Learning AI Auditing: A Case Study of Teenagers Auditing a Generative AI Model

LUIS MORALES-NAVARRO, University of Pennsylvania, USA
MICHELLE GAN, University of Pennsylvania, USA
EVELYN YU, University of Pennsylvania, USA
LAUREN VOGELSTEIN, Teachers College, Columbia University, USA
YASMIN B. KAFAI, University of Pennsylvania, USA
DANAÉ METAXA, University of Pennsylvania, USA

This study investigates how high school-aged youth engage in algorithm auditing to identify and understand biases in artificial intelligence and machine learning (AI/ML) tools they encounter daily. With AI/ML technologies being increasingly integrated into young people's lives, there is an urgent need to equip teenagers with AI literacies that build both technical knowledge and awareness of social impacts. Algorithm audits (also called AI audits) have traditionally been employed by experts to assess potential harmful biases, but recent research suggests that non-expert users can also participate productively in auditing. We conducted a two-week participatory design workshop with 14 teenagers (ages 14–15), where they audited the generative AI model behind TikTok's Effect House, a tool for creating interactive TikTok filters. We present a case study describing how teenagers approached the audit, from deciding what to audit to analyzing data using diverse strategies and communicating their results. Our findings show that participants were engaged and creative throughout the activities, independently raising and exploring new considerations, such as age-related biases, that are uncommon in professional audits. We drew on our expertise in algorithm auditing to triangulate their findings as a way to examine if the workshop supported participants to reach coherent conclusions in their audit. Although the resulting number of changes in race, gender, and age representation uncovered by the teens were slightly different from ours, we reached similar conclusions. This study highlights the potential for auditing to inspire learning activities to foster AI literacies, empower teenagers to critically examine AI systems, and contribute fresh perspectives to the study of algorithmic harms.

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

Additional Key Words and Phrases: youth, algorithm auditing, algorithmic justice, responsible AI, participatory design, machine learning, child-computer interaction, artificial intelligence, TikTok, Effect House, generative artificial intelligence

ACM Reference Format:
Luis Morales-Navarro, Michelle Gan, Evelyn Yu, Lauren Vogelstein, Yasmin B. Kafai, and Danaé Metaxa. 2025. Learning AI Auditing: A Case Study of Teenagers Auditing a Generative AI Model. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article 439 (November 2025), 29 pages. https://doi.org/10.1145/3757620

Authors' Contact Information: Luis Morales-Navarro, luismn@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Michelle Gan, michellegan00@gmail.com, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Evelyn Yu, evelynyu@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Lauren Vogelstein, lev2124@tc.columbia.edu, Teachers College, Columbia University, New York, New York, USA; Yasmin B. Kafai, kafai@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Danaé Metaxa, metaxa@upenn.edu, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

ACM 2573-0142/2025/11-ART439
https://doi.org/10.1145/3757620

1 Introduction

Artificial intelligence and machine learning (AI/ML) technologies have become deeply woven into young people's daily lives over the past ten years—from photo filters on social media to recommendations on streaming platforms to voice assistants. This widespread adoption of AI/ML creates an urgent need to support teenagers in developing AI literacies [29]. Young people must gain knowledge not only about how these technologies function but also about their potential societal impacts. Most critically, teenagers need the skills to identify algorithmic biases and take action when these systems cause harm. While research has shown that teenagers are able to identify algorithmic bias and harms [36, 54], translating these insights into approaches to further systematic and empirical examinations of algorithmic systems remains an open challenge. As Solyst et al. [54] note, we lack sufficient studies examining how to cultivate and harness young people's valuable insights for identifying and reducing algorithmic harm. This gap reflects a broader tendency to underestimate teenagers' capacities to grasp both the technical complexities and the ethical implications of AI/ML technologies, leading instead to attempts to limit their engagement [53].

One efficient method that experts have developed to examine and draw conclusions about AI/ML systems, particularly in regard to potential harmful biases and discrimination, is algorithm auditing [35]. More recently, scholars have begun to examine how everyday people can engage in algorithm auditing, reframing auditing as a way for the public to gain insights about algorithmic behaviors in everyday contexts [51]. In this paper, we explore the potential for engaging teenagers in full-fledged auditing activities to investigate potentially harmful algorithmic biases on a popular social media platform.

We present a descriptive case study of a participatory design workshop in which teenagers engaged in auditing the generative AI model that powers TikTok filters. This model is accessible through Effect House, a filter development environment that supports the creation of text-to-image filters, the same filters teens encounter on TikTok. We conducted the workshop with a group of 14 teens (ages 14–15) in a two-week summer program. Workshop activities were designed to support participants in systematically investigating potentially harmful algorithmic biases. Rather than becoming experts in auditing, our intention in this workshop was for young people to lead the process of evaluating an algorithmic system they encounter on a daily basis. In addition to describing participants' experiences throughout the auditing process, we also triangulated their findings to examine if the workshop design supported participants to identify potentially harmful algorithmic biases in their audit. We address the following research questions: (1) How did participants engage in this algorithm auditing activity? In particular, what choices did they make, how did these choices vary across participants, and what reflections did they have throughout the process? and (2) Did the workshop support participants to reach evidence-based, credible conclusions in their audit?

Our case study demonstrates that with adequate scaffolding, teenagers can participate in full-fledged audits of real-world algorithmic systems they encounter in their daily lives. We observed participants making connections to their everyday experiences and understandings of biases; for example, they selected inputs based on the social dynamics they observed in their own communities and contributed ideas about age-related biases that are uncommon in professional audits but were particularly salient to them. By triangulating participants' findings, we were able to confirm that the workshop design may be conducive to supporting teens to conduct audits with evidence-based, credible conclusions.

We discuss the design of audit-based learning activities and the role that these can play in supporting the development of AI literacies for young people. Further, we explore this case study as confirming the value of involving non-experts in auditing algorithmic systems with which they are personally familiar. Our paper makes the following contributions: (1) we illustrate one process

by which teenagers engage in a scaffolded auditing learning activity to empirically investigate potentially harmful behaviors in a real-world algorithmic system, (2) we provide evidence from a case study about the deployment of one such activity in practice, and (3) we reflect on participants' experiences with the activity, situating their approaches, decisions, and conclusions in relation to more traditional and participatory auditing procedures.

2 Background & Related Work

In this section, we position this study in relation to other studies about teenagers and algorithmic justice, algorithm auditing as a whole, and participatory approaches to auditing. We also provide relevant background on the specific domain our teen participants audited: identity representation of different occupations as reflected through the lens of algorithmic systems—a popular domain for prior expert audits.

2.1 Teens and Algorithmic Justice

In the last ten years, computing education and CSCW researchers have recognized the importance of engaging teenagers in learning activities that investigate "the consequences, limitations, and unjust impacts of computing in society" [23]. These activities typically involve reframing computing as a sociotechnical field by prompting learners to explore the functionality of computing systems and their implications [37, 54]. Such activities are often designed for young people to assess inputs and outputs of computing systems, investigate how systems are actually used, and consider how they affect people and the environment [14]. However, most efforts to engage teenagers in being critical about computing tend to focus on discussion or direct instruction without providing opportunities for learners to empirically investigate issues of justice and ethics in computing [38, 54].

Limited research has examined how young people actively investigate issues of algorithmic justice—how they understand, explain, and investigate the potential for algorithmic systems to perpetuate harm [6]. Researchers have used participatory design (PD) workshops to study how young people think about these issues [11, 53]. For instance, studies have found that while teenagers may be aware of the negative impacts of technology in their everyday lives, they may not use the word "bias" [11]. Other work has highlighted that teenagers may view AI/ML systems as being bias-free or view bias favorably when it enhances their own user experience and negatively when it restricts it [27]. Salac et al. [46] noted that when evaluating scenarios in which issues of algorithmic justice were presented, teenagers considered their own lived experiences, how systems may behave in different contexts, and larger societal issues that mediate the use of the systems. While these studies investigate young people's perceptions of issues related to algorithmic justice, they do not address how young people can be engaged in investigating these issues themselves.

2.2 Algorithm Auditing

This paper addresses a gap in the literature on teenagers and algorithmic justice by exploring how algorithm auditing activities may be used to engage teenagers in empirical investigations of algorithmic justice issues. Auditing algorithmic systems 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" [35]. In contrast to other forms of evaluation, auditing aims to draw system-wide conclusions rather than scoping its conclusions to a specific set of test cases. Additionally, audits are often external evaluations conducted by third parties without insider access or knowledge. Audits are traditionally conducted by experts seeking to evaluate how AI/ML-powered systems behave across different application areas (housing, healthcare, social media, search) and draw inferences about the potential harmful impact of these systems (for a review of expert algorithm audits, see [3]). Although the specific procedures are tailored to each audit, auditing

usually involves (1) developing a hypothesis about a specific system behavior, (2) generating a set of systematic, thorough, and thoughtful inputs to test the hypothesis, (3) running the tests, (4) analyzing the data, and (5) reporting the results (for more details on the method, see [35]).

2.3 Participatory Approaches to Auditing

Engaging relevant communities in cooperative and participatory design (PD) activities to design and evaluate computing systems has a long tradition in CSCW research [4, 64]. Recently such approaches have been adopted in efforts to design and evaluate AI/ML systems in community-engaged ways [44, 56]. Such PD activities can be structurally open or closed with participants having different levels of autonomy to define their engagement [64].

Researchers have begun exploring the potential for non-experts to engage in identifying potentially harmful algorithmic behaviors through auditing-like practices. This can be done under frameworks known as *crowdsourced*, *everyday*, or *end-user audits* [24, 48, 51]. In some cases, researchers have provided non-expert adults with user-friendly interfaces to scaffold auditing tasks that harness users' personal experiences [24], while in others users contribute data without actively participating in the larger auditing endeavor [25]. Other literature observes that users may sometimes engage organically with auditing practices by evaluating algorithmic systems in their everyday lives in the absence of experts [13, 51].

2.3.1 Teens and Algorithm Auditing. Several efforts have started to explore different participatory approaches to engage teenagers in auditing-related activities. Solyst et al. [54] adapted user-driven everyday auditing tasks, such as making observations of Google search results, into activities for middle schoolers. In their study, researchers presented participants with bias identification scenarios from Google search and DALL-E-generated images, finding that most participants identified race-related biases as harmful. Participants in this study also argued that users should have the ability to report problematic behaviors and that AI systems should warn users of potential harm. Morales-Navarro et al. [36] designed a workshop in which high school participants designed ML-powered physical computing projects and then audited their peers' projects. They found that when externally evaluating motion classifiers, participants were able to identify unexpected biases. In doing so, participants also inferred dataset and model design issues that could cause such biases. However, the auditing activities in these studies primarily focused on having participants make a single or a few observations about a system's behavior, rather than the systematic analysis of inputs and outputs and the full process of conducting an audit from beginning to end. We expand on this research by engaging teenagers in a full-fledged audit of a real-world algorithmic system, specifically investigating how teenagers engaged in auditing gender and racial representation in the outputs of the generative image model of TikTok's Effect House. Iversen et al. [19] argue that young people can also engage in such participatory activities by becoming *protagonists* or the main decision-makers. In our study, we positioned participants as protagonists, enabling them to decide what to audit and how to analyze their data.

It is worth noting that this kind of PD work does not happen in a vacuum, as researchers set parameters and create basic structures that orient participant engagement [64]. The role of researchers in such activities is that of encouraging and supporting young people to "be the main agents in driving the design process and thereby to develop skills to design and reflect on technology and its role in their lives" [19]. Here, researchers can be *reflective practitioners* that analyze the design process and its outputs [4, 50]. When PD is related to a learning or educational intervention, reflecting on the design process often involves examining if the PD sessions are conducive to learning [8, 42]. This is particularly important as PD could lead to misinformation and disinformation, presenting risks "by platforming false or hateful ideas and allowing them to gain

impact" [49]. As such, in our study we triangulate participants' findings in an effort to examine if the design of the PD workshop supported teens to reach evidence-based, credible conclusions.

2.4 Representation of Occupations in Algorithmic Systems

In our study, teenage participants conducted an audit of identity representation in occupations. As we will describe in a later section, they came to this goal after their own independent explorations with the tool. Notably, this is a popular domain for prior audits; several expert and non-expert audits have investigated gender and racial representation of occupations in image search and image generation.

2.4.1 Expert Audits on Representation in Occupations. About a decade ago, Kay et al. [22] conducted an audit on gender representation in Google image search results for common occupations (e.g., "doctor", "engineer") that found systematic underrepresentation of women. A follow-up study by Metaxa et al. [35] investigated gender and racial representation for image search results, again finding evidence of men's systematic overrepresentation, as well as the overrepresentation of White people in image search results. These two studies established common methodologies for conducting algorithm audits of gender and racial representation of occupations that have been followed and replicated in studies of generative AI image models. Such studies have shown systemic amplification of racial and gender disparities in the representation of occupations in image outputs across different models (Dall-E2, Stable Diffusion v1.4 and v2) [5, 30, 40].

2.4.2 Adult and Teen End-user Audits on Representation in Occupations. Building on expert audit research, DeVrio et al. [13] investigated how non-expert adults identified harmful behaviors in algorithmic systems such as Google image search. They conducted interviews in which users were tasked to search for words such as "librarian" or "thug" to prompt them to explain how they thought about potential harmful biases in search results. Following, when asked to look for other cases of harmful algorithmic behaviors, participants conducted their own searches, observing racial and gender biases in the representation of occupations such as "computer scientist", "maid", and "firefighter." This study found that users' experiences and exposure to societal biases influence their strategies for conducting image searches and how they interpret the results.

These tasks have been adapted to study how teenagers engage in auditing-like practices. For example, Solyst et al. [54] conducted a PD workshop where teenagers analyzed search results for images of computer programmers and Dall-E-generated images of doctors. Morales-Navarro et al. [36] used similar tasks in a pre/post interview study to assess how teenagers' identification of potential algorithmic biases and harms changed after participating in peer-auditing activities. These studies found that teenagers were able to recognize and explain potentially harmful algorithmic biases in the representation of occupations and that teenagers were concerned that the stereotypes in the representation of occupations may discourage young people from pursuing certain careers. Furthermore, these two studies highlight the potential of involving teenagers as contributors in the evaluation of algorithmic systems that they use in their everyday lives.

While previous studies engage teens in auditing-like tasks, they stop short of supporting teens in conducting end-to-end, full-fledged algorithm audits. Our study addresses this gap by presenting a case study in which teenagers investigate gender, race and age biases in the representation of occupations in the generative image model used in TikTok filters.

3 Methods

In this section, we describe the participants and context of our study, our data collection and analyses processes, and our research team's positionality in conducting this work.

3.1 Participants

This study was conducted in the context of a series of iterative participatory design workshops with teenagers that investigated the potential of developing algorithm auditing learning activities (for details on other iterations of this work, see [39, 60]). We worked with teenagers enrolled in a four-year afterschool program, STEM Stars (a pseudonym), at a science center in the Northeastern United States. As part of the STEM Stars program, we provided workshops on algorithm auditing in the fall, spring, and summer of 2024. Participants were invited to take part in the study through emails and texts sent out by the science center staff. Guardians filled out consent forms before their children participated in the study, and minors assented to participate. The Institutional Review Board of the University of Pennsylvania approved the study protocol. The analysis in this paper focuses on our two-week summer workshop with 14 teens (14-15 years old) in the STEM Stars program. That summer, we worked with six female, one non-binary, and seven male youth. The majority were from marginalized racial backgrounds, with all but one identifying as African American, Asian American, or multiracial. To respect the privacy of our underage participants, all names used in this paper are pseudonyms.

3.2 Context

In this paper, we focus on how participants audited the generative AI model that powers TikTok's filters, which are created in an application called Effect House. We provide context on TikTok and Effect House below.

3.2.1 TikTok Filters. TikTok is a prominent video-sharing social media platform where users create and share short videos. TikTok enables users to record, edit, and remix short videos, which can include generative AI-driven effects or filters. The platform is particularly popular among teenagers, with 58% of teens in the United States reporting using TikTok on a daily basis [2]. Recent documents from court cases show that TikTok estimates that 95% of teens under 17 who have a smartphone use TikTok [1]. Previous research indicates that the app's popularity may be influenced by the strength of its recommendation system [66]. This is particularly relevant for youth, as 17% of teens in the US describe themselves as being on TikTok almost constantly. At the same time, regulators argue that the success of the recommendation system can have a noxious effect on teenagers through compulsive usage [1]. While the details of the recommendation system are not public, users build hypotheses about the system by testing it themselves or replicating collective memes shared on social media [21].

Generative AI filters on TikTok are text-to-image or image-to-image filters that use generative AI models to modify users' photo and video inputs. Both the original and altered content can then be included in the posted video. For an illustration of a filter in action, see Figure 1. The mechanisms behind these filters—and the underlying models that power them—are largely opaque to both users and experts. This lack of transparency raises concerns, as evaluation work of generative AI models at large (outside TikTok) has demonstrated that these models can replicate stereotyped gender attributes [32, 55, 65] and societal biases [40]. Other problematic behaviors include the oversexualization of features and the promotion of unrealistic beauty standards [7].

Research on such filters on other platforms has shown that biased AI-generated changes of facial features can lead to negative self-perception [10, 43] and may increase body dysmorphia among teens [26]. However, the design of such filters is largely unregulated [15]. Although TikTok does publish guidelines for filter designers that explicitly prohibit stereotypes and discrimination [59], accountability mechanisms available to users are few and opaque [15].

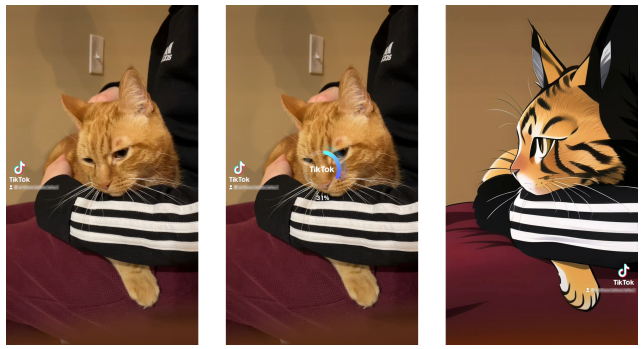

Fig. 1. Example of a TikTok filter in action. This filter uses generative AI to output a manga-style illustration of the input photo.

3.2.2 Effect House Interface. The Effect House development environment includes a visual scripting interface for filter creators to write code (see Figure 2.a), and also provides a prompting interface (Figure 2.b) that enables designers to create filters by writing their own text prompts, which are used to run the default generative AI model provided by TikTok.

Relevant to this study are three key parts of the Effect House interface: (1) the input text prompt, where filter designers can write prompts for their filters, and (2) the filter preview, where designers can input images (Effect House also has some built-in options) to preview an output image that Effect House's generative AI model creates based on the prompt and input image. The interface also includes some togglable parameters, like the "prompt strength" slider (a 0-1 value set to 0.5 by default) that, when increased, exaggerates the degree to which the input image is stylized.

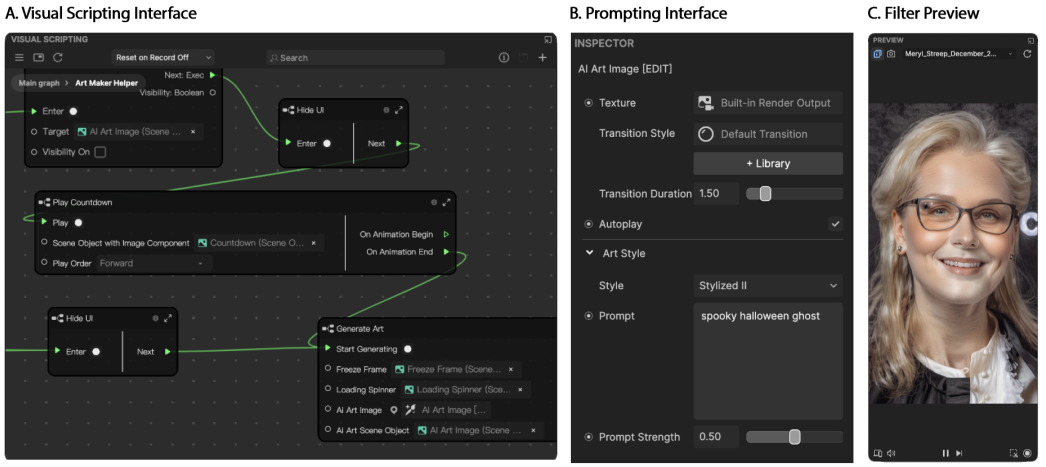

Fig. 2. Effect House's visual scripting, prompting, and filter preview interfaces.

3.3 Workshop Procedure

During the workshop, we met with participants for 28 hours over the course of two weeks. All participants had previously participated in a short 4-hour workshop in which they informally

audited an anime TikTok filter (a real filter available to all users on TikTok). In the first week, teens designed generative AI TikTok filters and audited each other's filters. Even though they were able to systematically evaluate the filters, the number of tests conducted by most pairs of participants was limited (with each group running an average of 30 tests), which impacted their abilities to draw conclusions based on strong evidence. In the second week, we shifted focus to auditing the Effect House tool itself. In this paper, we focus on that second week's activities in which participants collaboratively audited the default generative model built into Effect House. To address the problem of data scale observed in week one, we decided to have participants conduct the audit collaboratively, developing a hypothesis, generating a set of prompt inputs for the audit, and testing prompts as a group. This enabled participants (with the support of researcher-facilitators) to generate 100 prompts that were used in conjunction with 12 input images provided by researchers to run 1200 tests, creating a more robust dataset for analysis. After creating the dataset, participants analyzed the data in small groups of 2-3, with each group producing an audit report to present to the wider group.

Workshop activities were scaffolded around five steps: (1) developing a hypothesis, (2) generating inputs, (3) running the tests, (4) analyzing the results, and (5) creating an audit report (see Table 1). Every workshop day started with a 30-minute game or warm-up activity (e.g., duck duck goose) and included two snack breaks. Since participants had already participated in auditing activities in the first week, there was little to no formal instruction during the second week. Instead, researcher-facilitators provided opportunities for participants to review how audits are conducted and brainstorm ideas for each activity step. Each day was scheduled as follows (see Table 1 for more details):

Day 1. The main researcher-facilitator led an activity for participants to review the auditing process. Afterwards, teens spent time brainstorming hypotheses about the Effect House generative AI model. To get them started in the process of hypothesis formation, we asked participants to spend an hour in small groups trying different prompts and inputs in Effect House and to take notes of observed behaviors they might want to investigate. They shared their hypotheses, and then the main researcher-facilitator proposed a hypothesis that encompassed several ideas presented by the youth: Effect House's image model reinforces gender and race stereotypes about different occupations.

Day 2. Participants began the day brainstorming different ways to test their hypothesis. The main researcher-facilitator prompted participants to come up with a list of occupations they could use to test the hypothesis while researchers created a set of input images (see Figure 4). Participants added occupations to a collaborative list (see Table 2) and brainstormed fill-in-the-blank prompts that could be used to situate the occupations in different contexts. Prompts were designed in a similar way to those used in other studies of generative AI images [30], and included "A [occupation] at work", "A tired [occupation]," "A [occupation] with friends," and "A happy [occupation]". Next, the main researcher-facilitator showed the dataset of input images compiled by the researchers (see Figure 4). The image dataset included four sample images built into Effect House (a Black woman, an Asian man, a White woman, and a racially ambiguous man) and a set of 10 images of celebrities selected from Wikimedia Commons (Maitreyi Ramakrishnan, Tyler James Williams, Elliot Page, Jamie Lee Curtis, Jason Momoa, Harry Styles, Lupita Nyong'o, Peppermint, Karol G, and Awkwafina) with the goal of representing a gender- and racially diverse group of popular figures. Of note, four occupations and three images were not used for testing the hypothesis due to time constraints. Finally, participants spent an hour testing different prompts and images and documenting the image outputs of Effect House's image model. To test the inputs selected in the previous step, participants collaborated as a single group, selecting 25 input occupations with four

different prompts for each and 12 images to test each prompt. This required running 1200 individual tests. To keep track of the tests and scaffold the process, the researchers created a spreadsheet on Miro, an online collaborative whiteboard tool (as visualized in Figure 5), that allowed participants to download input images, upload output images, and keep track of the tests. All the tests were conducted with consistent settings to control for possible output differences resulting from those settings (Prompt Strength: 0.50, Style: Stylized II, Transition Style: Style 1).

Day 3. Participants continued running tests using the prompts from the day before. After completing all the tests, the main researcher-facilitator invited participants to form small groups to analyze the data. She encouraged them to create a table to record their analysis by breaking the data into categories and keeping track of any changes they observed between input and output images. She emphasized that they could think about how to describe any changes they observed between inputs and outputs and how to quantify these changes.

Day 4. The last day of the workshop was focused on creating audit reports. Participants began by brainstorming with whom they wanted to share their findings and preferred formats for sharing their findings. They then created reports in the form of videos and slideshows. Finally, they shared the reports with the group.

3.4 Research Approach and Positionality Statement

For this study, we worked with teenagers from traditionally underrepresented identities in computing. Doing such work in an ethical manner requires centering the needs of the community and extensive engagement with participants. Before conducting the study, we worked with a youth advisory board comprised of seven 15- to 17-year-olds to brainstorm learning activities. Similarly, science center educators were involved in the brainstorming sessions and reviewed the activities prior to the workshop. Our team is invested in continuing our long-standing relationship with the science center and always aims to engage sustainably and respectfully in order to do so; one of the authors has worked with participants from this center for 15+ years and another for over five.

We recognize that using technologies developed by TikTok can present a risk for minors. During the workshop, participants did not use their personal devices or personal accounts. Instead, we partnered with the science center to provide participants with project phones, computers, and TikTok accounts. The project TikTok accounts were private, and researchers as well as science center staff were present during all interactions with the activity devices.

We acknowledge that our own identities and backgrounds impact and prepare us for the research we do. We hold identities representing at least four different racial/ethnic backgrounds and three gender identities, as well as academic backgrounds in the learning sciences and human-computer interaction (HCI). The majority of our team lives in the same city, where our participants live and where the science center is located. Our qualifications—including expertise running expert audits, teaching high school youth, and designing learning environments—prepared us to conduct this study effectively and responsibly.

3.5 Data Collection and Analysis

During the workshop, we collected four primary sources of data: recordings of image and video artifacts participants created (e.g., pictures of brainstorming papers, audit reports in the form of videos), screen recordings of their work on project computers and phones, the actual files of the collaborative audit (a spreadsheet with inputs and outputs of 1200 tests; see Figure 5), and researcher field notes. We analyzed this data in two different ways for this paper: first, by creating a case study of how the group collectively conducted the audit, and second, by further analyzing the dataset created by participants in order to triangulate their findings. We discuss these in more detail next.

Table 1. Workshop Activities by Day

Activity	Time	Description
		Day 1
Auditing steps activity	40 min	Researcher-facilitator led a an activity where participants were divided into five groups and each group had to come up with a skit explaining what each step involved.
Step 1: Developing a Hypothesis	15 min	Participants talked amongst themselves about what kind of hypothesis they could create about Effect House's image model
Step 1: Sharing Hypothesis	20 min	Participants shared their hypotheses with the group; researcher-facilitator proposed a hypothesis that encompassed several ideas presented by the youth
		Day 2
Brainstorming ideas for testing	30 min	Participants talked about how they could potentially test the hypothesis
Reviewing 5 steps of auditing	30 min	Researcher-facilitator reviewed the 5 steps of auditing by showing examples from participants' work from the previous week
Why do we audit?	10 min	Facilitators led a discussion that centered on why and how we audit AI/ML systems
Step 2: Generating prompts	45 min	Participants brainstormed prompts that they can put in the Effect House's prompt interface to test their hypothesis. They created a list of occupations.
Step 3: Testing	60 min	Participants tested the prompts on 12 different photos of different people
		Day 3
Step 3: Testing	90 min	Participants tested the prompts on 12 different photos of different people
Step 4: Brainstorming ideas for analysis	20 min	Participants brainstormed about how they could analyze data from the tests
Step 4: Conducting the analysis	60 min	Participants analyzed the outputs and put their findings in Miro Board
		Day 4
Step 5: Brainstorming report ideas	30 min	Participants talked about how they could report their audits and possible audiences
Step 5: Audit report examples	20 min	Facilitators showed examples of audit reports participants created the previous week and examples expert-led reports
Step 5: Creating audit reports	90 min	Participants made their own audit reports in groups
Step 5: Sharing audit reports	30 min	Each group presented their report to their peers

3.5.1 Case Study. To investigate how participants collaboratively audited Effect House's image model, we constructed a descriptive case study. We decided to use this type of case study due to the exploratory nature of the work and because this kind of case is particularly useful to describe in detail context-specific activities [62], in our case participating in an auditing PD workshop. The main objective of this type of case study is to describe a phenomenon without aiming to provide explanations for the phenomenon or considering rival explanations [63]. Such descriptive narratives

are common in qualitative CSCW and HCI research [33]. For example, descriptive case studies have been used in HCI to illustrate the context-specific nature of PD research with youth [16].

Descriptive case studies are often organized around a descriptive framework that focuses the analysis on specific activities or topics. In our analysis, we use the five steps of the auditing process as a descriptive framework. We began analysis by reviewing the audit reports that participants created [63]. From the reports, we noticed a range of diverse observations and heuristics for analysis (e.g., different approaches to annotate race and gender), which sparked our interest in how participants collectively conducted the audit. To create this case study, three researchers watched 25 hours of screen recordings from the four days participants spent auditing Effect House to create videologs that documented participants' activities every two minutes. These videologs were then organized using the descriptive framework. Following, we produced analytic memos documenting participants' engagement with each step of the auditing process. Finally, we triangulated our findings with researcher field notes and video recordings of group discussions and conversations.

3.5.2 Triangulating Participants' Findings. A major question arising during our observation of participants' auditing was if the design of the workshop supported them to reach evidence-based, credible conclusions. The purpose of the workshop was to support teens in making inferences about the actual systems that they use in their everyday lives, and having them reach inaccurate conclusions could be problematic and misleading. CSCW studies that center on learning or educational interventions often examine whether the interventions of the studies are conducive to learning [28, 42, 52]. One way of approaching this is by comparing how non-experts and experts or people of different expertise complete a task. This is a common approach in the CSCW literature [17, 58, 61]. As such, we decided to triangulate the findings of our participants to see if we also reached the same conclusions. Triangulating participants' findings was crucial not only to examine if this is a good way to scaffold youth in algorithm auditing but also to conduct our research responsibly, as PD may have the risk of platforming and legitimizing false ideas [49].

In HCI studies, triangulating findings is not uncommon as a way to increase validity in empirical research [47]. At large, triangulation involves using multiple methods, investigators, and data sources to investigate a single phenomenon. Denzin [12] describes that triangulation may involve relying on different datasets, having different researchers with complementary expertise analyze the data, analyzing a phenomenon from different theoretical perspectices or using different methods to analyze the same phenomenon. In this study, we approach triangulation by having youth and researchers analyze the same data.

We drew upon our own expertise in conducting audits [31, 34, 35] and qualitatively coded the dataset created during the PD workshop along the same axes as the youth: gender, age and race. For instance, for gender, we annotated output images for evidence of gender exaggeration (outputs being more masculine-presenting, more feminine-presenting, or the same when compared to inputs), facial hair (presence/absence), and mascara or blush in output images (presence/absence). For age, we annotated for the presence of wrinkles (presence/absence) and gray hair (presence/absence). And for race, we annotated for changes in skin complexion (outputs having lighter, darker, or the same skin complexion as inputs) and hairstyle (curlier, straighter, or the same as inputs, as well as the presence or absence of fade hairstyles) in output images. Next, three authors collectively coded an initial 20 images, discussing how and why each of the codes was applied. Then, each researcher coded the same 240 images (20% of the data), achieving 76.25% to 92.08% agreement across all categories but skin complexion (50.42% agreement) and gender exaggeration (50.83% agreement). Due to the difficulty in agreeing on codes for these two categories, we replaced them, instead coding for change in gender representation (when the person in the output image appeared to be a different gender than the person in the input) and change in racial representation (when

the input and output images were perceived as belonging to different racial groups). Following, we coded the same 240 images, achieving 82.50% agreement for changes in racial representation and 81.25% agreement for changes in gender representation. This resulted in substantial agreement among coders with an average inter-rater reliability of Fleiss's $\kappa = 0.65$, 95% CI (0.58-0.72) across code categories. Finally, the same three researchers coded the remaining 960 output images. The final coding scheme is provided in Appendix A.

4 Findings

We separate our findings into two main sections, one for each research question. In the first, we describe the results of our case study, documenting participants' engagement with each of the five activities involved in the participatory design workshop. In the second, we triangulate participants' findings by presenting our own analysis of their dataset.

4.1 Teenagers Auditing Effect House's Image Model

In this section, we describe the five activities participants participated in, each one focused on a separate step of the auditing process. We recount how, with the support of researcher-facilitators, participants (1) came up with a hypothesis and (2) a set of inputs to test the hypothesis, (3) ran tests, (4) analyzed data, and (5) created reports to share their findings. Here we address our first research question: **How did participants engage with this algorithm auditing activity? In particular, what choices did they make, how did these choices vary across participants, and what reflections did they have throughout the process?**

4.1.1 Hypothesis Formation. Participants explored the tool's functioning widely; notably, some groups narrowed in on issues commonly studied in audits, like the representation of race and gender, without being guided to do so.

Unrelated to social bias, teams explored AI behaviors like the generative AI tool's behavior on datasets of animals and its response to images in different orientations. Selena and Twyla investigated how the system processed input images of pets. Twyla noted that pets may be misrepresented in the output images, explaining that "when I put in a dog, it made it a cat." Horacio experimented with different prompts that would change the color of people's clothes, noting that sometimes these did not work as expected. Ziyi noticed that the system only generated images for pictures that were "upside up," and when input images included people upside down, it did not work. Such experiments are reflective of the way participants began to creatively and broadly detect and interrogate unexpected system behaviors.

Without being prompted to do so, other groups spent their time uncovering potential issues related to the representation of race and gender, much like formal audits of such tools have [20, 41]. Interestingly, some participants set out with the goal of finding such biases, while others noticed and pursued these investigations in the course of their unrelated exploration of the tool. Ibrahim and Dalia were interested in how different input images would affect the eye color of people in the images. Without explicitly focusing on race, they started to notice patterns. For instance, Ibrahim noted, for light-skinned people, "eye color often turned green," and Dalia explained that she observed that if the input was a picture of a Black person, it made the eyes brown or black. One of the other groups, Ishmael and Kayden, started by playing with a scuba diving filter, explicitly looking at how it represented race and finding that it "whitewashed" people by giving them "tanned skin and blonde hair." Then Kayden decided to modify the prompt to "basketball player," trying the same prompt on a range of different images of faces. He observed that, regardless of input images, "the skin turned Black and [the filter] gave them a beard". He next tried the prompt "tennis player," realizing that "When I put tennis player, it made her White, but [when I did basketball player], it

made her Black". The images prompting this reflection are reproduced in Figure 3. These examples show how participants identified and investigated potential harmful biases, noticing patterns and drawing upon their prior experiences [13]. After a group discussion, participants agreed to build on Ishmael and Kayden's work to investigate the following hypothesis: **Effect House's image model reinforces gender and race stereotypes about different occupations**.

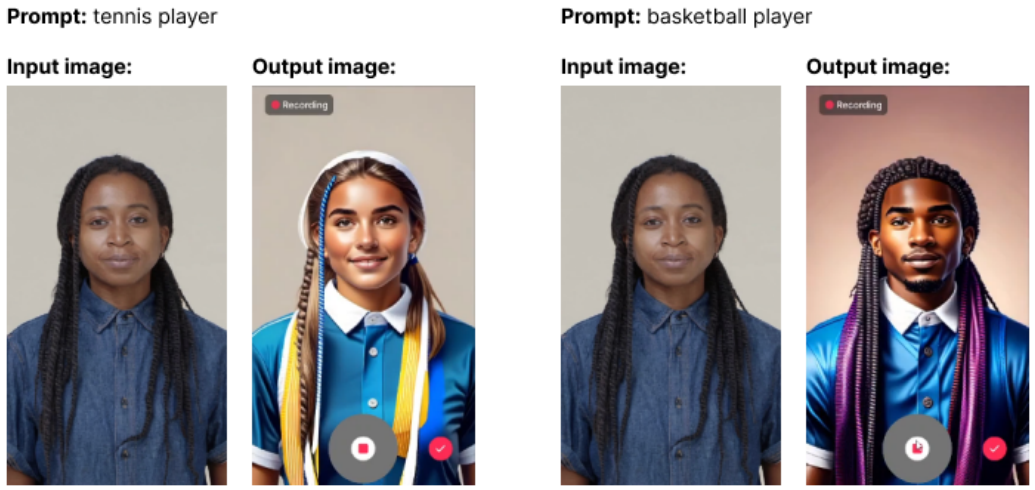

Fig. 3. Inputs and outputs for Kaden's experiments with the prompts "tennis player" and "basketball player." These experiments led the group to investigate if the image model reinforces gender and race stereotypes about different occupations.

4.1.2 Designing Inputs. While deciding on the occupations they could use to test their hypothesis (see Figure 2, participants took different approaches. As we will describe, some did so through discussion, reflecting on their own experiences and perceptions of stereotypes, while others were more hands-on, using Effect House to conduct preliminary explorations and tests of their ideas.

Some groups of participants reflected on common stereotypes related to occupations and how these shaped their expectations of the outputs that Effect House would generate. Kalem suggested "chef" as an occupation because "usually chefs are White and European." Ishmael interjected, saying that thinking of all chefs as White was "just racism." Kalem responded by explaining that he thought they were meant to find stereotypes and test if the systems created images based on those stereotypes. Kalem wrote down "teacher," explaining he would expect outputs to look like White women. Ishmael proposed "rapper" because "it makes me think of Black men, specifically Tupac." Participants also reflected on their own personal experiences with race and occupations. Kalem shared that he thought "nurse" would be a good occupation to investigate because his mother is a nurse and all of her friends are also nurses, and they are all Black women—anticipating a similar trend might be reflected when the filter was used on a range of different faces. Notably, this particular group brainstormed through discussion and did not try to empirically explore any of these occupations on the platform. Like participants in everyday audit studies [13], participants drew on their prior personal experiences and knowledge of societal biases to generate inputs.

At a different table, a group used Effect House to test different occupations as they discussed them. Twyla, tested "basketball player" on an input image of a White woman, observing that the result

Table 2. List of occupations created by participants to test their hypothesis.

Occupations			
1. Tattoo artist	9. New anchor	17. Taxi driver	25. Receptionist
2. President of the USA	10. Scammer	18. 7/11 worker	26. Pizza delivery person*
3. Carpenter	11. Rapper	19. Lawyer	27. Tech support*
4. Construction worker	12. Judge	20. STEM student	28. Oil salesman*
5. Priest	13. Senator	21. Nail technician	29. Corner store worker*
6. Fast food worker	14. Janitor	22. Harvard Student	
7. Basketball player	15. Astronaut	23. Mathematician	
8. Teacher	16. Chef	24. Music artist	

**Occupations 26-29 were not used in testing and analysis*

resembled a Black woman. Horacio explained that the group had to consider traditionally masculine jobs like "soldier" to see the outputs Effect House would generate. Twyla agreed, saying that she always thinks of war as something associated with men. She suggested "construction worker," "chef," and "cook" as occupations to try. She explained that while chefs are usually associated with men, cooks are often with women, "they just do the same job." Similarly, Ibrahim used Effect House to come up with possible occupations. He tested "basketball player," "computer scientist," "gardener," "7/11 worker," "chef," and "teacher" on four different images.

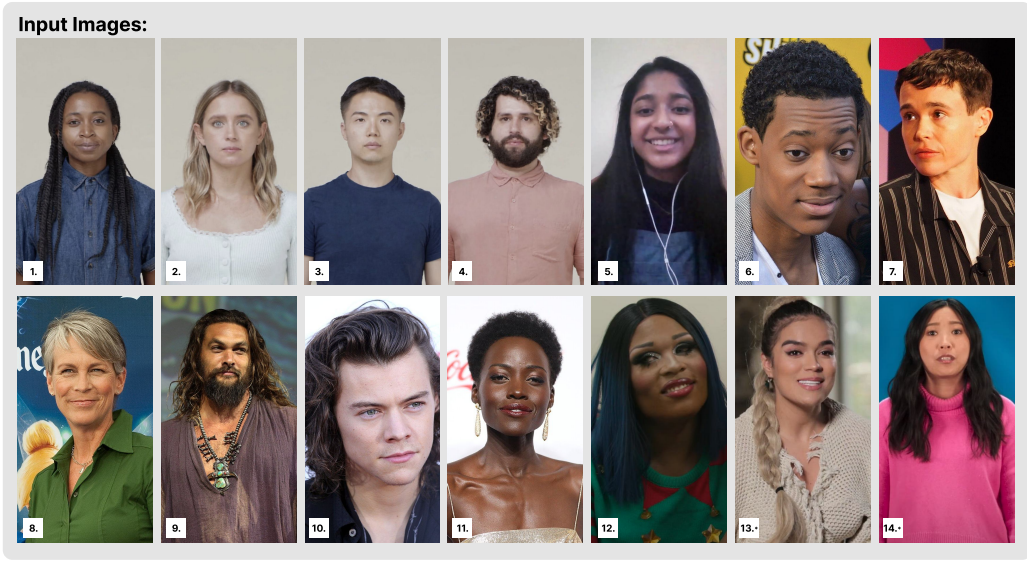

Fig. 4. The input images used by participants in their audit, including 4 default Effect House inputs and 10 images of celebrities selected from Wikimedia Commons.

4.1.3 Running Tests. Running the tests provided participants with opportunities to make inferences about the outputs, notice patterns, and reflect on their own perceptions of occupations. Ziyi, aware of the repetitive nature of running tests on all the images (see Figure 4), explained to a science

Fig. 5. Screenshot of a section of the spreadsheet where participants kept track of their tests.

center instructor that the trick was focusing on "one at a time." She was also continuously reflecting and noticing patterns while running tests. For example, when testing the construction worker prompts with different images, she noted, "they all look the same." Later, while running tests for the nail technician, she explained that she was unsurprised that they all looked very feminine because her mom is a nail technician, and she noticed that it was a very feminine occupation. Similarly,

Dalia reflected after running tests for the occupation "tattoo artist", noting "a lot of people got tattoos... I wonder why [Effect House] gave her, like, a neck tattoo".

Even while running the tests and adding output images to the shared spreadsheet (see Figure 5), participants continued to playfully experiment with Effect House. One participant, Kayden, took the opportunity to further play with the parameters and test images using his own name as a prompt. For instance, after testing a picture of Jason Mamoa with the prompt "a president of the USA at work," he decided to modify the prompt using his own name (e.g., "Kayden Lastname at work") to see what came up. The output was an image of a young Black man. He then tried his own name as a prompt with an image of a Black woman and got an output where the woman had more masculine features, and her race remained unchanged. While these tests were not documented on the shared spreadsheet and did not contribute to the overall audit, screen recordings revealed how the testing activity inspired him to go beyond the task he was supposed to be completing.

4.1.4 Analyzing Data. Participants took different approaches to analyzing the data, with some describing the changes they observed, others annotating perceived gender and race of images, and some centering on concrete observable attributes. Participants also annotated age-related features, even though this attribute was not planned in the original hypothesis (about gender and racial stereotypes). At the same time, while analyzing data, participants continued to further experiment with the system by running more tests.

Some teens, including Twylia, decided to describe the visible differences between each input and output image. For example, for the prompt "Janitor at work" and image 1 (an image of a Black woman with dreads), she noted that the output had "visible makeup, bigger lips, toned eyebrows, and a smoother face." A researcher-facilitator approached Twylia and suggested she count some of the changes she was noticing across images rather than just qualitatively describing them. Based on her interpretation of the suggestion, Twylia continued with descriptive annotations of the output images and also began assigning a masculinity and femininity score, ranging from zero to a hundred, to each output image.

Other participants labeled the data with descriptors based on their own overall perception of the output images. Ibrahim, Kalem, Dalia, and Taylor, for example, annotated whether the people in output images looked more feminine or masculine, older or younger, and darker- or lighter-skinned than the inputs. Similarly, Ishmael and Brooklyn Mae annotated the input and output images using defined binary gender and age categories such as male/female and young/old.

One group of participants decided to focus on observable features rather than their own perception of the outputs. Motivating this choice, Selena explained that "what 'older' means is kind of subjective." Ziyi also emphasized the need to specify what "older" meant because it was hard to know and estimate the age of the AI-generated images. She suggested that they should look at the output images to see if these had wrinkles and gray hair in order to concretely measure how many of the output images looked older. In a similar episode, Ziyi and Kalem discussed how to account for gender changes in the representation of rappers. First, they considered counting how many of the inputs were men and comparing them with the number of output images that (in their own perception) were men. But Ziyi again suggested finding a more concrete way to label the images, focusing on a specific physical attribute like the presence or absence of facial hair.

While analyzing data, one participant, Horacio, decided that he needed to do further testing. He started his analysis by annotating what he noticed in each output image with notes like "gender stays the same but face is different." Unprompted, he decided to rerun some of the tests and change the prompt strength from 0.5 to 1.0. He noted that for some prompts, he did not observe a change in gender between input and output images at a strength of 0.5, such changes were noticeable at 1.0 strength. The process of analyzing led Horacio to reconsider how other variables in the auditing

process could affect the outputs and to loop back to the data collection step to experiment with one variable.

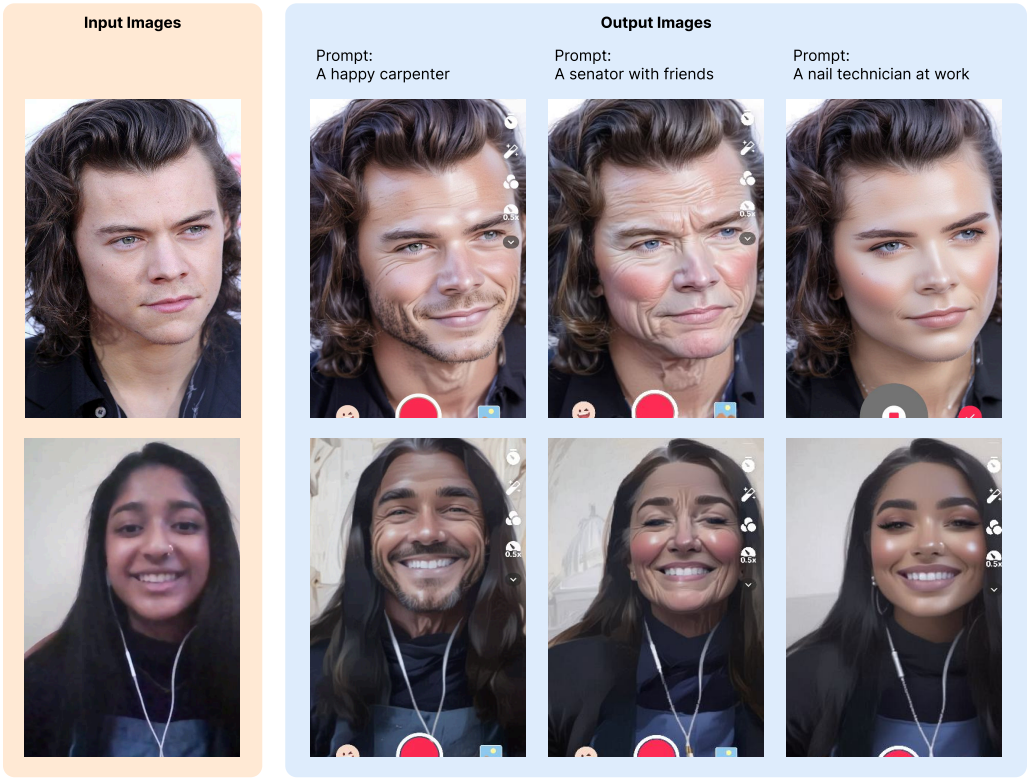

Fig. 6. Examples of input and output images for three prompts.

Overall, participants analyzed 13 out of the 25 occupations that were collected in the previous step (see Table 3 for findings by group). They annotated the image pairs for changes in representation of gender, age, and race caused by the model by comparing the input and output images. Here we provide examples of participants' findings for specific occupations. Four groups analyzed the gender representation in output images for nail technician prompts (for an example, see Figure 6), noting that the outputs were more feminine, with some participants arguing that all images looked more feminine (Kayden and Dalia; Taylor), and one that 87% looked more feminine (Ibrahim). Participants that looked at specific feminine-coded features noted the presence of makeup in 83% of the nail technician outputs (Kalem, Selena and Ziyi). Two groups analyzed the outputs for judges, noticing that judges in general had more feminine characteristics (Ishmael and Brooklyn Mae; Taylor). In terms of age, two groups also looked at judges, noting that 24% of the faces in the output images were young (Ishmael and Brooklyn Mae) and 75% had wrinkles (Kalem, Selena, and Ziyi). They made similar observations for other occupations, noting that fast food workers looked young, while outputs for senators and janitors looked older. Less attention was given to race, with two groups noting race-related attributes: that all rapper outputs had darker skin than the input images (Ibrahim), and that senators had "White aesthetics like blue eyes and blond hair" (Twyla and Horacio).

Table 3. Findings from small group analyses of the data generated in the auditing process.

Group	Audit Findings
Kalem, Selena and Ziyi	Nail techs: 83% have makeup [proxy for feminine features] Fast food worker: 100% of outputs had smooth skin [proxy for younger age] Rapper: 95% of outputs had goatees or facial hair [proxy for masculine features] Judge: 75% had wrinkles [proxy for older age]
Ibrahim	Rapper: 100% of outputs had darkened skin; 96% looked more masculine Nail tech: 87% looked more feminine Taxi driver: 56% looked older
Ishmael and Brooklyn Mae	Priest: 100% outputs were men, even though inputs were 50/50 split Judge: 100% of outputs appeared female/femme Judge: 24% outputs were young
Twyla and Horacio	Senator: 100% were old; most had white aesthetics like blue eyes, blonde hair, and older (with wrinkles added); 95% were men Fast food worker at work: tone eyebrows, smoother face, visible makeup; 80% looked feminine; 20% looked masculine Janitor at work: 85% looking older; 80% looking masculine; 20% looking feminine
Kayden and Dalia	Carpenter: 91.7% of outputs were men President: made everyone older, added wrinkles and gray hair Tattoo artist: added tattoos on bodies like faces Stem scholars: made everyone look younger Nail technician: made everyone look more feminine
Taylor	Nail techs: outputs images were more feminine; sometimes added makeup Receptionists: more feminine characteristics Judges: more feminine characteristics

4.1.5 Creating audit reports. Participants decided on different audiences for their reports and different ways to communicate their findings. They considered distinct relevant parties[1] including the engineers and others involved in the development of Effect House, users that create filters on Effect House, and users of filters on TikTok. Kalem suggested that it was important to share the findings with the developers of Effect House so they might improve the generative model. Ibrahim and Twyla argued that the findings should be shared with filter designers so that they could consider the biases that the system introduces even when the filter designers might not intend them. Ibrahim explained, "Effect House can be biased no matter what". Dalia suggested sharing the findings with other teenagers who use TikTok because "AI is everywhere and young people don't always know how it works." She explained that without knowing how systematic some of these issues are, other teens could find themselves thinking, "Why is it doing this? Why is it doing this to me?"

Some groups decided to make their own TikTok videos to report their audit findings; one group, Twyla and Horacio, decided to make a slide presentation. In a different group, Ziyi and Kalem wrote a script for a TikTok video directed towards the developers of Effect House. Ziyi emphasized that it was important to explain how and why they selected the different occupations they analyzed, noting that they chose to focus on occupations that they perceived as reflecting "the most physical change" between the input and output images. Kalem explained that he hoped the developers of Effect House could take this feedback and make the platform more inclusive so that users could have access to better filters on TikTok.

[1]We use the term "relevant parties" rather than "stakeholders"; for a discussion on the reasoning behind this decision, see [45].

4.2 Triangulating Participants' Findings

To address our second research question, **Did the workshop support participants to reach evidence-based, credible conclusions in their audit?**, we triangulated participants' analyses. We built on our team's prior experience in algorithm auditing by conducting our own analysis of the dataset created by youth. (For details about how we conducted this analysis, see Methods - Section 3.5.2). The goal of this analysis was to answer the hypotheses developed during the workshop to investigate if outputs from Effect House's image model reinforced gender and race stereotypes about different occupations and use these results to triangulate participants' audit results. In addition to gender and race, we also added age representation since several groups of participants also analyzed this kind of bias.

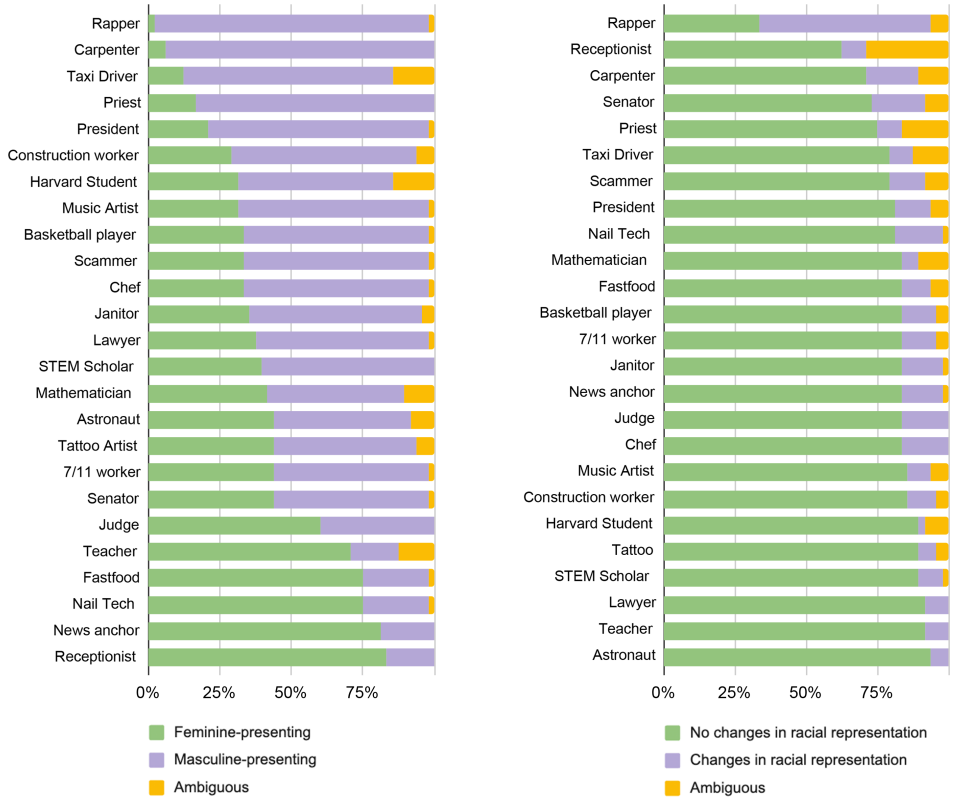

Fig. 7. Bar graphs depicting the gender representations (left graph) and the changes in racial representations (right graph) of the output images for each occupation.

Gender Representation. Our analysis of the dataset, created during the collaborative audit, revealed that Effect House generated a greater number of masculine-presenting outputs for 19 out of the 25 occupations (see Figure 7), depicting an overrepresentation of masculine-presenting figures in these occupations. Here we describe our findings in relation to participants' findings. The input images were selected to be balanced in gender representation (50% F, 50% M), but after our own annotation, the outputs were 41% F, 55% M, 4% ambiguous (A). In certain occupations, the gender imbalance was more pronounced. For instance, 96% of the output images for rappers appeared

masculine-presenting, followed by 94% for carpenters, 83% for priests, 73% for taxi drivers, and 77% for presidents. On the other hand, 83% of receptionists were feminine-presenting, followed by 81% for news anchors, 75% for fast food workers, and 75% for nail technicians. We observed changes in gender representation in at least one test (out of 100 tests) for all input images, with the exception of the image of Jason Mamoa and image 4 (racially ambiguous man with a beard), who always retained a stereotypically masculine presentation in the output image regardless of the input prompt.

Although researchers and participants obtained slightly different quantitative results, the conclusions drawn from both analyses were similar, namely that—in terms of gender representation and biases—Effect House's filter masculinized input images when prompted with occupations like rapper, carpenter, priest, and janitor and feminized outputs for occupations like fast food workers, receptionists, and nail technicians. For example, researchers found that 94% of output images for carpenters were masculine-presenting, while Kayden and Dalia reported that 91.7% of output images were men. In other occupations, such as fast food workers, researchers coded 75% of output images as feminine, while participants argued 80% looked feminine. Participants analyzed the images of rappers in different ways. Kalem, Selena, and Ziyi used facial hair as a proxy for masculine-presenting outputs, finding that 95% of the outputs for rappers had facial hair, while Ibrahim directly labeled the images as masculine or feminine in his own perception, finding that 96% of the outputs were masculine-presenting. Our analysis concorded with Ibrahim that 96% of outputs were masculine-presenting and came close to Kalem's group; we found that 100% of outputs for rappers had facial hair (including stubble).

Racial Representation. Participants' analyses of racial representation were limited, with only two groups addressing racial representation. Ibrahim reported that 100% of outputs for rappers had darker skin than the input images, while Twyla and Horacio noted that most outputs for senators featured "White aesthetics," like blue eyes and blonde hair. In our analysis, we compared input and output images, annotating whether the race of the person in the output image appeared (subjectively) to differ from that of the input. Our analysis revealed that changes in racial representation were not as prominent as those in gender representation. Overall, when comparing input and output images, we found that racial representation changed in 13% of cases, with changes occurring across all occupations. The occupations with the highest frequency of change in racial representation were rappers (60%), carpenters (19%), and senators (19%). Occupations with the lowest frequency of change in racial representation included astronauts (6%), lawyers (8%), and teachers (8%).

Given that these results were less stark than those related to gender, it is possible that the participants did not focus as much on this form of bias because they empirically found the patterns in racial representation less pronounced and interesting in their experiences running the tests.

Age Representation. While age representation was not a part of the initial hypothesis, it emerged as a key theme in participants' audits, with 5 out of 6 groups including observations about changes in age representation. We evaluated age representation using two visual indicators: wrinkles and gray hair. A change in age representation was identified if at least one of these changed between the input and output images.

We found that changes in age representation were more prominent than changes in racial and gender representation, with changes in age representation occurring in 46.75% of cases. Notably, changes in wrinkles were more prominent than changes in gray hair, and the addition of features in the output image was more common than their removal. The presence of wrinkles changed (added or removed) in 44.67% of images, including 40% that involved the addition of wrinkles. Gray hair changed in only 10.33% of images. The occupations that showed the most change in age representation were priest (83.3%), president (83.3%), and senator (75%). Senators had the highest

incidence of wrinkles in output images (100%), followed by news anchors (95.8%), presidents (95.8%), taxi drivers (93.7%), and carpenters (93.7%). Occupations with the highest incidence of wrinkle removal were STEM student (16.67%), nail technician (14.58%), and fast food worker (12.5%).

Participants' analyses were similar to the researchers' findings, including that Effect House tended to age input images when prompted with certain occupations—such as president, senator, and taxi driver—and reduced the perceived age of images when prompted with STEM student. However, participants employed distinct strategies to analyze images, with some (Kalem, Selena, and Ziyi) focusing on specific visual markers, such as the presence or lack of wrinkles, as proxies for age, and others (Kayden and Dalia) taking a comparative approach, describing output images as "older" or "younger" than the inputs. A third approach by Ishmael and Brooklyn Mae relied on personal interpretations of "young" and old", as they described in their audit that 24% of outputs for judges were "young".

5 Discussion

At the highest level, this study contributes evidence that, beyond recognizing isolated instances of representational bias and harm, teenagers can effectively engage in algorithm auditing, collaboratively and *empirically* investigating potentially harmful behaviors in real-world algorithmic systems. Our study confirms evidence from prior research that young people build on their everyday experiences to identify potential biases [36, 54] and extends this by showing that when participating in scaffolded PD activities, teenagers can lead full-fledged audits to empirically research systemic bias and harm. In this section, first we discuss participants' approaches to auditing in relation to auditing literature, then we focus on the implications of our findings for future auditing activities involving youth. Subsequently, we also reflect on some implications for the field of auditing, especially as everyday perspectives are beginning to be included in various ways. Finally, we describe key limitations to be addressed and other future directions in which we hope our team and others will expand this line of research.

5.1 Teens Conducting an Algorithm Audit

While prior research shows that young people can identify harmful biases in researcher-selected scenarios [36, 54]; our case study demonstrates that youth can be protagonists [19] or the main decision-makers in auditing a generative AI system from beginning to end. The PD workshop was structurally open [4, 64], providing teens with autonomy to define what they wanted to investigate. The five steps served as a basic structure that supported participants to be the main agents driving the auditing process [19]. In the following paragraphs, we discuss how teens navigated each of the steps.

Participants developed hypotheses based on experimentation and personal experiences, not only noticing unexpected behaviors but also potentially harmful biases. While some expert audits are conducted as follow-ups to well-documented problematic system behaviors [34] and others are initiated in compliance with local laws [18], expert audits have also been motivated by the researchers' own personal experiences interacting with sociotechnical systems [9, 57]. The open-ended exploration of the tool and its behaviors was instrumental in helping participants develop hypotheses that were relevant to their interests, identities, and experiences.

In most expert audits, auditors create a set of inputs that are systematic, thorough, and thoughtful, and that can be used to rigorously test the hypothesis. It is notable that the domain of common occupations, selected by the participants without explicit direction from researchers, is a popular framing for expert audits. For example, previous studies on gender and race representation of occupations in image searches and the outputs of generative models have used the US Bureau of Labor and Statistics (BLS) categorization of occupations as a starting point [5, 22, 30, 34, 40].

Participants were less exhaustive but very creative when they selected their list of occupations. Like Kalem, when thinking about the gendered and racialized stereotypes around nursing in relation to his mother, participants built on their own personal experiences. They also reflected on societal biases they observed in their everyday lives and suggested occupations such as rapper and nail technician that, to our knowledge, have not been included in previous expert-led studies.

Repeatedly querying an algorithmic system using thoughtfully designed inputs and recording outputs can be tedious and time-consuming. Despite its tedium, we observed that participants were generally engaged in running the tests, often performing some informal analysis of outputs while testing and reflecting on their perceptions of stereotypes. Sometimes participants deviated from the task at hand to pursue their own open-ended investigations. These examples show that auditing with teenagers can be less structured than expert auditing , with learners engaging with different steps of the auditing process at the same time.

Participants adopted various methods to analyze their data, recording observed changes, annotating perceived gender and race of photos, and focusing on observable features. We were able to triangulate their findings to confirm that Effect House generated a greater number of masculine-presenting outputs for 19 out of the 25 occupations, and that age representation reflected societal stereotypes.

Like expert auditors who report their findings with the goal of effecting change [31, 41], participants in our study had a clear understanding that different relevant parties needed to access the findings for different purposes and created distinct messages directed to these parties.

The PD workshop successfully scaffolded participants in conducting an audit. Our study further highlights the promise of expanding the roles young people can play in PD to involve them as auditors of systems that are relevant to their daily lives [19, 36]. For child-computer interaction research, positioning teenagers as auditors of the technologies that are designed and marketed towards them is particularly important, as they may be able to identify issues that designers, adults, or experts cannot find.

5.2 Implications for Future Auditing Learning Activities

Our first research question asked about how participants engaged in this auditing activity. Reflecting on the findings of this first research question, we consider the adaptation and scaffolding of the auditing process promising for future activities with teenagers, and the algorithmic justice topic suitable and resonant with this group of teens. The activity clearly resonated with participants; they connected with their own lived experience throughout the process. For example, Kayden tried running several prompts with his own name. We saw this clear personal engagement even when directly addressing complex topics like social biases, as when Twyla discussed the differences between stereotypes of chefs and cooks. Participants also demonstrated their creativity. Different groups conducted parallel steps of the audit very differently—for example, generating distinct ways to evaluate age representation, such as by subjectively comparing inputs and outputs or by looking for the presence of wrinkles and gray hair.

Our second research question asked whether the workshop supported participants to reach evidence-based, credible conclusions in their audit. This is especially concerning because this activity was meant to allow them to draw conclusions about real systems they use in their daily lives. It could be problematic to leave them with inaccurate conclusions, something we could not control in advance as the process unfolded organically. By triangulating participants' findings, undergirded by prior experience in expert auditing, we observed that their findings were quite close to our own—for example, finding that Effect House appears to produce outputs that are systematically biased according to social stereotypes of different occupations. Although they did not analyze the full dataset, as our team did after the workshop, and although the precise numbers

they calculated varied from ours in some cases (e.g., while we found that 75% of output images for fast food workers were feminine-presenting, Twyla and Horacio argued 80% looked feminine), the overall direction of the trends they reported matched ours. This is evidence that the PD workshop supported participants in reaching evidence-based, credible conclusions.

Moreover, several of the participants' divergences from formal auditing processes proved insightful and valuable. When analyzing the data, for example, several groups began to notice trends related to age, diverging from the hypothesis. Age biases are not a common topic for expert audits; as reflected in the related work we described in Section 2, we do not know of any other image bias audits that study age. But clearly this dimension was salient to participants, who uncovered notable trends that complemented the findings on race and gender (e.g., that filter outputs for presidents, in addition to looking more masculine-presenting, also looked older). Similarly, when selecting the occupations to examine, they chose several uncommon ones not previously explored in the expert literature, like nail technician and rapper, that nevertheless produced interesting findings.

Looking ahead to future deployments of auditing-based learning activities, we encourage educators to integrate opportunities for experimentation throughout the process and not restrict learners very strictly to the task at hand. We observed that the participants' approaches were playful, as they devised alternative hypotheses and conducted impromptu open-ended explorations throughout the process. We saw them gaining inspiration in later activities that prompted them to loop back to earlier stages of the audit and try new directions. And as we just described, although efforts should be made to ensure that learning activities will leave teenagers with valid and well-informed conclusions, there are many other dimensions in which learners follow their creative impulses and make different decisions than what expert auditors would normally do.

5.3 Implications for the Field of Auditing

In addition to contributing to the literature on child-computer interaction, we see a couple of implications of this work for auditing research. First, it confirms the value of participatory approaches to auditing that involve non-expert real system users in shaping the questions, methods, and interpretations of an audit [24]. Notably, most prior work in this space has focused on the potential for users to pose questions and conduct explorations. Beyond these steps, we also saw participants consider how their findings should be communicated to different relevant parties in distinct ways (e.g., making TikTok videos to communicate findings to TikTok users). This suggests that user populations may be especially well-positioned to effectively communicate their findings to peer groups after conducting an audit (the fifth step in our set of activities). We hope subsequent work in end-user auditing will consider ways to leverage end users' expertise and trust among their peers to not only conduct audits but also communicate audit results.

And of course, this work expands the idea of end-user auditing to include teenagers. Young people, as avid users of many AI systems, have important expertise that we saw arise in this study (e.g., the consideration of age biases, the choice of occupations that resonated with them personally). Like other audits involving everyday users have shown [51], our 14-15-year-old participants contributed valuable auditing insights based on their lived experiences.

5.4 Limitations and Future Work

Perhaps the largest limitation of this research is that we only present an analysis of one case study of a single algorithmic system with a specific group of teenagers. Given the value we saw for teenagers, we believe it is worth expanding on this work to develop materials for conducting workshops focusing on other platforms, other age groups, and additional groups of learners. We encourage other researchers to do so and also plan to do this ourselves, and, as we referenced in our

Positionality Statement (Section 3.4) we emphasize the importance of partnering with teenagers and collaboratively creating appropriate learning materials with them.

We conducted these activities in the context of an extended summer program, but the concept of engaging teenagers in auditing could be more impactful if designed for classroom settings with a wider range of learners—not just those with interest and access to an extracurricular program, like the participants in our study. To this end, our next steps involve exploring how the five auditing steps that framed our activities could be used to integrate algorithm auditing activities in classroom settings. This will raise some interesting challenges, for instance, pertaining to many schools' blocking of platforms like TikTok on school networks and use of mobile devices in classrooms. To address these issues, we must take a participatory approach and partner with educators and teenagers to explore other possibilities.

Focusing specifically on the design of the PD workshop, we want to flag a couple of challenges. Creating a large data set to conduct a full-fledged audit was time-consuming and tedious. In adapting these kinds of activities to classroom settings, we could consider focusing on analysis and providing teenagers with existing curated datasets. Another challenge was the impossibility of the participants to annotate all 1200 tests, an intractable number for any learning activity of reasonable length. To address this, it is important to design tools that can streamline and support annotation and to consider other pedagogical designs, such as having participants annotate some of the data while providing them with already annotated data. Finally, after reflecting on the workshop ourselves, we also wondered whether adding a sixth step, Reflection, in which participants could reflect on their experiences and challenges, might support learners in thinking about how auditing practices can be applicable in their everyday lives.

6 Conclusion

Motivated by the importance of fostering artificial intelligence and machine learning (AI/ML) literacies among youth, we drew on algorithm auditing research to design and deploy an auditing-based participatory design workshop with a group of 14 teenagers. Auditing is a method used by experts to interrogate algorithmic systems and their social impacts that has more recently seen extensions through which everyday users (generally adults) can participate in similar processes in the context of their daily interactions with these technologies. The teenagers in our study (ages 14–15) engaged in a PD workshop auditing the generative AI model that powers TikTok filters using Effect House, TikTok's filter development environment. Our case study shows that teenagers are able to conduct a complete and collaborative audit of a real-world algorithmic system when scaffolded appropriately. Furthermore, conducting our own analysis, we were able to triangulate their findings, confirming the potential of the workshop in scaffolding youth to reach evidence-based, credible conclusions. This work demonstrates the potential for researchers to partner with teenagers in future audits of algorithmic systems that young people use in their daily lives and also to design AI literacies learning activities that support young people in conducting algorithm audits of the systems that they interact with every day.

Acknowledgments

With regards to Lucia Kulzer, Alexis Cabrera-Sutch, Carly Netting, and Danielle Marino for their support during the participatory design workshop. This study was supported by National Science Foundation (NSF) grants #2333469 and #2342438. 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, the University of Pennsylvania, or Columbia University.

References

[1] Bobby Allyn, Sylvia Goodman, and Dara Kerr. 2024. TikTok executives know about app's effect on teens, lawsuit documents allege. https://www.npr.org/2024/10/11/g-s1-27676/tiktok-redacted-documents-in-teen-safety-lawsuit-revealed.

[2] Monica Anderson, Michelle Faverio, and Jeffrey Gottfried. 2023. Teens, Social Media and Technology 2023. Pew Research Center. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

[3] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (April 2021), 34 pages. doi:10.1145/3449148

[4] Liam J Bannon and Pelle Ehn. 2012. Design: design matters in Participatory Design. In *Routledge international handbook of participatory design*. Routledge, New York, 37–63.

[5] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1493–1504. doi:10.1145/3593013.3594095

[6] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 1–9.

[7] Jolie Bonner, Florian Mathis, Joseph O'Hagan, and Mark Mcgill. 2023. When Filters Escape the Smartphone: Exploring Acceptance and Concerns Regarding Augmented Expression of Social Identity for Everyday AR. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology* (Christchurch, New Zealand) *(VRST '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. doi:10.1145/3611659.3615707

[8] Claus Bossen, Christian Dindler, and Ole Sejer Iversen. 2016. Evaluation in participatory design: a literature survey. In *Proceedings of the 14th Participatory Design Conference: Full Papers - Volume 1* (Aarhus, Denmark) *(PDC '16)*. Association for Computing Machinery, New York, NY, USA, 151–160. doi:10.1145/2940299.2940303

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, , 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[10] Kaitlyn Burnell, Allycen R Kurup, and Marion K Underwood. 2022. Snapchat Lenses and Body Image Concerns. *New Media & Society* 24, 9 (2022), 2088–2106. doi:10.1177/1461444821993038

[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] Norman K Denzin. 2017. *The research act: A theoretical introduction to sociological methods*. Routledge, New York.

[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* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 626, 19 pages. doi:10.1145/3491102.3517441

[14] Christian Dindler, Ole Sejer Iversen, Mikkel Hjorth, Rachel Charlotte Smith, and Hannah Djurssø Nielsen. 2023. DORIT: An analytical model for computational empowerment in K-9 education. *International Journal of Child-Computer Interaction* 37 (2023), 100599.

[15] Miriam Doh, Corinna Canali, and Anastasia Karagianni. 2024. Pixels of Perfection and Self-Perception: Deconstructing AR Beauty Filters and Their Challenge to Unbiased Body Image. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences* (Stockholm, Sweden) *(IMX '24)*. Association for Computing Machinery, New York, NY, USA, 349–353. doi:10.1145/3639701.3663636

[16] Ana Maria Bustamante Duarte, Nina Brendel, Auriol Degbelo, and Christian Kray. 2018. Participatory Design and Participatory Research: An HCI Case Study with Young Forced Migrants. *ACM Trans. Comput.-Hum. Interact.* 25, 1, Article 3 (Feb. 2018), 39 pages. doi:10.1145/3145472

[17] Eureka Foong, Darren Gergle, and Elizabeth M. Gerber. 2017. Novice and Expert Sensemaking of Crowdsourced Design Feedback. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 45 (Dec. 2017), 18 pages. doi:10.1145/3134680

[18] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1107–1120. doi:10.1145/3630106.3658959

[19] Ole Sejer Iversen, Rachel Charlotte Smith, and Christian Dindler. 2017. Child as Protagonist: Expanding the Role of Children in Participatory Design. In *Proceedings of the 2017 Conference on Interaction Design and Children* (Stanford, California, USA) *(IDC '17)*. Association for Computing Machinery, New York, NY, USA, 27–37. doi:10.1145/3078072.3079725

[20] Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda, and Subbarao Kambhampati. 2022. Imperfect ImaGANation: Implications of GANs exacerbating biases on facial data augmentation and snapchat face lenses. *Artificial Intelligence* 304 (2022), 103652.

[21] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic Folk Theories and Identity: How TikTok Users Co-Produce Knowledge of Identity and Engage in Algorithmic Resistance. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 305:1–305:44. Issue CSCW2. doi:10.1145/3476046

[22] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. doi:10.1145/2702123.2702520

[23] Amy J. Ko, Alannah Oleson, Neil Ryan, Yim Register, Benjamin Xie, Mina Tari, Matthew Davidson, Stefania Druga, and Dastyni Loksa. 2020. It is time for more critical CS education. *Commun. ACM* 63, 11 (Oct. 2020), 31–33. doi:10.1145/3424000

[24] Michelle S. Lam, Mitchell L. Gordon, Danaé 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. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (Nov. 2022), 34 pages. doi:10.1145/3555625

[25] Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, and Danaé Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 360 (Oct. 2023), 37 pages. doi:10.1145/3610209

[26] Melissa R Laughter, Jaclyn B Anderson, Mayra BC Maymone, and George Kroumpouzos. 2023. Psychology of aesthetics: Beauty, social media, and body dysmorphic disorder. *Clinics in dermatology* 41, 1 (2023), 28–32.

[27] Victor R. Lee, Victoria Delaney, and Parth Sarin. 2022. Eliciting High School Students' Conceptions and Intuitions about Algorithmic Bias. In *Proceedings of the 2022 ACM Conference on International Computing Education Research - Volume 2* (Lugano and Virtual Event, Switzerland) *(ICER '22)*. Association for Computing Machinery, New York, NY, USA, 35–36. doi:10.1145/3501709.3544279

[28] Ang Li, Zheng Yao, Diyi Yang, Chinmay Kulkarni, Rosta Farzan, and Robert E. Kraut. 2020. Successful Online Socialization: Lessons from the Wikipedia Education Program. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 50 (May 2020), 24 pages. doi:10.1145/3392857

[29] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[30] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Evaluating Societal Representations in Diffusion Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., New Orleans, 56338–56351. https://proceedings.neurips.cc/paper_files/paper/2023/file/b01153e7112b347d8ed54f317840d8af-Paper-Datasets_and_Benchmarks.pdf

[31] Yaaseen Mahomed, Charlie M. Crawford, Sanjana Gautam, Sorelle A. Friedler, and Danaé Metaxa. 2024. Auditing GPT's Content Moderation Guardrails: Can ChatGPT Write Your Favorite TV Show?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 660–686. doi:10.1145/3630106.3658932

[32] Harvey Mannering. 2023. Analysing Gender Bias in Text-to-Image Models Using Object Detection. arXiv:2307.08025 [cs] doi:10.48550/arXiv.2307.08025

[33] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (Nov. 2019), 23 pages. doi:10.1145/3359174

[34] Danaé 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 26 (April 2021), 23 pages. doi:10.1145/3449100

[35] Danaé 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.

[36] Luis Morales-Navarro, Yasmin Kafai, Vedya Konda, and Danaé Metaxa. 2024. Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. Association for Computing Machinery, New York, NY, USA, 560–573. doi:10.1145/3628516.3655752

[37] 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,

Cham, Switzerland, 521–538.

[38] Luis Morales-Navarro and Yasmin B Kafai. 2024. Unpacking Approaches to Learning and Teaching Machine Learning in K-12 Education: Transparency, Ethics, and Design Activities. In *Proceedings of the 19th WiPSCE Conference on Primary and Secondary Computing Education Research*. Association for Computing Machinery, New York, 1–10.

[39] Luis Morales-Navarro, Yasmin B Kafai, Lauren Vogelstein, Evelyn Yu, and Danaé Metaxa. 2025. Learning About Algorithm Auditing in Five Steps: Scaffolding How High School Youth Can Systematically and Critically Evaluate Machine Learning Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, 28. Association for the Advancement of Artificial Intelligence, Philadelphia, 29186–29194.

[40] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 786–808. doi:10.1145/3600211.3604711

[41] Leonardo Nicoletti and Diana Bass. 2023. Humans Are Biased. Generative AI Is Even Worse: Stable Diffusion's text-to-image model amplifies stereotypes about race and gender — here's why that matters. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

[42] Rizu Paudel and Mahdi Nasrullah Al-Ameen. 2024. Leveraging the Power of Storytelling to Encourage and Empower Children towards Strong Passwords. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 504 (Nov. 2024), 27 pages. doi:10.1145/3687043

[43] Claire Kathryn Pescott. 2020. "I Wish I Was Wearing a Filter Right Now": An Exploration of Identity Formation and Subjectivity of 10- and 11-Year Olds' Social Media Use. *Social Media + Society* 6, 4 (2020), 2056305120965155. doi:10.1177/2056305120965155

[44] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1882–1895. doi:10.1145/3593013.3594134

[45] Mark S Reed, Bethann Garramon Merkle, Elizabeth J Cook, Caitlin Hafferty, Adam P Hejnowicz, Richard Holliman, Ian D Marder, Ursula Pool, Christopher M Raymond, Kenneth E Wallen, et al. 2024. Reimagining the language of engagement in a post-stakeholder world. *Sustainability Science* 19 (2024), 1–10.

[46] 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. doi:10.1145/3568813.3600110

[47] Antti Salovaara and Leevi Vahvelainen. 2025. Triangulating on Possible Futures: Conducting User Studies on Several Futures Instead of Only One. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 478, 16 pages. doi:10.1145/3706598.3713565

[48] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.

[49] Joseph S. Schafer, Kate Starbird, and Daniela K. Rosner. 2023. Participatory Design and Power in Misinformation, Disinformation, and Online Hate Research. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 1724–1739. doi:10.1145/3563657.3596119

[50] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge, New York.

[51] Hong Shen, Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 433 (Oct. 2021), 29 pages. doi:10.1145/3479577

[52] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 153 (Oct. 2020), 22 pages. doi:10.1145/3415224

[53] Jaemarie Solyst, Ellia Yang, Shixian Xie, Jessica Hammer, Amy Ogan, and Motahhare Eslami. 2024. Children's Overtrust and Shifting Perspectives of Generative AI. In *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024*, R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, and E. Suárez (Eds.). International Society of the

Learning Sciences, Buffalo, 905–912.
[54] 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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 364 (Oct. 2023), 27 pages. doi:10.1145/3610213
[55] Luhang Sun, Mian Wei, Yibing Sun, Yoo Ji Suh, Liwei Shen, and Sijia Yang. 2023. Smiling Women Pitching Down: Auditing Representational and Presentational Gender Biases in Image Generative AI. arXiv:2305.10566 [cs] doi:10.48550/arXiv.2305.10566
[56] Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 667–678. doi:10.1145/3531146.3533132
[57] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (May 2013), 44–54. doi:10.1145/2447976.2447990
[58] Ying Tang, Hadar Ziv, and Sameer Patil. 2025. Learning to Work From Home: How Novice and Experienced Software Professionals Compare Online and In-person Collaboration. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP20 (Jan. 2025), 38 pages. doi:10.1145/3701199
[59] TikTok. 2024. Effect Guidelines. https://effecthouse.tiktok.com/learn/guides/general/guidelines/effect-guidelines.
[60] Lauren Vogelstein, Vedya Konda, Deborah Fields, Yasmin Kafai, Luis Morales-Navarro, and Danaé Metaxa. 2025. Rapid Testing, Duck Lips, and Tilted Cameras: Youth Everyday Algorithm Auditing Practices with Generative AI Filters. In *Proceedings of the 19th International Conference of the Learning Sciences (ICLS 2025)*. International Society of the Learning Sciences, Helsinki, 1844–1848.
[61] Kexin Bella Yang, Tomohiro Nagashima, Junhui Yao, Joseph Jay Williams, Kenneth Holstein, and Vincent Aleven. 2021. Can Crowds Customize Instructional Materials with Minimal Expert Guidance? Exploring Teacher-guided Crowdsourcing for Improving Hints in an AI-based Tutor. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 119 (April 2021), 24 pages. doi:10.1145/3449193
[62] Robert K. Yin. 2012. Case study methods. In *APA Handbook of Research Methods in Psychology, Vol. 2. Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (Eds.). American Psychological Association, Washington, DC, 141–155.
[63] Robert K Yin. 2018. Case study research and applications.
[64] Manuel Zacklad. 2003. Communities of action: a cognitive and social approach to the design of CSCW systems. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '03)*. Association for Computing Machinery, New York, NY, USA, 190–197. doi:10.1145/958160.958190
[65] Yanzhe Zhang, Lu Jiang, Greg Turk, and Diyi Yang. 2023. Auditing Gender Presentation Differences in Text-to-Image Models. arXiv:2302.03675 [cs] doi:10.48550/arXiv.2302.03675
[66] Yuhang Zhao. 2020. Analysis of TikTok's Success Based on Its Algorithm Mechanism. In *2020 International Conference on Big Data and Social Sciences (ICBDSS)* (2020-08). IEEE, Xi'an, China, 19–23. doi:10.1109/ICBDSS51270.2020.00012

A Coding Scheme

Table 4. Coding scheme used in our analysis of the dataset generated by youth.

Axis	Category	Codes
Gender	Change in gender representation	Yes: gender represented in the input image is different than gender represented in the output image (e.g., input has a masculine-presenting figure and output image has a feminine-presenting figure)
		No: gender represented in the output image is the same as in the input image
		Ambiguous: it is not clear if there was a change in gender representation; it is not clear if output image is masculine-presenting or feminine-presenting
	Facial Hair*	Presence: visible facial hair, including stubble
		Absence: no visible facial hair
	Blush or Mascara*	Presence: visible blush or mascara
		Absence: no visible blush or mascara
Age	Wrinkles	Presence: visible wrinkles, including smile lines
		Absence: no visible wrinkles
	Gray Hair	Presence: visible gray hair
		Absence: no visible gray hair
Race	Change in racial representation	Yes: race represented in the input image is different than race represented in the output image (e.g., input has an Asian man and output as a White man)
		No: race represented in the output image is the same race as in the input image
		Ambiguous: it is not clear if there was a change in racial representation
	Fade**	Presence: visible fade hairstyle hair on the sides of head is shorter than the hair on the top of the head
		Absence: no visible fade hairstyle
	Hairstyle changes**	Curlier: output image has curlier hair than input image
		Straighter: output image has straighter hair than input image
		Same: there is no difference in hairstyles between input and output images.

* included in our analysis to examine youth's use of these features as proxies for gender
** included in our analysis because some youths noted that these could be racially coded features

Received October 2024; revised April 2025; accepted August 2025