at the differences between designer-led usability testing and external evaluation, highlighting that when peers play the role of testers in designer-led tasks, it is the designers that benefit from gaining insights on how to improve their own projects to meet the needs of their users. On the other hand, external evaluation of software provided opportunities for evaluators to "apply the insights gained from their own design process" [30, p. 128]. These research studies highlight how testing and evaluating applications can also support youth in their learning of computing. Since then having children design artifacts that can be tested by their peers has become a common activity in many CCI projects.

Several CCI studies have mentioned the importance of engaging youth in testing AI/ML models, but they often lack detailed findings on how young people evaluate models and what they can gain from the testing process [9, 26, 33, 63]. This is also the case in the AI/ML education literature, where training models has received the most attention [47]. The few studies that have investigated how youth test their own models show promising results. Several studies argue that when youth test models, they build hypotheses and explanations for model behaviors and also come up with new ideas for how model performance could be improved [19, 61, 62]. Recent work shows that testing can support young people in identifying issues related to data diversity, class imbalance, and data quality [57, 58]. Yet testing is not always systematic and in-depth. Sometimes young people, after identifying cases in which models do not perform as expected, instead of trying to fix the models, change their testing practices [67]. Other studies have shown that youth rarely test their models, only doing it when prompted by researchers, or that sometimes they think that by simply testing they can improve model performance without making changes to training datasets, model parameters, or retraining [66]. A couple of studies also engage youth as testers of their peers' models in designer-led testing activities [21, 47], that is, when the designers of the applications guide their peers in the testing process. Notably, none of these studies involved youth as external evaluators that evaluate models from the outside in.

What we learn from these previous studies is that there is already a rich tradition in CCI research of engaging children and youth as testers and evaluators, from traditional software to machine learning applications. In introducing youth as auditors, we are adding a new "role" to the repertoire that is distinct from previously examined roles. Here we describe how auditing is different from other forms of testing and evaluation, review existing research that involves non-experts in algorithm auditing, and research on algorithmic justice and youth that inform our approach to positioning youth as peer auditors.

2.2 Youth as auditors of AI/ML applications

To begin, we note that auditing differs from traditional testing and evaluation in several ways [45]. In algorithm auditing, traditionally, the emphasis is on the system itself rather than how users react to or interact with it, though recent work is beginning to include users as part of the system being audited [36]. Unlike other forms of testing, auditing is systematic, with the intention of drawing conclusions at the level of the system rather than about individual test cases. Finally, audits are generally external evaluations done by independent third parties from the outside-in, based on externally-measured system behaviors.

Traditionally, teams of expert auditors conduct audits using methods such as scrapping, automatically collecting and analyzing data from online sources, or "sock-puppets," in which researchers collect data by imitating user behaviors [4]. For example, an expert audit by Metaxa and colleagues [44] investigated gender and racial representation disparities in Google Images by scrapping and analyzing image search results. They found evidence of under-representation of women and people of color in queries of common job occupations in search relative to the U.S. workforce. Some other audits involve non-experts through crowdsourcing, collecting data in distributed and centralized ways. Here, the involvement of non-experts, for example, could include asking users to install a browser extension that automatically queries a system and logs the resulting data [48].

Recently, HCI researchers have started investigating how non-experts engage with algorithm auditing by involving them beyond crowdsourcing data. Studies have looked at how users engage in emergent, everyday auditing practices, without the participation of experts, on social media platforms [52]. Other work has investigated approaching auditing from a socio-technical perspective by auditing both system and non-expert auditor practices [35, 36]. These studies highlight that non-expert audits can uncover problematic algorithmic behaviors that experts may not be able to find [35, 52]. Notably, all of these studies have been conducted with adults. Another important finding has been the collaborative nature of non-expert auditing which often involves sharing problematic findings with others via social media—a practice that might connect well with youth engaged in similar tasks.

More closely related to our work, DeVrio and colleagues [13] have investigated how non-expert adults involved in auditing tasks make sense of potentially harmful behaviors in algorithmic systems. For instance, they had users conduct Google image searches during think-aloud interviews in which participants were tasked with looking for specific images using keywords that may show potential harmful biases. Following, they asked participants to search for other keywords that may also generate problematic results. The study showed that users' findings and interpretations are based on their prior experiences and exposure to societal biases. Furthermore, users came up with ideas to reduce harmful biases, including increasing representation diversity in the content and in the order in which results are displayed.

2.2.1 *Youth's perspectives towards algorithmic justice.*

While, to our knowledge, no studies have investigated the role of youth as auditors, a handful of studies have researched youth's perspectives towards algorithmic justice and potential harmful biases. Researchers have engaged youth in discussions in relation to high-stakes policing surveillance technologies [59] and hypothetical robot interactions [10]. For instance, Coenraad [11] and Salac et al. [49, 50] investigated youth's perceptions of algorithmic fairness. They found that youth "demonstrated an awareness of visible negative impacts of technology and provided examples of this bias within their lives" [11] but did not have the words to discuss bias or how "invisible bias" emerged. After introducing examples of threats to equity, youth were able to discuss visible and invisible issues of equity. Salac and colleagues [49] presented children and youth with scenarios of algorithmic unfairness to prompt their understandings of how the systems worked. The scenarios included bias towards female nurses in image search, a voice assistant not understanding a student with an accent, and a case of inequitable access to school supplies. They explain that children used human and technical lenses to make sense of the issues they were presented with and, at the same time, brought up their own identities and lived experiences in discussing the scenarios. Teenagers examined potential sources of bias and considered the effects these could have in different contexts and on individual people as well as communities.

Solyst and colleagues [55] have also investigated youth's perspectives towards fairness, finding that youth have a desire for agency to participate in the design of technology and define how applications should work. In a different study, they engaged youth in activities to identify algorithmic biases and propose ways to address these [56]. Here participants interacted with examples of image search on Google and image generation in DALL-E, finding that youth identified various types of biases and different potential harms that these could cause. Furthermore, in computing education research, audits have been discussed for their potential as productive opportunities for critical inquiry in which learners investigate the limitations and implications of computing applications [46]. For example, inspired by algorithm auditing research, Walker and colleagues [64] adapted Buolamwini's [8] ideas about "evocative audits" into a learning activity in which young African American students used art to reflect on the harm that algorithms may inflict on their communities.

2.2.2 *A new role for youth.*

The previous research findings on algorithm auditing with non-experts and current work on algorithmic justice and youth provide us with a promising foundation to conceptualize the role of youth as auditors in the tradition of CCI research. Here, it is possible to imagine different ways in which youth could be positioned as auditors of applications. For instance, building on expert audit research on sock puppet auditing [4], youth could be guided to learn about auditing by creating fictional personas to collect data and evaluate how systems behave differently depending on who uses them. Building on non-expert auditing work on emergent audits in social media [52], CCI researchers could investigate how youth audit popular applications (such as TikTok) both in "the wild" and in auditing workshops. Similarly to DeVrio, Solyst, and Salac's work [13, 49, 56], CCI researchers could design and co-design tools and learning activities to engage youth in auditing the technologies they use in their everyday lives. Finally, building on AI/ML testing activities [21, 47, 57] youth could audit each other's applications. We further discuss this approach in the next subsection.

2.2.3 ***Youth as peer auditors***. In this paper, we examine positioning youth as peer auditors of AI/ML applications. As peer auditors, youth can audit applications designed by their peers by collaboratively and iteratively querying the systems to evaluate their behaviors against expected behaviors. Like the consultants in the LOGO studies, youth can "play doctor" and assist in identifying and diagnosing problems [32]. Playing the role of an auditor may have similar benefits to those already identified when youth test their own applications, including identifying issues related to data diversity, class imbalance, and data quality, building hypotheses and explanations for model behaviors, and coming up with new ideas for how model performance could be improved [19, 47, 57, 60, 61]. In the case of classifiers, the context of this study, peer auditing involves iteratively querying the system, comparing auditor-expected classification outputs to system classification outputs, and analyzing the results to make inferences about system behavior (see Fig. 1).

We conducted a workshop in which youth first designed applications that used classifiers. Following, they wrote project factsheets that specified the objectives of the projects and their labels/classes. Then, they exchanged projects with their peers. Peer auditors iteratively queried the systems and documented each query. To ensure a wide variety of queries, every five minutes the auditors of the projects changed, with youth rotating through all projects but their own. Finally, auditors analyzed the data gathered and wrote a report in which they were asked to describe when the applications worked as expected, when they observed unexpected behaviors, and possible next steps to improve system behaviors.

To evaluate the potential benefit of peer auditing activities, before and after the workshop, we conducted a pre/post clinical interview study in which participants were presented with auditing tasks and asked to explain what they were thinking as they completed them [17].

3 METHODS

3.1 Participants

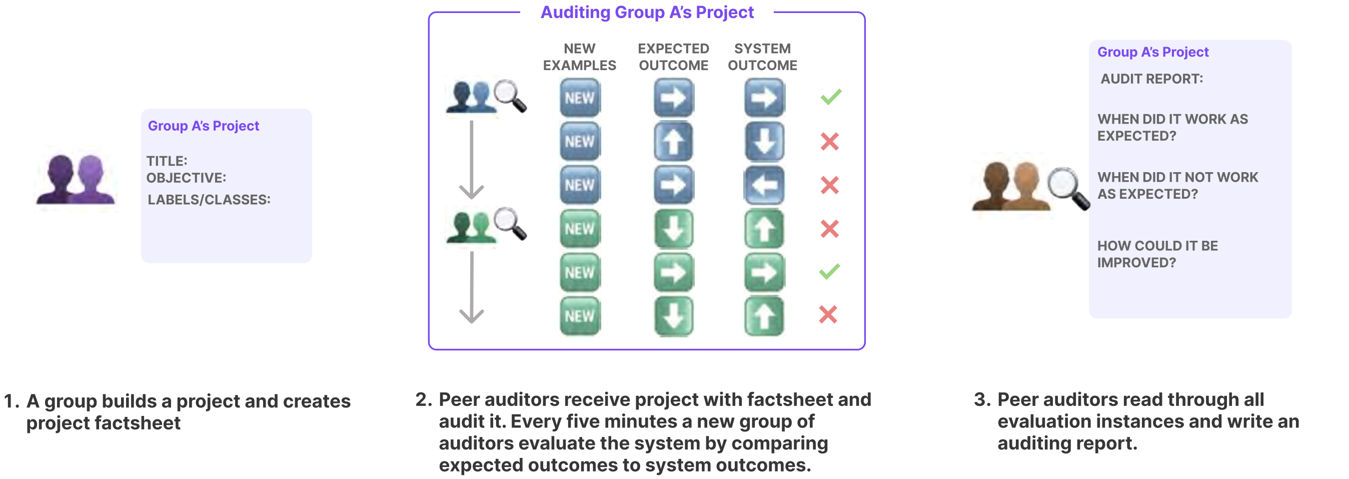
We held a two-week in-person workshop at a science center in the Northeastern United States with fifteen youth (ages fourteen to fifteen) who had shown interest in STEM by taking part in an after-school program meant to increase participation for historically underrepresented communities. Thirteen obtained guardian consent and assented to participate in research. Participants were already acquainted with one another, having participated in the science center program for at least a year. Out of the participants, six identified as female and seven as male. Of the participants, seven identified as Black, five identified as White, three as Latinx, two as Asian, and one as Native American, with five choosing more than one category. Eleven participants had taken computing classes at school or attended out-of-school CS workshops. None had taken workshops or courses on AI/ML (see Table 1). Science center staff sent out paper handouts and emails inviting youth to take part in the study. Before the study began, guardians completed consent forms that included a brief explanation of the research, and youth gave their assent to participate. The institutional review board of the university approved the study protocol. All names mentioned in the paper are pseudonyms.

3.2 Workshop activities

During the workshop, participants learned about ML in the context of designing, testing, and auditing physical computing (e-textiles in particular) applications. Each workshop session had a duration of 3.5 hours, which included a 30-minute community-building activity and a 15-minute snack break. In the first week of the workshop, youth participated in structured activities to learn about machine learning classifiers, e-textiles, and how to create projects that incorporate ML and physical computing. The physical computing activities provided practical experience for youth to learn how to program the micro:bit microcontroller, use sensors and actuators, construct circuits, and sew with conductive thread. Afterwards, youth participated in hands-on activities to learn about AI/ML, different types of models, the ML pipeline [23], and data design practices [57] for training and testing image, audio, and pose classifiers created using ml5.js (a beginner-friendly machine learning javascript library), as well as Teachable Machine and a similar application for training and testing models with sensor data. They then used Bluetooth to send the classifiers' outputs to the micro:bit.

Auditing played an important part in the work workshop durin both weeks. On the fourth day of the first week, participants were introduced to algorithm auditing and participated in an auditing activity for an image classifier they had designed. For this activity, youth in pairs first prepared a factsheet describing the expected behavior of their project and then handed it over to their peers for auditing. After receiving a project, youth proceeded to evaluate their peers' classifiers, and every five minutes, they exchanged projects to have a wide range of auditors evaluate the projects from the outside in. While auditing, they kept track of individual testing instances on a table. Finally, youth wrote an audit report for the designers of the projects in which they synthesized their findings and made recommendations on how to improve the model. During the second week, we had another auditing session in which, following the same format as in week one, participants audited each other's projects and created audit reports.

To illustrate the activities of the workshop, we describe some of the final projects and peer auditors' key findings. Jackie Star and Emily created a drawing game (Fig. 2 A). The game involved players trying to match drawings displayed on the screen of a micro:bit attached to a pen. The project used an image classifier to classify drawings made by users; if the user drew the right shape, the micro:bit played celebratory music and prompted the player with another shape. Among other issues, auditors identified that the project did not work well with "curvy squares that don't have super sharp angles." Andrés created a sports game that detected different basketball moves. He attached the micro:bit to a glove and used data from its accelerometer to train a move classifier (Fig. 2 B). Auditors found that the project constantly misclassified moves when users



1. A group builds a project and creates project factsheet
2. Peer auditors receive project with factsheet and audit it. Every five minutes a new group of auditors evaluate the system by comparing expected outcomes to system outcomes.
3. Peer auditors read through all evaluation instances and write an auditing report.

Figure 1: The process of peer auditing involves: (1) youth creating factsheets for the projects they designed, (2) exchanging projects with peers for auditing, which involves iteratively querying the system and comparing auditor-expected classification outputs to system classification outputs, and (3) writing audit reports.

Table 1: Self-reported demographic information.

Pseudonym	Age	Gender	Race & Ethnicity	Previous CS experience
Kayla	14	Female	Black	Yes
Lou	15	Female	Black	No
Jerome	15	Male	Native American & Black	Yes
Bryan	15	Male	Asian & White	Yes
Jackie Star	15	Female	White	Yes
Fatimah	14	Female	Black	Yes
Andrés	14	Male	Latinx	Yes
Richard	14	Male	White	Yes
Iván	14	Male	Latinx & White	No
Emily	14	Female	Black	Yes
Luke	15	Male	Black & Latinx	Yes
Stephanie	15	Female	Black & White	Yes
Walter	15	Male	Asian	Yes

were six feet or taller. Iván and Walter created a fighting game (Fig. 2 C & D) that imitated Mortal Combat, where users controlled their players by kicking, doing uppercut punches, or superman punches. The game recognized and classified poses. Auditors noted that the game only worked well when played against plain white walls and when only one person was in the frame.

3.3 Interview design and data collection

We conducted pre interviews a week before the workshop and post interviews on the last day of the workshop. The interviews consisted of two pre-determined auditing tasks to evaluate image classifiers and a text-to-image generative model, accompanied by prompts that were specifically designed to elicit students' ideas (e.g., about bias, data and model design, and justice) in an open-ended manner. The interviews were conducted in pairs and deliberately structured to resemble conceptual change research interviews [17, 38, 53]. Each interview had a mean duration of 23 minutes, with individual interviews ranging from 12 minutes to 31 minutes. We recorded audio and participant screens during the interviews.

In each interview, we presented youth with two image classifiers that were intentionally faulty. This task was adapted from prior research on ethics in AI/ML education in which youth were presented with faulty cat and dog classifiers and their datasets [3]. We prepared classifiers with inter- and intra-class imbalances (e.g., in the berry classifier, the training data included pictures of strawberries in all shapes, sizes, and colors, while blueberries and blackberries were limited to a few very similar pictures), spurious relationships (e.g., in the drawing tools classifier, all pictures of pencils in the training data included human hands and none of the pictures of other tools included hands; pictures of markers or paint brushes with hands were misclassified), and overfitting issues (e.g., in the pet classifier, all bunnies in the training data were white, as such,

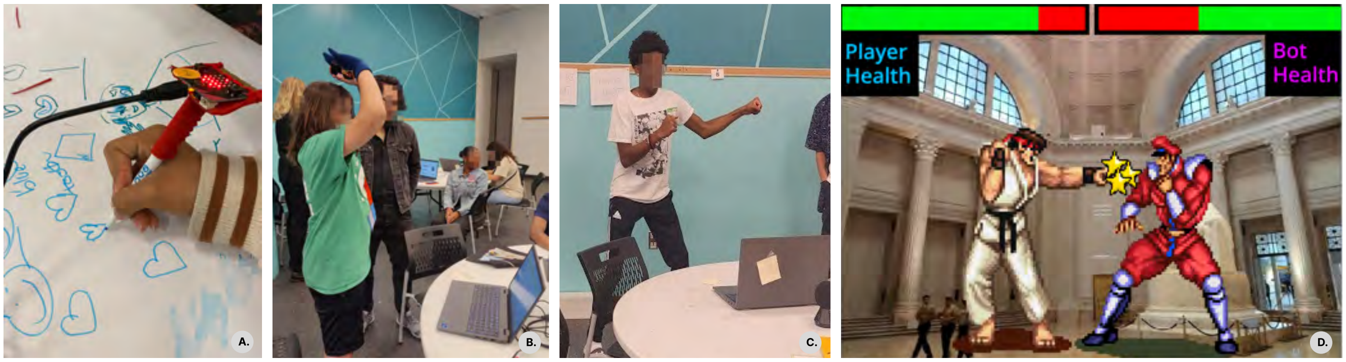


Figure 2: Auditing youth's final projects. (A) A participant audits Jackie Star and Emily's drawing game. (B) A participant audits Andrés sports game. (C & D) A participant audits Iván and Walter's fighting game.

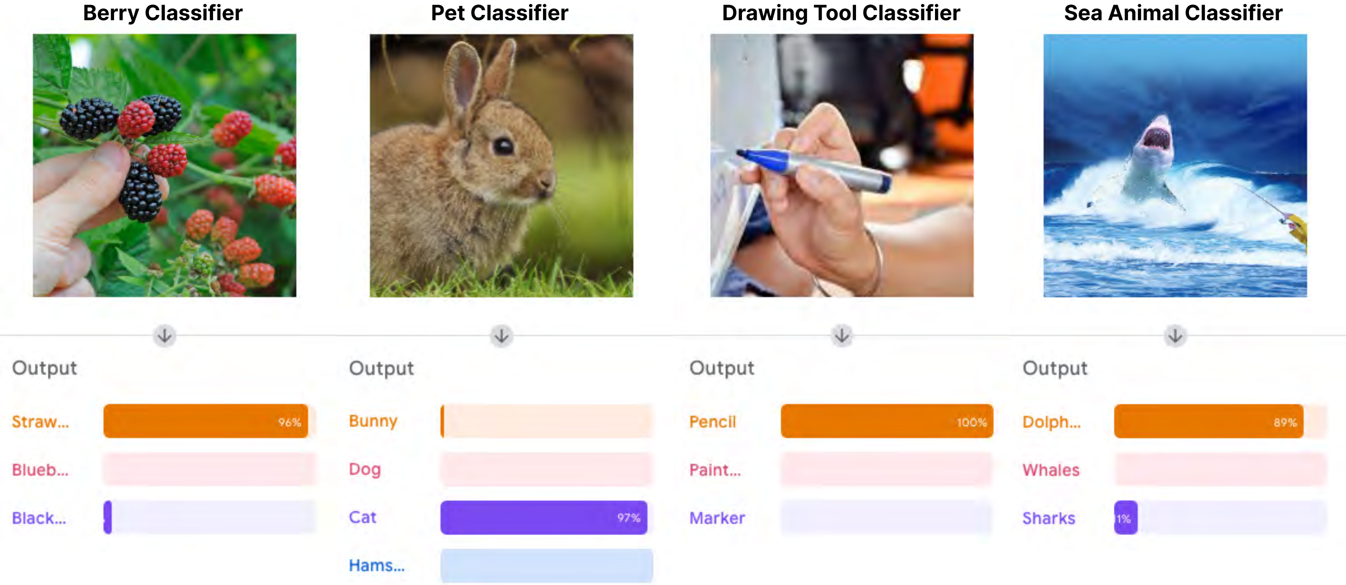


Figure 3: Faulty classifiers used in pre/post auditing tasks included a berry classifier, a pet classifier, a drawing tool classifier, and a sea animal classifier. The figure shows cases in which the classifiers did not work as expected.

bunnies of other colors were misclassified). During the interviews, we first asked participants to interact with the classifier and explain its functionality (for correct and incorrect results) (see Fig. 3). After a few minutes, we showed them the data used to train the classifiers and asked them to explain how it worked, why it did not work with certain images, and to share any ideas they had for how to fix them.

For the text-to-image tasks, participants were asked to evaluate the outputs generated by DALL-E mini [12]. This task was adapted from everyday algorithm auditing studies that have had users conduct image searches on Google to identify potentially harmful behaviors [13], or used results from image searches and DALL-E generated images to prompt participants to reflect about algorithmic justice [56]. In each interview, we asked participants to generate images for five topics (e.g., weddings, beautiful women, librarians, scientists) that had shown potential problematic results in previous studies and to come up with new examples that may yield problematic results (see Fig. 4). We asked participants to share their thoughts about the results, whether they thought they were biased or discriminatory in harmful ways that might negatively impact people, and what they would do if they had the option to change or adjust the results.

Additionally, in the pre interview we asked participants if they had heard of harmful bias or discrimination in algorithmic systems or in applications used in their everyday lives and to provide any examples they were familiar with. In the post, we asked participants to tell us "a little bit about what you learned or noticed when auditing other people's projects" and whether auditing was helpful or not in the process of making applications and addressing potential

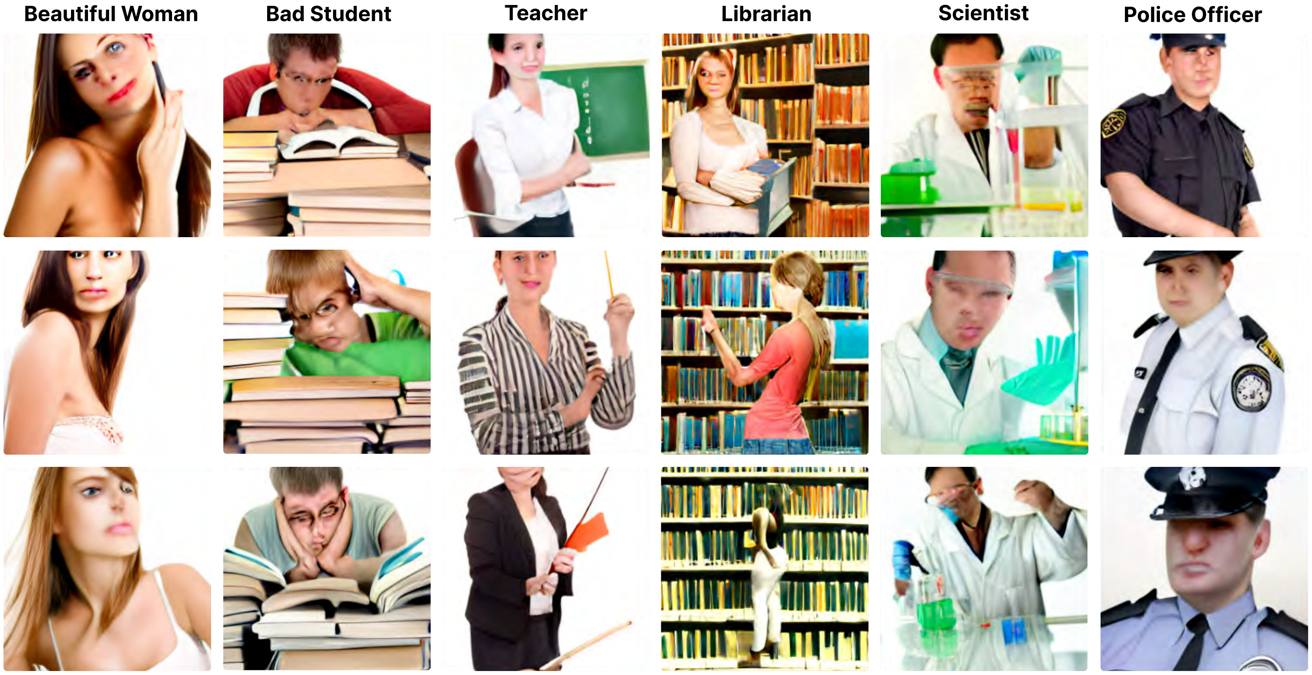


Figure 4: Images generated by DALL-E mini for the following prompts: "beautiful woman," "bad student", "teacher", "librarian", "scientist", and "police officer".

problematic behaviors. Across both tasks and the additional question, we used follow-up prompts to elicit details from participants: "What do you think could be the causes of these issues?´´ "Can you tell me more about that?" "What makes you think so?" "Can you give me an example?" "Why did you say that?"

3.4 Analysis

We conducted two rounds of inductive-deductive thematic analysis [7]. We used an automated transcription tool and then verified the transcripts for accuracy. As part of the initial analysis, two researchers inductively coded one-third of the data (consisting of three sets of matching pre/post interviews) identifying 77 emergent codes. Following this, codes were grouped to develop a codebook (9 codes and 45 subcodes); this process was also informed by previous research [13, 56]. The codes included bias, data design, project reflection, antropomorphizing, auditing, algorithmic justice, identifying issues, prior experiences, and model design. Each code was specified through subcodes. For example, bias was broken down into subcategories, including race bias and gender bias. In a second phase of analysis, two researchers applied the coding scheme to all pre and post interviews (see codebook in Appendix A). During the coding process, the researchers analyzed the data together, actively communicating with each other, discussing disagreements, and striving to reach consensus. They also had ongoing discussions with a third researcher who was knowledgeable about the data and the coding framework. We coded 1250 instances in which we observed the themes. Because this is an exploratory study with only a few participants, we focused on reaching agreement during the coding process by coding together and resolving any differences through extensive discussions rather than relying on inter-rater reliability, also in keeping with recommendations in prior work [43].

4 FINDINGS

In this section, we begin by discussing how youth identified potential algorithmic biases, then move on to their considerations of harm and justice during the interview auditing tasks. Then we discuss youth's inferences about data and model design issues, as well as their suggestions for how to address them during the interview auditing tasks. Finally, we consider youth perspectives on the benefits of peer auditing activities.

4.1 How did youth's identification of potential algorithmic biases and harm change?

4.1.1 ***Identifying algorithmic biases.*** Notably, all 13 participants were able to identify potential algorithmic biases in the classifier task during the post-interview compared to 11 participants in the pre-interview. Participants identified potential biases related to body shapes, breed (in the case of animals), color, size, shape, and context/location. In the pre, for instance, Kayla argued that the pet classifier was biased against German Shepherds as it classified them as cats while other breeds of dogs (Dalmatians, Golden Retrievers, and Corgis) were correctly classified. While fewer instances in which youth discussed potential biases were identified in post (92 instances) than in pre (131 instances), in post, participants related biases to data and model design issues. We further discuss this in the next section. The only subcode in which the number of both participants and instances increased in post was for context/location bias, going from 5 participants (11 instances) to 12 participants (36 instances). Jackie Star, for example, noticed that, when testing a "tools for drawing" classifier in the post, any image that contained a hand holding the tool was classified as a pencil. She explained that the model was biased because of the context in which pencils were probably portrayed in the training data.

In the DALL-E task, overall, the number of participants that identified biases increased from 9 in pre to 12 post. Across most subcodes (age, body appearance, color, gender, context/location, race, and relevancy), the number of students that identified biases increased. Race and gender biases were the most commonly identified by participants. For racial biases, 9 participants noted these in pre and 12 in post. Fatimah, in post, argued that DALL-E was biased in favor of White people because all words related to professions (librarians, teachers, lawyers) generated images of White people. As an experiment, she tried to generate images for "thug" because "I was expecting, like, some Black dude, you know, but it was a White guy that looked like Eminem with a hood on." This example also shows how youth's expectations were also based on their personal biases; we discuss this below. Seven participants noticed gender biases in pre and 11 in post. In post, for example, Stephanie noted that gender in the images depended on the prompts used to generate images, with all police officers generated by DALL-E being male and all librarians being female.

Participants (2 in pre and 5 in post) identified biases in favor of irrelevant, dated results. Lou claimed that the results were biased in favor of "early 2000s pictures" because of the "hairstyles and clothes" of humans generated by DALL-E as well as the "outdated values" represented in the images. Andrés, using as an example the images generated with the prompt "beautiful woman" (see Fig. 4), explained that the results did not represent beauty standards today because "Lizzo, everyone calls her beautiful but none [of the generated pictures] looked like her." Fatimah and Emily also commented on how the representation of gender and sex was outdated, noting that the pictures did not include gender and sex non-conforming people and queer couples. Kayla also noticed biases in favor of older/outdated representations of humans in the pictures generated for both "good students" and "bad students," explaining that all images included books and "chalkboards and everything" when she expected "more like actually [students] working probably like less of the books, cause most of it is digital now."

While completing the tasks during post, nine participants talked about trying to break the classifiers, or DALL-E, as an approach to auditing. As Iván explained, this involved "challenging it [the classifier] to see what would break it."

When identifying potential algorithmic biases, participants sometimes reflected on their own personal biases and their perceptions of societal biases. This was particularly salient in the DALL-E task. In post 11 participants voiced being aware of their personal experiences and biases when evaluating outputs, compared with 7 participants in pre. For example, when looking at the outputs of DALL-E for teachers, Iván reflected, "from my personal experience, teaching as a very female-dominated profession." In a similar instance, when looking at the outputs for librarians, Kayla said, "It makes sense that it will all be women? I've personally never heard of a male librarian... I really haven't. I've never seen a male librarian." We observed 7 participants discuss how societal biases were reflected in the outputs during both pre and post. Here for example, Fatimah discussed that the outputs reflected what popular media looks like, saying that "representation in media is necessary." In a similar instance, when looking at the outputs for gamers, Iván noted "a lot of YouTube channels it has... I feel like it's mainly run by White guy gamers."

4.1.2 *Considering justice and harm.*

The number of participants who talked about justice and potential harm increased from 7 in pre to 12 in post. While completing the DALL-E tasks, youth showed diverse understandings of harm as context-dependent, being able to think about harm in terms of how they could be affected by algorithmic systems and how these could affect other people.

A common concern was the representation of professions and how it may discourage people from pursuing certain careers. Luke explained how a lack of representation can be harmful: "For the scientists, like kids saying they want to be scientists, looking up scientists and not seeing anybody like them can kind of be like, whoa, if nobody that looks like me is a scientist, then should I really become one?" Similarly, Luke argued that beauty standards portrayed by AI-generated images could affect people's mental health and self-esteem.

Some participants argued that these systems could exclude people but are not necessarily harmful. In pre, Iván explained, "they're biased, they're not like making anyone look bad, but they're more like excluding people." Notably, his perspective changed and in the post he expressed that exclusion could be potentially harmful "I feel like at this point right now, it's not harmful. But as it evolves, it will be [. . .] if these issues aren't addressed by adding more diversity." At the same time, only one participant, Richard, expressed in both pre and post that AI/ML systems could not be harmful. In pre, Richard said, "I think if you're getting harmed by an AI, I don't know, that's more of a personal problem." In post, he explained "I don't think it can be harmful. I do think it's discriminatory. You're not gonna, like, get offended by the AI."

Walter and Jackie Star explained that harm depends on the context in which AI/ML systems are used. Jackie said: "Yeah, it just excludes. Like in this context, with just generating pictures. I don't know if it's really impactful." Similarly, Walter argued that harm depends on whether "someone's using this in an actual like, like a practical use".

During the post interviews, participants brought up cases of algorithmic injustice they were already familiar with. Walter talked about how racial biases in image generation could also be present in how people are recognized and classified in policing systems that could be biased "towards protecting, like, White males or something like that." Kayla also gave a similar example of how a "Black man who had never done anything wrong in his life" could be identified as a criminal in a biased facial recognition system. Lou noted that in medicine, if AI/ML systems do not recognize Black patients, it could be dangerous as people could be misdiagnosed.

4.2 How did youth’s inferences about data and model design change?

4.2.1 ***Making inferences.*** As participants interacted with the auditing tasks, they explained what they observed by coming up with ideas about data and model design issues that could impact the performance of the systems. In post, each participant identified an average of 12.8 possible data and model design issues, compared to an average of 6.9 issues in pre.

In the classifier tasks, all youth identified potential model and data design-related issues in the post interview (compared to 11 participants in pre). All participants but one identified more data design issues in the post than in the pre. At the same time, the number of issues identified increased across all subcodes (i.e., model features, data composition, data diversity, data context, data sources, and class balance) except data quantity. This shows that through the workshop, in which they designed and audited applications, youth may have developed a more nuanced understanding of how data quality impacts model performance, moving beyond the popular adage that data quantity drives model performance. In the DALL-E task, the number of youth that identified data and model design issues increased from 8 in pre to 10 in post.

In pre, participants described their understanding that models base their performance on some of the features of the data. When interacting with the pet classifier, Fatimah argued that it was important to “provide more features” to the model so that it would know what to look for and not make decisions just based on color. Other participants voiced similar ideas, talking about how the models classified images based on “key factors and traits.” Similarly, in the post, they talked about “main identifiers” and how some features “mattered more than others.”

While participants often made inferences about data diversity in general terms (11 in pre, 13 in post), in post, they referred more often to data composition (5 participants in pre, 10 in post), context (1 participant in pre, 10 in post), and sources (7 participants in pre, 8 in post). In terms of data composition, for example, Lou talked about how different camera shots influenced performance, noting that all close-ups in the sea-life classifier were classified as sharks. Jackie Star agreed, “Yeah, definitely a bias towards sharks if it was close up to a face, because that’s probably all that it really is like taught on.” When looking at the data set, Richard also noted that all pictures of sharks were taken from the same angle. For the same classifier, in terms of data context, Iván noticed that all pictures of dolphins were of dolphins out of the water. Data context was also discussed in the DALL-E task, particularly with regards to weddings, with students like Fatimah speculating that the data was probably all from the same context because “certainly with Indian weddings, there’s different traditions and different ceremonies that happen for weddings, not just white dress.” The sources for the data used to train models were also discussed, with participants speculating that the data for DALL-E represented what they commonly see on certain YouTube videos (of gamers and weddings) or pictures from stock images or magazines. Jackie Star reflected that data sets are curated by humans that decide on where to source data from; it “shows more human bias than AI bias because if it’s like trained off of like pictures, and that’s kind of like the pictures that it’s seen [...] I think it’s more of like a human problem that the bots are just learning from,” she explained.

4.2.2 ***Coming up with next steps.*** Eight participants in pre and 10 participants in post came up with concrete next steps related to model and dataset design that could be taken to address the issues they identified. Next steps went from adding more data to balancing classes in pre to more nuanced ideas about data composition and augmentation in post. Walter, for example, reflected that to improve the performance of the sealife classifier, it was important to make sure each class had images composed in diverse ways. He discussed that shark images should not just include close-ups but also "zoomed out like the whole body and good lighting.” For whales, he argued that the model probably needed more pictures of whales “out of the water while jumping.” Jackie Star, in post, also voiced some ideas about data augmentation, such as rotating images or making images black and white.

4.3 What benefits did youth find in auditing applications and having their applications audited?

During the post interviews, youth reflected on their experiences auditing each other’s applications and having their applications audited during the workshop. In particular, they valued how auditing provided them with new perspectives, gave designers ideas on how to improve projects, and helped them think about their own projects in new ways.

Eleven participants talked about how auditing provided them with new perspectives related to model functionality and how to improve model performance and their own projects. “This is not taxes; it’s more like a game,” Richard said, describing the role of the auditor as that of someone whose goal is to identify “all the problems.” Overall, youth agreed that auditors were able to bring in new perspectives because they were unfamiliar with the projects and how these were created. Here, Iván noted that auditing involved “not just getting more diverse user input, but feedback from people that don’t think like you.” Lou explained, “you also get different standpoints because people think in so many different ways that, like, you wouldn’t have thought of something and now you can incorporate that.” Luke voiced a similar idea, highlighting that auditors “may see things that [designers] have not seen.” Jerome further reflected on the collaborative nature of auditing, saying that “it’s more than one perspective [...] different viewpoints come together.”

Participants also reflected that the ideas that auditors provided on how to improve projects were helpful. Jackie Star noted that “people were like, well, you could have added more variety to this class,” giving her concrete steps on how to improve her projects. Similarly, Fatimah claimed that the feedback from auditors “helped me humble myself, helped me realize, okay, there are changes I can make, or actually my project is doing much better than I thought it would.”

Seven youth also mentioned that auditing helped them look at their own projects from different perspectives, making connections between what they saw other people do and what they were doing in their own projects. Jerome explained that after auditing, “you can turn around and improve that yourself." Iván explained that after auditing, "I use the logic that I use in their project of challenging it to see what would break it on our project." Lou claimed that after auditing, she was able to avoid other people's mistakes and prevent some of the issues she observed in other people's projects in her own project.

5 DISCUSSION

In this paper, we investigated the potential benefits of positioning youth as peer auditors of AI/ML activities. Here we discuss adapting algorithm auditing methods for youth by grounding them in the rich history of CCI and the findings of our clinical interview study.

5.1 Peer auditing in child-computer interaction

In the case of our study, we built on previous research that positioned youth as evaluators of their peers' applications [30]. We observed several similarities between our work and previous work. For instance, youth benefited from cognitive distance [29], being able to "take perspective" [1] of their own applications and those of their peers. This enabled them to provide recommendations for their peers and to apply what they saw as auditors to their own projects. Like in youth as software consultants research, playing the role of peer auditors was similar to "playing doctor" as youth identified and diagnosed issues. Yet, in our study, youth took a more adversarial approach, describing how, for some of them, the goal was to try to "break" the applications or find "all the problems". This approach differs from the stance of expert auditors—which is about understanding systems with frequent emphasis on problematic behaviors, reflecting both the unrealistic expectations that novices may have and how their understanding of auditing may be influenced by pre-existing ideas about auditing in other fields. For instance, audits in taxation are often perceived as a threat, with people trying to avoid "being caught" by auditors [2, 5, 42] (such perceptions are also portrayed in popular media, e.g., Everything Everywhere All at Once). Further research is needed to better understand how non-expert auditors see their own role.

Auditing is a sociotechnical process. Our study confirms findings from previous work [13] that show that participants' interpretations about algorithmic biases are guided by their personal experiences and their understandings of societal biases. The fact that in post youth were more aware of how their personal experiences and biases influenced their perspectives of algorithmic biases suggests that auditing activities may support youth in taking perspective about both algorithmic systems and their relationship to these. This highlights the importance of thinking about auditing as sociotechnical and furthering our research not only on auditing algorithmic systems but also understanding how non-experts, including youth, audit them [36].

Future CCI research on youth as auditors should also build work on youth as testers and evaluators of expert-designed applications. Positioning youth as auditors of technologies designed and marketed towards them is particularly important, as they may be able to identify issues that designers and adults cannot find. At the same time, recent work conducted with adults on emergent audits, in which users evaluate systems in decentralized and distributed ways to understand their behaviors could be replicated with youth.

5.2 Auditing for algorithmic justice

Our study showed that algorithm auditing tasks used in research with adults [13] cannot only be used with youth to study their perspectives towards algorithmic justice [56], but also be adopted in pre/post interviews to assess the potential benefits of auditing interventions. In particular, we noticed that the classifier task and the DALL-E task had unique affordances in prompting youth to think aloud about different things. The classifier task, which resembled much more closely the workshop activities (designing and auditing applications that used classifiers), prompted youth to make well-informed inferences about data and model design issues. At the same time, the DALL-E task enabled youth to make connections between what they did in the workshop and generative models. This task also prompted participants to reflect on harm by making connections to societal biases and their personal experiences. It may be more difficult to talk about issues of algorithmic justice when talking about classifying bunnies than how certain professions are represented in the outputs of a generative model.

While our findings show that even in the pre interview some participants were able to identify potential biases, it is notable that in post all participants identified potential biases. It was not surprising that some youth were able to identify potential biases in pre, as previous research shows that both adults and teenagers participating in cooperative inquiry sessions and think-aloud interviews can engage with these topics by building on their rich experiences as users of AI/ML-powered applications [13, 49, 56]. The fact that all youth identified biases and made inferences suggests the value of having youth design and audit applications.

Like in previous work with teenagers [49, 56], participants shared their perspectives about algorithmic justice and potential harm. After designing and auditing their peers' applications (in post), they voiced their opinions about harm and justice more frequently. Our findings show that youth's perspectives are diverse, with some recognizing how systems could affect people in concrete ways, others arguing that harm is context-dependent or highlighting the difference between exclusion and harm, and one claiming that the burden of harm lies on the user and not algorithmic systems. These perspectives were informed by participants' positionalities (in terms of race and gender) and their lived experiences. Further research should explore how youth's identities shape their beliefs about justice and harm.

5.3 Auditing and computational empowerment

Lastly, we want to address a larger point about algorithm auditing that connects to on-going discussions about computational empowerment [16]. Computational empowerment focuses on the construction and deconstruction of computing technologies—in our case AI/ML applications, that youth interact with. Deconstruction involves describing, evaluating and reflecting on the values and intentions embedded in sociotechnical systems and considering their possible implications [16]. Auditing activities may be particularly well suited to support the deconstruction process. We note that all youth in the post interview were able to make inferences about data and model design. Whereas this finding is similar to those of research on youth testing their own applications [47, 57] it is worth noting that the inferences were made from the outside-in, on models that participants had not designed and did not know about prior to the task. This suggests that auditing activities, beyond being helpful to identify potential harmful biases, may be productive in supporting people to understand and make sense of blackboxed AI/ML systems. The inferences made by participants show that they made connections between potential biases identified and concrete issues in the design of the models; that is, they thought about biases not as abstract but as the product of decisions made when building models in the way datasets are designed and model features and parameters are decided.

6 LIMITATIONS

In this paper, we used pre/post clinical interviews to investigate changes in the way youth identified bias and harm and made inferences about data and model design during auditing tasks. As such, we did not focus on the practices that youth engaged with when auditing each other's projects or the findings of the audits they conducted. Future research on peer auditing must include analyzing auditing activities microgenetically, moment-by-moment, to identify key practices and perspectives that youth may have. Such analysis could also provide useful insights into what motivates youth when auditing and what their attitudes and dispositions are towards auditing. Similarly, we did not evaluate if the issues and potential next steps proposed by youth were adequate; this should be done in future research.

One further limitation of the findings is that we did not have a control group in which youth only designed applications. As such, it is not clear if our observations are the product of peer auditing activities, the design of applications, or both. Future studies could use the same clinical interview protocol across three treatments: one in which youth only design applications, one in which youth only audit applications, and a third one in which they design and audit applications.

Finally, our study, like most studies related to youth and algorithmic justice, was conducted with a very small number of youth under very specific circumstances. Considering how youth identities and lived experiences may shape their beliefs about algorithmic justice, future research could intentionally sample youths with diverse experiences and backgrounds to explore how these may relate to their perspectives towards auditing. At the same time, peer auditing and youth algorithmic justice research at large should scale up and move from afterschool workshops to formal classroom settings.

7 CONCLUSION

In this paper, we introduced youth as peer auditors of AI/ML applications. Our research illustrated how youth were able not only to identify various potential biases related to gender and race but also to connect these to more complex issues of data design. Moreover, peer auditing provided youth with valuable insights for designing their own AI/ML applications. Thus, algorithm auditing expands the repertoire of roles available to children and youth in the design of computing applications in child-computer interaction research. While our study was focused on how youth conducted algorithm audits, its opportunities and limitations, and the ways they built on personal experiences, this study also points towards the possibility of including peer auditing in learning activities. Here we see a particular promise to develop algorithm auditing activities that could promote computational empowerment.

8 SELECTION AND PARTICIPATION OF CHILDREN

We recruited youth already enrolled in a STEM afterschool program in a city located in the Northeastern United States. Youth were invited by the organizer of the STEM program to participate via email and through paper handouts. Parents received consent forms prior to the study, which included a brief explanation of the research, and youth assented to their participation. Research protocols and data collection methods were approved by the IRB board of the University.

ACKNOWLEDGMENTS

This analysis and writing of this paper was supported by National Science Foundation grant #2333469. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF or the University of Pennsylvania. We thank Phillip Gao and Isabelle Gu for support in data collection and Deborah Fields and Lauren Vogelstein for their feedback.

REFERENCES

[1] Edith Ackermann. 2012. Perspective-taking and object construction: Two keys to learning. In *Constructionism in practice*. Routledge, , 25–35.

[2] Arun Advani, William Elming, and Jonathan Shaw. 2023. The dynamic effects of tax audits. *Review of Economics and Statistics* 105, 3 (2023), 545–561.

[3] Safinah Ali, Blakeley H Payne, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2019. Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International workshop on education in artificial intelligence k-12 (eduai'19)*, Vol. 2. , , 1–4.

[4] Jack Bandy. 2021. Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction* 5, CSCW1 (2021), 1–34.

[5] Marcelo Bergolo, Rodrigo Ceni, Guillermo Cruces, Matias Giaccobasso, and Ricardo Perez-Truglia. 2023. Tax audits as scarecrows: Evidence from a large-scale field experiment. *American Economic Journal: Economic Policy* 15, 1 (2023), 110–153.

[6] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), .

[7] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association, .

[8] Joy Buolamwini. 2022. *Facing the Coded Gaze with Evocative Audits and Algorithmic Audits.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[9] Debra Burhans and Karthik Dantu. 2017. ARTY: Fueling creativity through art, robotics and technology for youth. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. , , .

[10] Vicky Charisi, Tomoko Imai, Tiija Rinta, Joy Maliza Nakhayenze, and Randy Gomez. 2021. Exploring the concept of fairness in everyday, imaginary and robot scenarios: a cross-cultural study with children in Japan and Uganda. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference.* , , 532–536.

[11] Merijke Coenraad. 2022. "That's what techquity is": youth perceptions of technological and algorithmic bias. *Information and Learning Sciences* 123, 7/8 (2022), 500–525.

[12] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. 2021. Dall· e mini. *HuggingFace. com. https://huggingface. co/spaces/dallemini/dalle-mini (accessed Sep. 29, 2022)*, (2021), .

[13] Alicia DeVrio, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* , , 1–19.

[14] Griffin Dietz, Jennifer King Chen, Jazbo Beason, Matthew Tarrow, Adriana Hilliard, and R Benjamin Shapiro. 2022. ARtonomous: Introducing middle school students to reinforcement learning through virtual robotics. In *Interaction Design and Children.* , , 430–441.

[15] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle gauge: a self-report instrument for evaluating children's engagement with technology. In *Proceedings of the Interaction Design and Children Conference.* , , 614–623.

[16] Christian Dindler, Rachel Smith, and Ole Sejer Iversen. 2020. Computational empowerment: participatory design in education. *CoDesign* 16, 1 (2020), 66–80.

[17] Andrea A disessa. 2007. An interactional analysis of clinical interviewing. *Cognition and instruction* 25, 4 (2007), 523–565.

[18] Liz Dowthwaite, Helen Creswick, Virginia Portillo, Jun Zhao, Menisha Patel, Elvira Perez Vallejos, Ansgar Koene, and Marina Jirotka. 2020. " It's your private information. it's your life." young people's views of personal data use by online technologies. In *Proceedings of the interaction design and children conference.* , , 121–134.

[19] Stefania Druga and Amy J Ko. 2021. How do children's perceptions of machine intelligence change when training and coding smart programs?. In *Interaction design and children.* , , 49–61.

[20] Allison Druin. 2002. The role of children in the design of new technology. *Behaviour and information technology* 21, 1 (2002), 1–25.

[21] Utkarsh Dwivedi, Jaina Gandhi, Raj Parikh, Merijke Coenraad, Elizabeth Bonsignore, and Hernisa Kacorri. 2021. Exploring machine teaching with children. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC).* IEEE, , , 1–11.

[22] Jerry Alan Fails, Mona Leigh Guha, Allison Druin, et al. 2013. Methods and techniques for involving children in the design of new technology for children. *Foundations and Trends® in Human–Computer Interaction* 6, 2 (2013), 85–166.

[23] Rebecca Fiebrink. 2019. Machine learning education for artists, musicians, and other creative practitioners. *ACM Transactions on Computing Education (TOCE)* 19, 4 (2019), 1–32.

[24] Adele Goldberg. 1979. Educational uses of a dynabook. *Computers & Education* 3, 4 (1979), 247–266.

[25] Idit Harel. 1991. *Children designers: Interdisciplinary constructions for learning and knowing mathematics in a computer-rich school.* Ablex Publishing, .

[26] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can children understand machine learning concepts? The effect of uncovering black boxes. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* , , 1–11.

[27] Nathan Holbert, Matthew Berland, and Yasmin B Kafai. 2020. *Designing constructionist futures: The art, theory, and practice of learning designs.* MIT Press, .

[28] Golnaz Arastoopour Irgens, Hazel Vega, Ibrahim Adisa, and Cinamon Bailey. 2022. Characterizing children's conceptual knowledge and computational practices in a critical machine learning educational program. *International Journal of Child-Computer Interaction* 34 (2022), 100541.

[29] Yasmin Kafai and Idit Harel. 1991. Learning through design and teaching: Exploring social and collaborative aspects of constructionism. *Constructionism* , (1991), 85–106.

[30] Yasmin B Kafai. 1998. Children as designers, testers, and evaluators of educational software. In *The design of children's technology.* , , 123–145.

[31] Yasmin B Kafai. 2012. *Minds in play: Computer game design as a context for children's learning.* Routledge, .

[32] Yasmin B. Kafai and Idit Harel. 1991. Children's learning through consulting: When mathematical ideas, programming knowledge, instructional design, and playful discourse are intertwined. In *Constructionism.* Ablex Publishing, , 85–110.

[33] Magnus Høholt Kaspersen, Karl-Emil Kjær Bilstrup, Maarten Van Mechelen, Arthur Hjort, Niels Olof Bouvin, and Marianne Graves Petersen. 2022. High school students exploring machine learning and its societal implications: Opportunities and challenges. *International Journal of Child-Computer Interaction* , (2022), .

[34] Alan C Kay. 1996. The early history of Smalltalk. In *History of programming languages—II.* , , 511–598.

[35] Michelle S Lam, Mitchell L Gordon, Danaë Metaxa, Jeffrey T Hancock, James A Landay, and Michael S Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–34.

[36] Michelle S Lam, Ayush Pandit, Colin H Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–37.

[37] Dev Raj Lamichhane, Janet Read, and Scott Mackenzie. 2023. When Children Chat with Machine Translated Text: Problems, Possibilities, Potential. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.* , , 198–209.

[38] Victor R Lee and Deborah A Fields. 2017. A rubric for describing competences in the areas of circuitry, computation, and crafting after a course using e-textiles. *The International Journal of Information and Learning Technology* 34, 5 (2017), 372–384.

[39] Florence Kristin Lehnert, Jasmin Niess, Carine Lallemand, Panos Markopoulos, Antoine Fischbach, and Vincent Koenig. 2022. Child–Computer Interaction: From a systematic review towards an integrated understanding of interaction design methods for children. *International Journal of Child-Computer Interaction* 32 (2022), 100398.

[40] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems.* , , 1–16.

[41] John H Maloney, Kylie Peppler, Yasmin Kafai, Mitchel Resnick, and Natalie Rusk. 2008. Programming by choice: urban youth learning programming with scratch. In *Proceedings of the 39th SIGCSE technical symposium on Computer science education.* , , 367–371.

[42] Giulia Mascagni. 2018. From the lab to the field: A review of tax experiments. *Journal of Economic Surveys* 32, 2 (2018), 273–301.

[43] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–23.

[44] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. 2021. An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[45] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344.

[46] Luis Morales-Navarro and Yasmin B Kafai. 2023. Conceptualizing Approaches to Critical Computing Education: Inquiry, Design, and Reimagination. In *Past, Present and Future of Computing Education Research: A Global Perspective.* Springer, , 521–538.

[47] Luis Morales-Navarro, Meghan Shah, and Yasmin B Kafai. 2024. Not Just Training, Also Testing: High School Youths' Perspective-Taking through Peer Testing Machine Learning-Powered Applications. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1.* , , .

[48] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[49] Jean Salac, Rotem Landesman, Stefania Druga, and Amy J Ko. 2023. Scaffolding Children's Sensemaking around Algorithmic Fairness. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.* , , 137–149.

[50] Jean Salac, Alannah Oleson, Lena Armstrong, Audrey Le Meur, and Amy J. Ko. 2023. Funds of Knowledge used by Adolescents of Color in Scaffolded Sensemaking around Algorithmic Fairness. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1* (Chicago, IL, USA) *(ICER '23)*. Association for Computing Machinery, New York, NY, USA, 191–205. https://doi.org/10.1145/3568813.3600110

[51] Mike Scaife and Yvonne Rogers. 1999. Kids as informants: Telling us what we didn't know or confirming what we knew already. *The design of children's technology* , (1999), 27–50.

[52] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[53] Bruce L Sherin, Moshe Krakowski, and Victor R Lee. 2012. Some assembly required: How scientific explanations are constructed during clinical interviews. *Journal of Research in Science Teaching* 49, 2 (2012), 166–198.

[54] Cynthia Solomon, Brian Harvey, Ken Kahn, Henry Lieberman, Mark L Miller, Margaret Minsky, Artemis Papert, and Brian Silverman. 2020. History of logo. *Proceedings of the ACM on Programming Languages* 4, HOPL (2020), 1–66.

[55] Jaemarie Solyst, Shixian Xie, Ellia Yang, Angela EB Stewart, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2023. "I Would Like to Design": Black Girls Analyzing and Ideating Fair and Accountable AI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* , , 1–14.

[56] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 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 (2023), 1–27.

[57] Tiffany Tseng, Matt J Davidson, Luis Morales-Navarro, Jennifer King Chen, Victoria Delaney, Mark Leibowitz, Jazbo Beason, and R Benjamin Shapiro. 2024. Co-ML: Collaborative Machine Learning Model Building for Developing Dataset Design Practices. *ACM Transactions on Computing Education (TOCE)* , (2024), .

[58] Tiffany Tseng, Jennifer King Chen, Mona Abdelrahman, Mary Beth Kery, Fred Hohman, Adriana Hilliard, and R Benjamin Shapiro. 2023. Collaborative Machine Learning Model Building with Families Using Co-ML. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.* , , 40–51.

[59] Sepehr Vakil and Maxine McKinney de Royston. 2022. Youth as philosophers of technology. *Mind, Culture, and Activity* 29, 4 (2022), 336–355.

[60] Maarten Van Mechelen, Rachel Charlotte Smith, Marie-Monique Schaper, Mariana Tamashiro, Karl-Emil Bilstrup, Mille Lunding, Marianne Graves Petersen, and Ole Sejer Iversen. 2023. Emerging technologies in K–12 education: A future HCI research agenda. *ACM Transactions on Computer-Human Interaction* 30, 3 (2023), 1–40.

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

[62] Henriikka Vartiainen, Tapani Toivonen, Ilkka Jormanainen, Juho Kahila, Matti Tedre, and Teemu Valtonen. 2021. Machine learning for middle schoolers: Learning through data-driven design. *International Journal of Child-Computer Interaction* 29 (2021), 100281.

[63] Iro Voulgari, Marvin Zammit, Elias Stouraitis, Antonios Liapis, and Georgios Yannakakis. 2021. Learn to machine learn: designing a game based approach for teaching machine learning to primary and secondary education students. In *Interaction design and children.* , , 593–598.

[64] Raechel Walker, Eman Sherif, and Cynthia Breazeal. 2022. Liberatory Computing Education for African American Students. In *2022 IEEE Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT).* IEEE, , , 85–89.

[65] Jason Yip, Kelly Wong, Isabella Oh, Farisha Sultan, Wendy Roldan, Kung Jin Lee, Jimi Huh, et al. 2023. Co-design Tensions Between Parents, Children, and Researchers Regarding Mobile Health Technology Design Needs and Decisions: Case Study. *JMIR Formative Research* 7, 1 (2023), e41726.

[66] Abigail Zimmermann-Niefield, Shawn Polson, Celeste Moreno, and R Benjamin Shapiro. 2020. Youth making machine learning models for gesture-controlled interactive media. In *Proceedings of the interaction design and children conference.* , , 63–74.

[67] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K Kane, and R Benjamin Shapiro. 2019. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM international conference on interaction design and children.* , , 121–132.

[68] Lauren Zito, Jennifer L Cross, Bambi Brewer, Samantha Speer, Michael Tasota, Emily Hamner, Molly Johnson, Tom Lauwers, and Illah Nourbakhsh. 2021. Leveraging tangible interfaces in primary school math: Pilot testing of the Owlet math program. *International Journal of Child-Computer Interaction* 27 (2021), 100222.

A CODEBOOK

anthropomorphizing

- Definition: Attaching human characteristics to the model.
- Example: "It's thinking this is a shark" Richard

auditing.breakIt

- Definition: Indicating that auditing involves finding moments when the project breaks.
- Example: "Challenging it [the project] to see what would break it " Iván

auditing.newPerspectives

- Definition: Voicing that auditors provide new perspectives on how the project works.
- Example: "it is helpful to get feedback and other perspectives from other people. Who may see things that we have not seen." Luke

auditing.nextSteps

- Definition: Coming up with next steps to improve model performance.
- Example: "Add more variety in the data. Like these are all the same kind of rabbit." Kayla

auditing.noticePatterns

- Definition: Identifying patterns in the outputs to build explanations for model behaviors.
- Example: "which I feel like that's kind of customary for DALL-E because I feel like that the images it's been fed with are more probably just white people instead of diverse like images." Iván

auditing.applyingKnowledge

- Definition: Indicating that insights gained while auditing peers' projects can be helpful to improve their own projects.
- Example: "I felt like just going in, and being able to see other people's projects and see what they can improve on. It's like, what can I improve on it is like, your telling somebody else, what they can improve on and you can turn around and improve that yourself." Jerome

audtiting.placeResponsibility

- Definition: Referring to something or someone as responsible for the outputs.
- Example: "I think it shows more of like human bias than like, AI bias maybe that that's like what like the like, because if it's like trained off of like pictures, like you were showing, and that's kind of like the pictures that it's been like that it's seen being put up on the internet." Jackie Star

bias.age

- Definition: Identifying potential age related biases in model outputs.
- Example: "it's all like let's say young woman like for like the other ones there's like more young people but like they should I feel like they should I add like more older people are like mid-age these are these people look really young" Walter

bias.appearance.body

- Definition: Identifying potential body appearance related biases in model outputs.
- Example: "it's all like let's say young woman like for like the other ones there's like more young people but like they should I feel like they should I add like more older people are like mid-age these are these people look really young" Walter

bias.appearance.body

- Definition: Identifying potential fashion related biases in model outputs.
- Example: "like the general like, in the lab, mixing stuff together, lab coats, goggles gloves. Again, is mainly just just white people." Jerome

bias.breed

- Definition: Identifying potential breed related biases in model outputs.
- Example: "Yeah like the Corgi you can see is all... the Corgi is [identified as] a dog and the German Shepherd as the cat. Why would it think that?" Kayla

bias.color

- Definition: Identifying potential breed related biases in model outputs.
- Example: "Yeah like the Corgi you can see is all... the Corgi is [identified as] a dog and the German Shepherd as the cat. Why would it think that?" Kayla

bias.gender

- Definition: Identifying potential gender related biases in model outputs.
- Example: "The models were all white girls with straight hair.." Jackie Star

bias.location/context

- Definition: Identifying potential location/context related biases in model outputs.
- Example: "All the dolphins are jumping on the water none of this whales are so it's creating a bias towards the dolphins because they're all technically in the same act that having the same like actions and as dolphins." Fatimah

bias.position

- Definition: Identifying potential position related biases in model outputs.
- Example: "These cats are mainly perched up. Yeah, I feel like in the same position." Emily

bias.race

- Definition: Identifying potential race related biases in model outputs.
- Example: "For the doctor that was always white male, or any other professions like lawyer teacher well teachers mean female or like librarians female but it was majority white people and no black or any race." Fatimah

bias.relevancy

- Definition: Identifying potential relevancy related biases in model outputs.
- Example: "This was probably how what a good student was like defined as back in the day." Stephanie

bias.shape

- Definition: Identifying potential shape related biases in model outputs.
- Example: "Because I had just like a prominent shape like yeah, shape or triangles and cones. It's easy to see compared to the other two, which can be kind of a bit similar. It made the [blackberry] look like strawberries." Kayla

bias.size

- Definition: Identifying potential size related biases in model outputs.
- Example: "this thing just thinks everything that has big ears as a cat." Richard

bias.socioeconomic

- Definition: Identifying potential biases related to socioeconomic status in model outputs.
- Example: "They're all like white again, like probably middle class... upper middle class white people... mainly white men or boys I guess. " Kayla

dataDesign.classBalance

- Definition: Inferring data composition related issues in the design of training datasets.
- Example: "I would add more pictures of like sharks. Like zoomed out like the whole body color, good lighting and the whale add more like out of water where it is jumping maybe like the dolphins nothing because it seems to get like a dolphins pretty accurately." Walter

dataDesign.context

- Definition: Inferring data context related issues in the design of training datasets.
- Example: "There may be like hands involved or like, like a background, like a lot of greenery so it knows what's the difference... because this one's just covered, the background is just white. So there's nothing, nothing in the background." Walter

dataDesign.diversity

- Definition: Inferring data diversity related issues in the design of training datasets.

- Example: “More representation definitely just different environments different people different skin colors, races genders and even for this maybe like same sex couples that could be something” Fatimah

dataDesign.edgeCase

- Definition: Identifying potential edge cases.
- Example: “I don’t think it’s really used to like pencils or markers that are like black and white because all the other examples are brightly colored yeah.” Jackie Star

dataDesign.lighting

- Definition: Inferring lighting related issues in the design of training datasets.
- Example: “Like better lighting for this one, because they’re all like underwater.” Kayla

dataDesign.quantity

- Definition: Inferring data quantity related issues in the design of training datasets.
- Example: “We would add more hands to holding new markers, we will do some darker markers. There’s only like two darker markers.” Emily

dataDesign.source

- Definition: Inferring issues related to the sources of data used in training datasets.
- Example: “Maybe based off of like stock images, like that’s what it was trained on.” Kayla

ID.failure.explanation

- Definition: Identifying cases in which the model did not perform as expected and providing an explanation for such performance.
- Example: “Yeah, it’s so they go based off of the sort of since the cat is white. It goes to the bunny since all of the pictures of the bunny are white.” Luke

ID.failure.noExplanation

- Definition: Identifying cases in which the model did not perform as expected without providing an explanation for such performance.
- Example: “That’s not a shark.” Luke

ID.success

- Definition: Identifying cases in which the model performs as expected.
- Example: “They all have like the big ears. I think that’s why they got these two right because they have big dog ears.” Kayla

justice.exclusion

- Definition: Arguing that system outputs exclude some people.
- Example: “Because people can look at this, these pictures of doctors and rich people and see that they look nothing like them? And then feel discouraged? As if Oh, that’s not an opportunity that I can convey?” Luke

justice.harmful

- Definition: Arguing that system outputs produce harm.
- Example: “Mainly the librarian one is harmful because it shows a bunch of women... it’s not like men can be librarians, "don’t do that... it’s not a masculine job, you shouldn’t have that job if you’re a man." I guess it’s kind of saying.” Kayla

justice.notHarmful

- Definition: Arguing that system outputs do not produce harm.
- Example: “I think if you’re getting harmed by an AI, I don’t know. That’s more of a personal problem.” Richard

justice.potentiallyHarmful

- Definition: Arguing that system outputs could be potentially harmful.
- Example: “I feel like at this point right now, it’s not harmful but like, as it evolves, it will be is like beautiful. I feel like no one’s actually going to compare themselves to these awful images, and be like, Wow, if I don’t if I don’t look like this person, my God and of the world, but I feel like as it evolves, it will be more harmful if these issues aren’t addressed by adding more diversity.“ Iván

modelDesign.features

- Definition: Inferring parameters or features used by the model.
- Example: “I don’t think it’s taking in color, it doesn’t care if the strawberry is white or whatever the color of strawberries it’s more looking at the texture.” Richard

priorExperiences.personalBias

- Definition: Considering personal biases in evaluating system behaviors.
- Example: “It makes sense that it will all be woman. I’ve personally never heard of a male librarian... I really haven’t.” Kayla

priorExperiences.societalBias

- Definition: Considering societal biases in evaluating system behaviors.
- Example: “What about like cook because I feel like that could go in either direction. Yeah, because it could be stereotyped with women being in the kitchen…” Iván

programmed

- Definition: Indicating that the machine has been programmed to work in a certain way.
- Example: “mean, I’ve seen... I saw like this one thing, I don’t remember where it was, it was like, it asked the chat GPT to do a prompt for a smart scientist. And it just put out that the scientists had to be white. And I found that was kind of interesting, because it was that was the biases that it was programmed with, when that’s not true.” Iván