The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI

JAEMARIE SOLYST, Carnegie Mellon University, USA
ELLIA YANG, Carnegie Mellon University, USA
SHIXIAN XIE, Carnegie Mellon University, USA
AMY OGAN, Carnegie Mellon University, USA
JESSICA HAMMER, Carnegie Mellon University, USA
MOTAHHARE ESLAMI, Carnegie Mellon University, USA

Youth regularly use technology driven by artificial intelligence (AI). However, it is increasingly well-known that AI can cause harm on small and large scales, especially for those underrepresented in tech fields. Recently, users have played active roles in surfacing and mitigating harm from algorithmic bias. Despite being frequent users of AI, youth have been under-explored as potential contributors and stakeholders to the future of AI. We consider three notions that may be at the root of youth facing barriers to playing an active role in responsible AI, which are youth (1) cannot understand the technical aspects of AI, (2) cannot understand the ethical issues around AI, and (3) need protection from serious topics related to bias and injustice. In this study, we worked with youth (N = 30) in first through twelfth grade and parents (N = 6) to explore how youth can be part of identifying algorithmic bias and designing future systems to address problematic technology behavior. We found that youth are capable of identifying and articulating algorithmic bias, often in great detail. Participants suggested different ways users could give feedback for AI that reflects their values of diversity and inclusion. Youth who may have less experience with computing or exposure to societal structures can be supported by peers or adults with more of this knowledge, leading to critical conversations about fairer AI. This work illustrates youths' insights, suggesting that they should be integrated in building a future of responsible AI.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; • **Social and professional topics → Computing / technology policy**; • **Applied computing → Education**.

Additional Key Words and Phrases: youth, K-12, adolescents, FATE, fair AI, responsible AI, algorithm auditing, workshop, computing education

ACM Reference Format:
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 (October 2023), 27 pages. https://doi.org/10.1145/3610213

Authors' addresses: Jaemarie Solyst, jsolyst@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Ellia Yang, elliay@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Shixian Xie, shixianx@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Amy Ogan, aeo@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA, 15213; Jessica Hammer, hammerj@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213; Motahhare Eslami, meslami@andrew.cmu.edu, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213.

2573-0142/2023/10-ART364
https://doi.org/10.1145/3610213

1 INTRODUCTION

Recently, artificial intelligence (AI) has become ubiquitous in a range of popular technologies, from social media and entertainment to high-stakes decision-making tools. Youth and families regularly interface with AI in these contexts. However, AI and data-driven technology can reflect societal biases, which cause real-world harm [50]. Youth are among those harmed by AI when, for example, it encodes societal racism [26]. Since almost all teens report that they have access to a smart phone [2], they are regularly exposed to biased algorithmic outcomes and design decisions. This is especially the case for those from underrepresented backgrounds in tech, such as Black communities [49], women [76], and non-binary users [60]. With this in mind, we focus on diverse groups who are underrepresented in tech.

Within the CSCW community and broader HCI audience, there have been conversations around fairness and harms in socio-technical systems and the importance of direct stakeholder engagement in supporting AI practitioners to identify and mitigate harms in AI systems [45], as well as recent efforts to define what 'age-appropriate AI' looks like for youth [77]. However, few studies have included youth as impactful stakeholders or explored possible routes to include minors' feedback in responsible AI.

While minors are often overlooked as stakeholders in responsible AI efforts, they have demonstrated significant potential to confront AI harm. For example, recently, youth in the UK protested against AI that automatically graded them in a way that disadvantaged the working-class students. Parents supported their children by further pressuring the government. As a result of these protests, the UK government terminated the use of the grading algorithm due to inequities it caused [40]. Youth have also developed strategies for working around algorithmic systems to achieve their goals. Some teenagers, for example, believed that the Facebook News Feed curation algorithm promoted posts with commercial keywords; therefore, they added random product names to their posts as a means of influencing the algorithm and gaining more visibility in their friends' feeds [13, 27]. Examples like these show the potential for youth to understand and manipulate the workings of algorithmic systems and to have agency in defining what fair and effective AI might look like.

Despite this potential, there is a dearth of research on how to foster the critical insight youth have to offer in surfacing and mitigating algorithmic harm. One potential barrier to this may be rooted in the perception that regular users of algorithmic systems do not have enough knowledge to understand the technical and ethical complexities of these systems. Recent years have seen the emergence of a new phenomenon, "everyday algorithm auditing," in which regular users can be a part of the process of surfacing algorithmic bias, providing insight about both problematic and ethical AI behavior [21, 58]. To date, however, this approach has only been studied with adults.

Youth have faced barriers to their involvement in responsible AI and have been under-explored as a group that can contribute to surfacing and mitigating algorithmic biases. We believe these barriers are rooted in three key notions about youth: *(1) youth cannot understand the technical aspects of AI, (2) youth cannot understand the ethics issues around AI, (3) youth need protection from serious topics related to harm and injustice.* In the first case, we agree that many youth come from non-technical backgrounds and may have less lived experience with technology than adults. We therefore seek to understand to what extent this is actually a barrier to their agency in everyday algorithm auditing. In the second case, we recognize that many processes around ethical and policy-related topics are not inclusive of children (e.g., in the case of voting in the United States [55]). However, youth are capable of complex moral reasoning [17, 69], and therefore may be able to participate in conversations about ethics and technology. Finally, we note that many attempts to protect children from serious topics result in censorship of important ideas [6] and may lead to continuation of injustice [9]. Rather than avoid these issues, we consider that there are methods

to make these topics youth-inclusive, which supports engagement rather than removing youth agency. In particular, we include families in this work to understand how children may additionally be supported by their parents in engaging in these topics.

This work takes the stance that since youth are regular users of AI and face its impacts, they are stakeholders in the future of AI. We therefore interrogate the above notions and aim to empirically explore them by asking the following questions:

- **RQ1**: How might youth perceive and identify algorithmic bias and its impacts, and to what extent are their perceptions accurate?
- **RQ2**: How do youth ideate possible designs to surface and mitigate harms in AI?
- **RQ3**: How can youth be supported in understanding and addressing algorithmic bias?

We conducted four workshops with thirty youth participants in first grade through 12th grade, as well as with six parents, to understand how youth may be involved in surfacing and mitigating AI bias and harm. We recruited a diverse group of youth from marginalized populations including girls and Black families. Our focus on participants from underserved backgrounds relates to the increased harm they face from bias in AI, such as in the case of Black families facing unfair screening results for child maltreatment predictions [14] and gender bias that is continually documented in AI-driven technology (e.g., [42]). We engaged participants in a series of activities to investigate their perception of fairness in AI, ability in identifying AI biases and the nuances of harm severity, and their design ideas for a future of more fair AI.

We find that youth are capable of identifying, contemplating, and articulating complex notions of fairness in AI, even around serious issues of bias and harm. When participants were presented with various cases of AI bias which differed in type and the level of the harm they might inflict, participants, even those without strong technical backgrounds, were able to identify these biases, despite some examples being more nuanced and hidden in social and cultural norms. For example, when seeing a screenshot of Google Image search results for 'wedding', youth brought up several types of biases including the lack of LGBTQ+ couples and interracial marriages, along with including mainly fancy and culturally Western weddings. Countering the idea that children may not be able to engage in the conversation around problematic AI, we saw many participants were passionate about voicing their opinions about bias in AI. This showed the promising ability of youth in navigating a world where (flawed) algorithms are significantly influencing digital and social structures.

Further, we observed that youth went beyond identifying AI biases by discussing the potential ways to mitigate such harm. This includes an organic discussion coming from participants about what fairer AI should look like, and whether it should represent the current state of society (which could reflect harmful but existing bias) or an ideal future state. Youth also came up with many ways to design systems for surfacing and mitigating harm, including report and feedback mechanisms, users' agency in adjusting potentially harmful algorithmic outputs, and AI technology being transparent about its shortcomings in its design.

We identified places where youth could be supported by their parents or peers who may have higher prior knowledge, relevant lived experiences, or knowledge about societal systems, and therefore see opportunities for designing situated systems where youth are not operating alone with the algorithm but rather leveraging parental and peer support to engage. Our findings contribute to a growing body of knowledge investigating how youth and families may be involved in creating ethical technology, with a focus on algorithm auditing and emphasis on those with underrepresented identities in tech.

2 RELATED WORK AND BACKGROUND

2.1 Harm of Biased AI

Recent harms of AI include bias in a number of popular media platforms and technology. For example, Twitter used AI-driven computer vision, which cropped out people with darker skin from images [30]; Google Search results riddled with stereotypes [34]; and voice recognition algorithms recognizing some groups poorly, including people with certain accents and more feminine voices [5], as well as children [59]. AI has also played a detrimental role in unfair decisions regarding child welfare predictions by falsely predicting cases of child maltreatment in Black families [14]. Amazon additionally developed a gender-biased resume screening tool, which gave women lower applicant scores and less opportunity to be hired [18].

Ninety-five percent of teens reported having access to a smartphone [2], meaning that many are likely to experience the impacts of algorithmic bias. For instance, youth may be exposed to racist content on AI-driven technology [10], and personalized recommendations may lead to segregation in the content Black and White teens are exposed to [71]. Minors are at a particularly vulnerable age and susceptible to bias, as they are developing their identities as young people [26]. This work aims to more deeply understand how children, potentially supported by their parents or peers, might perceive bias in AI.

2.2 Youth and Fairness in AI

With AI being relatively new in computing education efforts [74], many youth do not have access to AI literacy opportunities. Past research has explored how youth from underrepresented backgrounds perceive fairness in AI. While children may not know exactly how AI works, they are capable of ideating futures where AI is used to solve societal problems they care about, and they define fairness in AI as encapsulating equality, equity, kindness, and without bugs or technical malfunctions [64, 66, 67]. However, this raises the question: *How might youth understand **un**fair and biased AI?* Past studies have investigated how youth identify fairness in technology. Prior work has seen that children aged 9-12 were able to understand unfairness in AI examples that they have experienced directly (e.g., if their culturally Black American name was often autocorrected by word editors) [67], but they sometimes struggled to understand algorithmic bias that was both unintentional and scalable (e.g., bias that showed up in search algorithms and job ads), being more able to grapple with harm that was intentional and embodied (e.g., an AI-powered robot programmed to carry out evil deeds [66]). Recent work also explored threats to 'techquity' (technology + equity) with Black youth (8-12 years old) and found that learners need to be supported in considering both the visible (e.g., scams, negative impacts on users related to mental health) and invisible (e.g., privacy and tracking of data) harms of AI technology [15]. While newer and developing, there have been recent efforts to create formal (in-school) and informal (out-of-school) learning opportunities for youth comprising both technical and ethical components of AI (e.g., [22, 43, 80]), with some research focusing on centering learners from marginalized backgrounds (e.g., [66]). Other research has resulted in the development of tools to help with scaffolding children in thinking about fairness in AI, e.g., that use 'explainability' to prompt critical thinking [46]. Finally, Druga et al. [23] describes family as a space for youth and family members to learn about and reflect on AI together. With a focus on surfacing and mitigating algorithmic bias, we aim to build on this work and understand if and how parents might support their children in understanding bias in AI.

2.3 Algorithm Auditing and Everyday Users

Algorithm Auditing can be defined as surfacing and reporting issues or problematic behavior of algorithms [58]. In the context of this paper, it may be thought of as surfacing biases and potential

harm in AI. The original concept of algorithm auditing was created by developers and researchers, and required substantial technical knowledge. For example, in code audits, a developer uses tests to explore what vulnerabilities exist in the technology, while a scraping audit involves issuing inquiries to observe system behavior [58].

In practice, users are often the first to notice and sometimes surface harmful algorithmic behavior. They could also have critical insights that developers may lack when considering bias in AI [21]. One study by Shen et al. [63] suggests that adult users can effectively detect subpar behavior and algorithmic bias in their frequent interactions with AI. Crowd-sourced and collaborative sensemaking may be particularly fitting for everyday users to contribute to algorithm auditing [21, 58]. DeVos et al. [21] suggests that a four-step process can take place in everyday user-driven auditing, where users can (1) initiate auditing by recognizing the potential harm or faulty behavior of an algorithm, then (2) raise awareness (e.g., by reporting or posting on social media about it), (3) hypothesizing and testing by trying different inputs with an algorithm, all in order for (4) remediation to take place.

We emphasize that this process requires the ability to give reactions and feedback but does not require a strong technical understanding of AI. Prior literature on user auditing would suggest this is the case with everyday users [21, 63]. Despite this work with users, including those without technical backgrounds, children remain an under-explored group. Further, it is unclear what design or form youth-inclusive systems should take. This work looks to explore how youth-facing systems may be designed.

2.4 Potential Barriers to Youth Participation

We observe that youth have already led public protests against biased AI systems that affect them, such as the UK-based 'F the Algorithm' movement [40]. This example showcases the potential for youth to have agency and play a critical role in defining fairer futures with AI. At the same time, we observe that youth are often excluded from full participation in both social and socio-technical systems, as we illustrate with examples from other domains.

Given that the field of algorithm auditing is still developing, we seek to intervene early on behalf of youth. These societal defaults would be easy to replicate, particularly in the absence of empirical data on youth capabilities. For example, in studies understanding fairness notions in a screening tool for children and families, youth are not a part of the conversation, even though they are at the center of the AI-driven decisions [12]. In the spirit of algorithm audits, we seek to expose these defaults and provide the necessary empirical data, so that we can more effectively support youth agency in this domain.

2.4.1 Technical knowledge. Do children lack technical knowledge or educational opportunities to participate in discussions around fair AI? Understanding AI from a technical aspect can be difficult, even for adult engineers [51]. It requires a certain level of fluency with data science and data-savvy knowledge, mathematics, and often programming for implementation. There are documented barriers for youth understanding AI and computing topics, such as engaging in and having misconceptions about systems thinking [81] and understanding what problems a computer can and cannot solve [75].

Currently, there is also limited access to AI education, possibly due to gaps in access to computing resources [78], as well as perceived complexity and a relatively new emphasis on AI in computing education [62]. Parents may also be uncertain about how to best support their children in gaining access to computing opportunities [68]. Even in more mature STEM and technical fields, access is not equally distributed, as many opportunities are only available to children in higher-resourced

communities. Additionally, parents, especially from marginalized backgrounds, face challenges in finding out about and enrolling their children in educational STEM programs [36].

While these limitations are real, we interrogate if this means that youth then cannot have more critical agency in their interactions with AI-powered systems in which they are stakeholders. For example, the healthcare system has historically put youth and their medical experiences on the sidelines. However, recent work has explored better supporting youths' agency by rethinking participation they have in a healthcare setting [32]. Hong et al. [31] designed methods for youth patients to have an active role in sharing their narratives in order to better center their experiences and needs when receiving medical care. We think that youths' ability to engage in playing an active role in fair AI systems can follow a similar pattern, which we explore in this work.

2.4.2 Ethical and moral reasoning. Many social systems exclude youth on the grounds that they are neither sufficiently rational nor morally developed enough to participate in serious conversations and processes around defining just futures and policies. For example, almost every country, including the country in which this study takes place, requires that voters be at least 18 years old [70]. Arguments for lowering the voting age have shown how political decisions that children have no say in can actually impact them greatly [55]. Only a few countries have lowered their voting age to 16, where there have been reports of subsequent positive civic outcomes [25].

Do children lack moral reasoning skills? Moral development literature shows that children are able to identify behavior that they view as unfair starting from infancy [17, 69], suggesting that they are indeed capable of strong moral reasoning early on. Starting from the age of 11, children can begin to reason with empathy, awareness of others, and with more complex ideas of fairness [53]. However, much younger children and toddlers at first may not have as developed moral or social reasoning, such as Theory of Mind (the ability to think about what others are thinking and experiencing) [65], which begins at the age of 4.

While very young children are still developing moral reasoning, many youth from slightly older elementary ages through high school do have skills and judgments that are closer to or aligned with adults (e.g., [33, 56]). In some cases, youth are also able to lead as moral and ethical agents. For example, Greta Thunberg has led climate justice initiatives to protect the natural world, where her identity as a young person has played a salient role in her work [73]. With youth being current and future users of AI, we believe that algorithmic justice may additionally involve the engagement and leadership of youth.

2.4.3 Protection from serious topics. Children are often perceived as needing protection from serious topics. In fact, many parents go out of their way to avoid talking about topics around race and bias, despite children as young as six being affected by stereotype threat related to socioeconomic status [20] and gaining awareness of or endorsing stereotypes based on race and gender [38]. Further methods of protecting children include censorship. For example, the Florida Senate passed the 'Don't Say Gay' bill in 2022, which bans schools from discussing LGBTQ+ topics and gender identity in classrooms up to third grade [6], despite many students having gained awareness of these topics already by then.

Parents are also documented to protect their youth *with* technology. Specifically, middle and upper class parents have used technology (e.g., baby monitors, phones, and methods of digital tracking) to surveil and control their children on and offline [47]. Parental control with technology differs by child identity, for instance, girls and boys are treated differently in parents' attempts to shield them from online content and experiences [48].

However, parents' overprotection and avoidance of some serious topics may inhibit their children's ability to contemplate and approach serious topics with critical social thought. Prior work has shown that youth often benefit from addressing serious topics around bias in age-appropriate

ways. For example, when youth have the opportunity to talk about bias and barriers related to STEM access, they are supported in thinking critically about societal systems and can step into the roles of potential change agents and justice advocates (e.g., [4, 19]). And conversely, lack of adequate acknowledgement and avoidance of addressing bias has been argued as perpetuating injustices, such as White supremacy [9].

Overall, there is a misalignment between 'protection' from serious content and granting emboldening experiences toward sociocultural and economic topics that are highly relevant to youth, many of which they are already subject to or aware of. We believe that allowing youth to critically address and discuss unfairness rather than ignore it could grant them, especially those from marginalized backgrounds, agency to define a fairer future where they can be protected from harmful AI behavior.

3 METHODS

3.1 Participant Groups and Recruitment

To collect data, we ran an IRB-approved educational workshop study on AI and fairness with four different groups (30 youth and 6 adult parent participants). The workshops ranged from 45 minutes to 2 hours long in duration, depending on the structure of each program. All data collection sessions took place in a middle-sized city on the East Coast of the United States. We discussed with partnering organizations how we could shape our research to fit the existing or ideal program needs. For two organizations, these were already programs being run to provide STEM exposure, and our session was one of multiple but the only one with a research component. In the program with families, we took a co-design approach with the school to create workshops around their needs. In the rest of this section, we describe each group (de-identified names), the general format of the workshop, compensation, and accompanying information about recruitment. We note that ethics around child compensation in research have been debated, since it can impact incentives and decision making for minors and their parents. However, when calculated carefully, monetary payments compensate participants' time, efforts, and potential inconveniences (e.g., commute) [79]. Taking guidance from prior work with child populations and families (e.g., [7]), we aimed to compensate participants at IRB-approved rates that supported their engagement with our programs and acknowledged their contributions to our research if it aligned with our partnering organizations' goals, values, and structure. Preferences and logistics regarding compensation differed across contexts that we describe below for each workshop.

The first group, "TechView" (N = 8), was with high school girls visiting a private research university with strong STEM programs for a day-long event on a Saturday. The event was hosted by a university student-led organization focused on gender diversity in STEM for middle school, high school, and bachelor's students. Participants attended a number of technology-related lectures and activities, including our workshop (a 45-minute session introducing learners to fairness in AI). To recruit, the organization sent information about the event to local high schools and asked them to forward it to their students. The event and our specific workshop session were free, and all participants who registered by the deadline were admitted. Students were not compensated for their participation, a decision that was made in deliberation with the research team, program organizers, and IRB. The structure of the program was not fitting to compensate participants, as multiple STEM topic sessions were run during the same time blocks in different rooms. Therefore, it would have been unfair if participants who happened to be assigned to our section by the program organizers were paid for their participation in research, when other learners assigned to different sections running at the same time did not also have the chance to do so. Additionally, prioritizing program regulations, we did not gather in-depth learner demographic data, so they are not included in tables with participant demographics. These learners may have come from higher socioeconomic

Table 1. ScienceJam participants were racially diverse, and all were girls or non-binary.

Participant Code	Age	Self-described Race	Prior Self-Described CS and Tech Experience (summarized)
SJ1	11	White	First experience in the ScienceJam Program
SJ2	11	White	Knows how to code, reading a book about AI
SJ3	11	White	Can code a little
SJ4	13	White and Jewish	Family member is a programmer and has conversations about CS
SJ5	11	Asian	Went to a few ScienceJam nights before
SJ6	11	Didn't share	A little bit of programming
SJ7	11	Asian and White	No previous experience before ScienceJam
SJ8	12	Middle Eastern and White	No previous experience before ScienceJam, recently started Girls Who Code
SJ9	11	Black	Played video games (but does not code)
SJ10	11	Black	Girls Who Code summer program and made a video game
SJ11	12	Black	Maybe
SJ12	11	Asian	Not really

backgrounds and higher exposure to STEM given that their parents could transport them to the visit day, and the learners self-selected into the event.

The second group, "ScienceJam" (N = 12), was structured as an after-school program workshop with middle school students with strong backgrounds in computing and tech. Our data collection session was run in a 90-minute session as part of a series of workshops by a student-led organization at the same university as the first group (TechView). The student organization that created this program focused on hosting weekly educational, after-school experiences for middle school students to support gender diversity in STEM, aimed at girls and non-binary learners. To recruit, they emailed school counselors and teachers, asking them to pass on information about the program to families and those who had participated in previous years. In the week prior, the students had a session on training data in machine learning with a focus on a technical understanding of features in data. Anyone who applied to join ScienceJam was allowed to attend. ScienceJam was a free program, and students who participated in our session were not compensated. In an in-depth conversation with the program organizers and a staff diversity coordinator, we initially encouraged payment for participants who attended our session, but they did not see payment as aligned with their goals and were concerned that compensation felt coercive or unfair between learners who did and did not opt into data collection.

The two last groups, "CompuFam A" (N = 7) and "CompuFam B" (N = 9), were the same workshop run twice at a K-8 charter school (publicly funded, independently ran), which was 92% Black in the school and 99% Black in the affiliated youth development center, both of which we recruited from. The school was eligible for free lunches. Workshops consisted of three Black (CompuFam A) and multiracial Black-majority families (CompuFam B) each, with children in middle and elementary school. Parents ranged in their exposure to STEM and computing, and many children had some type of previous exposure to computing but generally much less overall than those in ScienceJam and TechView. The workshops were structured as 120-minute sessions with a break embedded. The research team hosted the family workshops over two Saturdays, which were one of several planned programs to sustain STEM and computing opportunities in collaboration with the school. Different families participated in each workshop. Participants were recruited by the school sending out messages with recruitment text to families. A researcher also tabled at a family night, giving out information to families about the workshop. When we discussed the program design in meetings

Table 2. CompuFam workshops included six families, where the participant codes denote: P for parent, C for child, and A or B for the group they were in.

Participant Code	Parent	Age	Gender	Self-described Race	Prior Self-Described CS and Tech Experience (summarized)
PA1	n/a	38	M	Black	None
CA1	PA1	10	Didn't share	Black	Learning how to code in regular school hours, has used TinkerCAD
PA2	n/a	34	M	Black	None, uses technology like Microsoft word, PowerPoint, and an iPhone
CA2	PA2	13	M	Black	Learned in a camp class how to code and make a picture with shapes
PA3	n/a	30	F	Black	Worked in technology services position and have watched software developers
CA3	P3	9	F	Black	Knows how to use technology
CA4	P3	7	M	Black	Knows how to use a computer and plays Minecraft
PB1	n/a	39	F	White	20 years in IT systems, networking, telecoms, with a military background
CB1	PB1	10	M	Black	Learned a little bit of coding in first grade but not a lot
PB2	n/a	38	F	White	Studied web design and familiarity with HTML, Javascript, and Flash
CB2	PB2	9	F	Mixed	Learned a little coding in classes
PB3	n/a	34	F	Black	A couple classes in coding, basic understanding
CB3	PB3	13	F	Black	Did some coding in school and used a 3D printer
CB4	PB3	11	F	Black	Learned about technology
CB5	PB3	12	F	Black	A little background in TinkerCAD
CB6	PB3	9	F	Black	None

with the center director, compensation was requested for participating families to acknowledge their contributions to our research and help offset barriers to participation. This program also differed from the others, in that we designed and ran the full workshop (rather than adding a session to an already existing program), and sessions were longer than those for TechView and ScienceJam. We used IRB-approved rates: parents were compensated $50 and children $20 for participating.

For informed consent and assent, we shared documents with parents before the programs. There were two forms — parents filled out consent forms, and children, since they were not 18+, filled out assent forms with their parents. After receiving a description of our study through recruitment emails from our partnering organizations, they could read attachments further detailing the protocol and research procedures. Parents and children could also review the same information on hard copies that we handed out prior to the sessions. We requested that parents carefully review the research participation procedures with their children, encouraging them to ask us any questions. If children agreed to the research procedures, they gave assent and signed with the support of their parents. At the beginning of workshop sessions, we introduced our research team and reminded participants that they could opt out of any research activities. When we took photos or audio recordings during the session, we verbally asked for additional permission before doing so, ensuring that participants were aware of data collection and approved it.

3.2 Workshop Material

Workshops as a methodology are used within HCI to support participants in sharing their ideas, needs, and values in technology. They can be particularly fitting in research with children, as

workshops help to navigate power dynamics, emboldening youth participants to have agency in the experience [57]. We took a mixed methods approach to leverage triangulation from individual participants' and group-wide sources of data, namely focus group discussions, design-based artifact creation, short interviews, and observations to investigate our research questions. We designed the study to give insight into which ideas came from children versus adults, or developed over time with parents' support. For example, group discussions were with everyone in the room. While family members in CompuFam A and B sat near one another in the group-wide discussions, we emphasized that parents allow their children to speak first in order to amplify youth perspectives before adult perspectives and intergenerational sensemaking. Each workshop was led by the same researcher with a background in culturally responsive computing, HCI, and AI literacy, with 2-3 research assistants who helped take notes and have follow-up conversations with learners about their ideas and artifacts.

In this paper, we focus on three algorithm auditing activities that were embedded in a larger workshop about bias in AI. Workshop session content was formulated with time allotment in mind and flexibility. In the longer data collection sessions, these activities functioned as the first part of a larger workshop on AI and Fairness, which later addressed more technical topics about algorithms and responsible AI. For the scope of this paper, we focus only on the activities described below, which were the same across all four sessions.

3.2.1 Activity 1: AI and Fairness Discussion. In order to understand what knowledge participants were coming in with, we started with an open-ended group conversation with everyone in the room about AI and fairness. The discussions were facilitated by a researcher who displayed questions on slides. We first asked participants what fairness meant and how they might define it. We then aimed to understand participants' backgrounds with AI by asking them: *"Do you have any artificially intelligent (AI) technology in your home?"* and *"Can you think of any examples of AI?"* After some participants answered the question, we also gave additional examples: digital assistants such as Alexa, selfie filters that use face recognition, and Google Search.

3.2.2 Activity 2: Bias Identification. We then ran an individual activity and follow-up group discussion where we showed participants images of AI and asked if they thought the AI could be "unfair or harmful" in any way (Table 3). We varied the type of known biases in these examples based on prior scientific literature or examples from the news. We also added two bias examples as controls, which were generated through iterative discussions among the research team. Some of these examples are shown in Figure 1, which includes the Google Image 'wedding' search results (Figure 1a), Google Image 'trees' results (our intended control or more neutral example) (Figure 1b), Dall-E generative AI for 'rich doctor walking on a street' (Figure 1c), and Google Search bar suggestions for the input 'why are asian' (Figure 1d). In Table 3, we also note the researcher-evaluated level of harm as a way to give context to the examples and later on in the findings.

Participants individually wrote whether they thought the example we showed could be fair/not harmful or unfair/harmful and a reason why. After reviewing all examples, we discussed possible biases in each example in a group conversation with everyone in the room to debrief the activity.

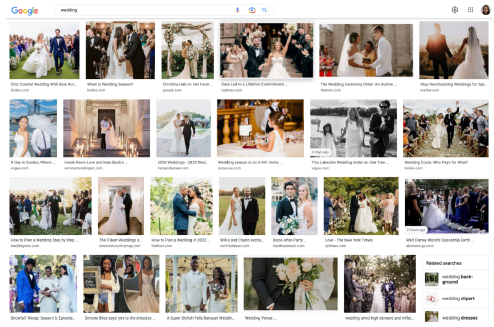
3.2.3 Activity 3: Designing for Surfacing and Mitigating AI Harm. We then ran an individual crafting activity, where we asked each participant to design a technology to report potential AI biases. We printed out the examples from the AI Bias Identification activity (see Table 3) and asked participants to use the images as a base for augmenting with a system, process, or interface where users can raise a concern about an algorithmic system. Parents were instructed to make their ideas youth-inclusive. We provided scissors, glue, markers, and pencils. After they finished crafting, some having explored multiple ideas with different printouts, we discussed their designs in a one-on-one conversation

Table 3. List of AI examples shown in the Bias Identification activity. The Harm Level column is what the research team evaluated the example as to help structure the study and later on findings.

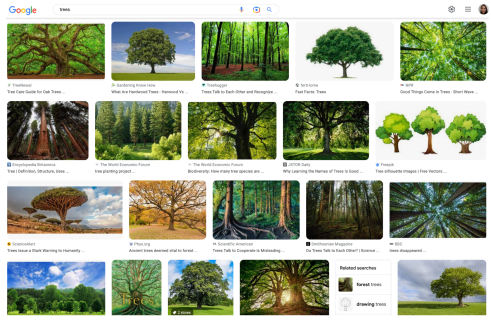
AI Example	Type of Existing Biases	Harm Level	Basis
Coded Bias: Face Recognition, a Black woman's face is not detected by computer vision until she puts on a white mask.	Racial bias and colorism.	High	This was a headliner example that went viral (e.g., [44]), also featured in a documentary [37] and a highly cited academic paper [11].
Dall-E AI image generator: "rich doctor walking on a street"	Only White men shown.	High	This was an exploratory example, given problematic behavior of generative AI in the news (e.g., [24], [1]).
Google Search Bar: "why are asian …", with the first suggestion being "-s so good at math".	Racial and cultural bias.	High	This was an exploratory example, based on headlining articles from news sources (e.g., [41]), referencing the bias in suggestions.
Google Image: "computer programmer"	Lack of women and people of color.	High	This was a common gender-profession based example established in prior literature ([39], [8]).
Google Image: "secretary" (clip art) Note: we chose not show the non-clip art results, due to strong explicit content (shown to all groups except CompuFam B).	Lack of representation of men or masculine people, all White people.	High	This was another similar but different gender-profession example [8], with the opposite effect of the prior example.
Google Image: "wedding"	Nuanced example with no LGBT marriages, other cultures, or physical disabilities shown. People of color are shown.	Medium	This example has been shown in prior work and has a number of different facets [21].
Google Image: "food"	Nuanced example, with some representation and some lacking. There is a lack of cultural diversity in the food but still some variety.	Low	This was an exploratory, nuanced example we found with some representation on a less emotionally salient topic.
Google Image: "trees" (shown to ScienceJam and CompuFam A & B)	Images of trees we evaluated as inoffensive.	Intended control	We searched for an example with as little bias as we could find for a control.

with each participant. We asked them to open-endedly explain what they made, did, or thought about during the activity. In CompuFam A, parents and children were split for this activity, such that all children sat together and all parents sat together at tables. However, we found that this was disruptive for some families, so for CompuFam B, families sat together, but each participant completed individual designs. We requested that parents focus on their own projects and not guide their children so that we could understand adult and child design insights separately. Researchers who went around to answer questions and support during the activity observed that parents were quite focused on their own ideas.

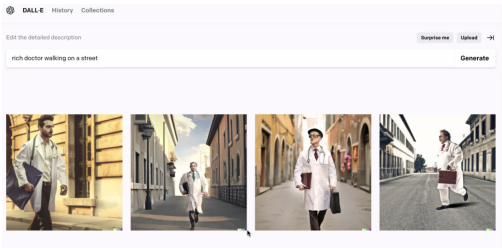
We chose to use a crafting activity for several reasons. First, prior work has used crafting to help scaffold youth in ideating fair AI [22], as it can engage younger participants in deep and deliberate thought. Further, it positions participants in the role of the designer, which is aligned with our goal of offering an emboldening workshop experience. Third, having youth engage in designing a system can investigate not only *whether* they can participate in surfacing harm and



(a) Google Image: "wedding"



(b) Google Image: "trees"



(c) Dall-E AI image generator: "rich doctor walking on a street"

(d) Google Search Bar: "why are asian"

Fig. 1. Some examples we showed for the Bias Identification activity

raising awareness, but *how* they might envision themselves doing so. We not only get to see their direct thoughts about the examples they chose to augment but also additional information about their mental model of auditing processes and stakeholder roles, which we highlight in our findings.

3.3 Data Capture and Analysis

In each data collection session, researchers took detailed notes on their observations of the session. Notes from conversations in the ScienceJam workshop and audio recordings were taken from CompuFam A and B. Photographs were taken of participants' artifacts if they consented to it. Due to program regulations, we did not record anything from TechView, but researchers took notes on their observations and conversations with learners. Furthermore, since some insights are purely from the researcher notes taken during the live data collection, we were not able to mark down all participant IDs for all quotes we gathered.

Using the transcripts from the recorded conversations, artifacts from participants (e.g., crafting projects and lists of answers for the Bias Identification activity), and research notes, we applied an open-ended thematic approach [16] and consensus-based analysis [28] to analyze our data. Data from three participants were excluded from the Bias Identification activity, namely SJ11, CA4, and CB6. SJ11's handwritten answers were illegible to the researchers, CA4 had too many answers with uncertainty as to which was the corresponding AI example for each answer, and CB6's (youngest participant) answers indicated that there was confusion about the task. Three researchers reviewed the data for each session and separately developed different themes. They then discussed the final themes iteratively until a consensus was reached. We present these themes in our findings below.

3.4 Research Approach and Positionality Statement

For this project, we worked with participants with marginalized identities. We recognize that there are a number of concerns for research with underserved communities, in which there can be an unfair exchange between researchers and participants, where researchers gain more than the communities, particularly in the short-term [29, 54]. To address this, we collaborated with organizations that had ongoing opportunities for youth involvement, aiming to prioritize what the programs needed and fit our research around those needs. For example, we tailored each data collection session to fit the organization's standards and requests. In the school where we ran the CompuFam workshops, we also began a longer-term collaboration to support the development of educational technology offerings. Given the school's communicated aspirations, we aimed to fill a gap in supporting their staff in creating educational computing literacy experiences for learners and families with the workshops serving as part of this effort.

We also realize that our own identities and backgrounds impact the work we carry out. The team of researchers for this paper comprises women with a range of identities and academic backgrounds. Our academic domain specialties include computer science, learning science, design, culturally responsive computing, and responsible AI. We come from diverse socioeconomic backgrounds; our racial demographics include Asian, Middle Eastern, and White, with American, Asian, and Jewish cultural backgrounds. While one author has deep knowledge of the city in which the data collection took place, none of the authors have extensive experience with the area in which the school for CompuFam was located nor identify as Black. Our affiliation with a private research institution may have impacted power dynamics and data collection, and our team consisted of all adults, while most of our participants were youth.

The team took care to iterate heavily on the workshop content, including reviewing material with stakeholders from our community partners. We aimed to introduce serious ideas in ways that were both youth- and family-friendly. For example, the search results of 'secretary' were too explicit for children, which is why we opted for clip art. We had extensive conversations within the team, with one author having had training in culturally responsive teaching and two authors being parents of young children. Centering the participants was a core value of the work carried out.

3.5 Limitations

Discussion is integral to the workshop format of our data collection sessions. However, sometimes this can lead to a lack of clarity about which participants know what and who holds which opinions. For workshops where parents were present, it is possible that parent views impacted their children's ideas. We also acknowledge that our own backgrounds, experiences, and biases as researchers can impact the results of qualitative thematic analysis. Finally, while we strove to get a diverse range of participants, all participants were self-selected in some way, with participants from TechView and ScienceJam opting into STEM and computing-focused programs, and CompuFam participants showing interest and opting into the computing workshops.

4 RESULTS

In this section, we introduce a discussion of fairness both with and without the context of AI behavior, as well as how youth participate in algorithm auditing activities. Themes are based primarily on youth data, given our effort to emphasize the importance of parents initially refraining from guiding their children in the CompuFam sessions and letting the young participants express their views first. Parent data is specified when applicable. After engaging in group conversations, younger children's understandings were not just supported by their parents but also other participants,

including peers and parents of other children, which highlighted the impact of parental and peer support in promoting comprehension and involvement.

We saw that *equality* was the most common idea of fairness amongst all participants, with slightly more complex ideas for older participants (e.g., equity vs. equality). We next report on (1) how participants recognized bias in AI, suggesting their high sensitivity to detecting bias in even more nuanced examples, (2) youth-inclusive algorithm auditing ideas and (3) possible constraints of algorithm auditing for youth, including exposure to societal systems. For the first topic, we examine how participants noticed and described biases that they found in examples from the Bias Identification activity. Next, we discuss common algorithm auditing ideas presented by our participants, and finally, we share certain challenges that we saw in youth engaging in this process as well as differences in ideas between youth and parents. Themes and subthemes presented in the following sections are illustrated with examples from our data.

4.1 Youth Perceiving and Recognizing Bias in AI (RQ1)

Youth mostly thought of fairness as equality but also considered some more complex ideas such as equity, especially if they were older or had higher prior knowledge in computing. When it came to detecting problematic behavior in AI, youth were perceptive and articulate, with many discussing bias that they saw in the examples in great depth. Sometimes higher-harm examples sparked passion toward changing or providing feedback about the AI. For context, all participants agreed that they had some kind of AI in their homes or on their phones. Many could also name examples of AI, which included physical forms of AI or apps and features on apps, such as "robots" (CA2), "smart locks" (CA3), chatbots (SJ4), and face filters (SJ7).

4.1.1 Ideas of fairness emphasized equality. In asking participants about how they identified fairness, *equality* was usually the first and most common idea to arise, with a participant from the TechView group saying, *"everyone gets treated the same."* Treating others well, e.g., *"equality and respect"* (SJ1), came up across youth and family groups as well. Other ideas included inclusivity (e.g., *"Unfair would be not hanging out with people"* (CA3)), or consent (e.g., *"Something that is unfair is collecting user data without telling them"* (SJ4)). More complex ideas related to equity also came up, generally raised by older participants (such as the high school students and parents), which took into account people's different backgrounds, e.g., *"[standardized tests are] not fair for everyone, because not everyone has the same resources. It's the base standard for everyone, but it's not actually fair"* (PA1). Younger participants tended to agree after they heard these more complex ideas, even if they did not think of them on their own at first. This included if it was from an adult other than the child's parent, e.g., another child's parent or older peers. Many of these ideas are reflected in prior literature (e.g. [64, 67]) but help to add context to the rest of the findings.

4.1.2 Strong potential of youth in identifying and articulating biased and harmful AI behavior. From the Bias Identification activity, we found that the majority of the participants' decisions about whether an example exhibited potentially harmful bias matched the existing evaluations of those cases from prior literature. Figure 2 shows the detailed results of participants' perception of fairness for each case. Overall, we saw that the AI examples with the most salient race-based bias also posed the highest level of harm (e.g., Coded Bias: Face Recognition, Google Search: 'why are asian', Dall-E: 'rich doctor walking on a street') resulted in the most participants feeling that it was unfair.

In addition to the ability to identify algorithmic bias, the majority of participants were able to articulate, often in great detail, how the examples we showed contained bias. All participants thought critically about the examples shown, especially the groups of girls with higher prior exposure to computing and STEM. For example, when viewing the Google Images: 'wedding' results, one middle school participant (SJ6) pointed out that, *"There's only young people. Also, only*

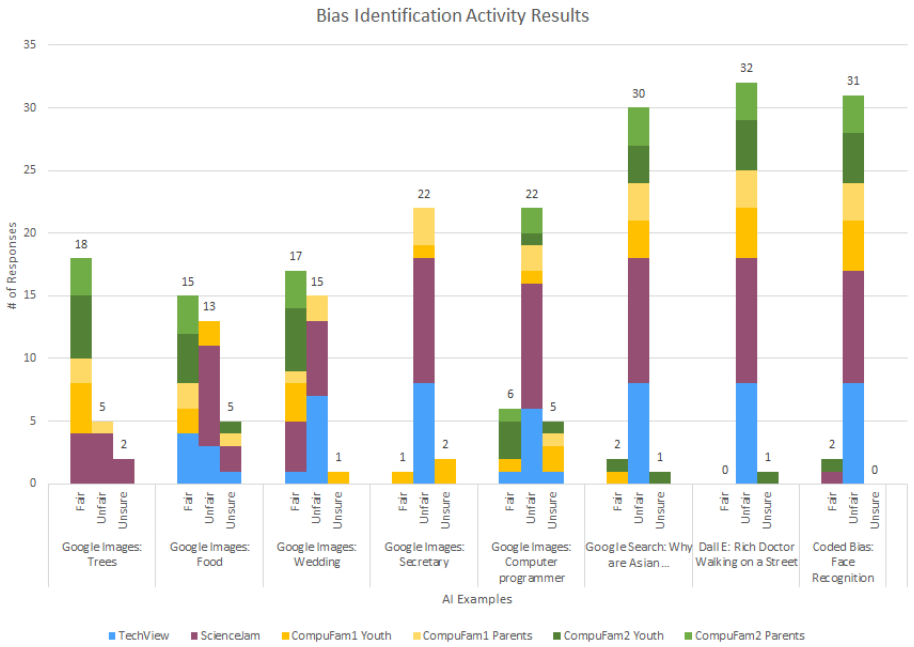


Fig. 2. Graph shows the distribution of fair and unfair decisions, broken down by group. Due to the iterative nature of the workshops and differing structures of the programs, not all groups saw all of the same examples.

White people are marrying to White people, and Black people are marrying to Black people. There are also no gay or lesbian marriages." SJ7 added, *"Also everyone is wearing the traditional white dress, black suit. Most people don't have completely traditional marriages."* This depth of understanding and articulation of various types of bias, even in nuanced cases such as the 'wedding' example, was both surprising and promising in revealing the ability of youth to navigate the complexities of harmful algorithmic behavior.

We further saw that even for examples that we chose as lower harm, participants found bias to surface, e.g., when looking at the Google Images: 'food' example, one participant (SJ6) mentioned that there was *"mostly only food from [American] culture, which is not the most fair thing."* This same sentiment came up in the TechView group. Additionally, we saw that a number of people were able to discuss potential bias in the Google Images: 'trees' example, which we believed was the most unharmful example. In two of the three groups the example was shown to, CompuFam B and ScienceJam, participants noted the lack of diversity among the trees depicted, which were green, visually appealing, and healthy-looking, as well as appearing to be from certain climates that excluded examples such as desert trees. This was seen as bias by some of the participants.

4.1.3 Recognition of harm severity. Participants agreed that the more obvious and higher-harm examples were unfair, with a few exceptions (e.g., CA4, who was the youngest learner of the entire study). Youth were not only capable of detecting bias but able to understand the extent of possible harm from more extreme examples and how it could be detrimental.

For the higher-harm biased AI examples, we saw some participants display strong emotional reactions, such as anger, disgust, or disbelief. This was especially the case for the face recognition example from Coded Bias and Google Search: 'why are asian', with multiple people from all groups

describing the AI as *racist*. For example, SJ10 wrote on her Bias Identification answer sheet *"NO, NO, NO, she is Black but face [identification] doesn't work, but as soon as she puts on a WHITE face mask, she is detected (makes me MAD)."* Even if the AI example was based on a 'positive' stereotype, (e.g., Asians being good at math), youth still believed this to be harmful due to the inaccuracy of the stereotype: *"Just too stereotypical and racist. There are Asians that are not good at math."* (CB1). Learners also stated that 'positive' stereotypes for one group can also result in perceived negative stereotypes for others: *"They are making it seem like no other race can be good at math,"* (CB4).

There was less agreement and more uncertainty about the level of harm, as expected, for the nuanced and control examples, in particular, the Google Image search results for 'food', 'trees', and 'wedding.' In the original discussions, at least one participant in each group could point out how there was bias in these examples. However, in the follow-up discussion, some participants noted that they were unsure or believed they were fair. For example, there were some mixed opinions on the 'wedding' images, in line with the content, which had some diversity (e.g., people of color), but was lacking representation in other ways (e.g., omitting people of color, non-traditional marriages). We noted a potential misconception that younger children might have, which was conflating harm with the actual content. CB1 and CB2 suggested that the 'wedding' example was fair due to the image content having "happy" people, while no parents or older children reasoned this way.

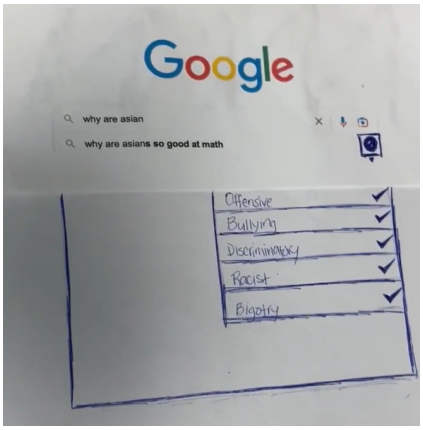
4.2 Ideas and Designs for Surfacing and Mitigating AI Harms (RQ2)

After participants identified algorithmic harms, they became deeply engaged in ideas for reducing such harms. They organically brought up an ongoing debate in AI fairness: *Should technology reflect a harmful, but existing, societal bias or an ideal future state?* [50] Participants also believed that in most cases, the creators of the technology (e.g., the companies) were responsible for fixing it, especially for higher-harm scenarios. For more nuanced scenarios, they believed users also had the agency to work around problematic AI behavior.

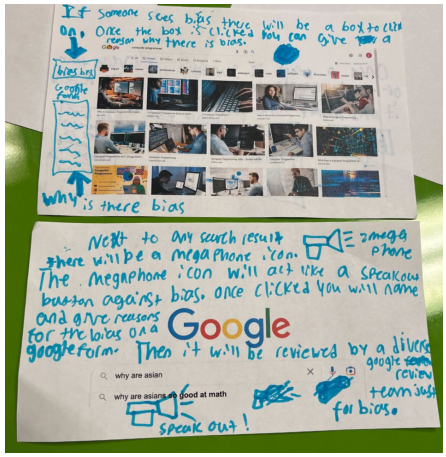
When it came to designing systems for surfacing and mitigating harm, participants came up with a variety of solutions, from providing report and feedback mechanisms to surface harm, to giving users agency to adjust algorithmic outputs by introducing diversity and inclusion to such outputs, as well as suggesting additions where the AI technology itself could contribute to raising awareness among users about its shortcomings and potential harms.

4.2.1 What should more fair AI look like? From current to ideal state. When we showed Google Image search results for 'programmer,' which showed mostly men, all four groups organically (without prompting from researchers) raised the question of *'Should fair AI (e.g., in the case of search results) reflect the current state of things or an ideal state?'* One high school student from the Techview group who was unsure about whether or not this example was fair stated,*"If there are mostly men who are programmers, then that's what Google will show."* This prompted further discussion about the sources of AI bias. In the student-only groups (TechView and ScienceJam), youth hypothesized that these biased AI results may be based on past and current biased norms and data, while in the CompuFam workshops, parents (PA1, PB1) usually raised this idea. For example, SJ2 considered the role of current biases in forming biased results, suggesting for the Google Search: 'why are asian' example that she was *"pretty sure Google bases these off of actual searches."*

As a result, some youth brought up the idea of mitigating such existing biases in the input of AI systems to avoid biased outputs: *"There are biased training data [leading to] algorithmic bias. We need to be able to stop the gender, race, and age discrimination within those algorithms so that it's a better and more accurate presentation"* (SJ4). Interestingly, SJ4 continued to state that this revision of inputs to algorithmic systems can actually make a difference in the future, like a feedback loop that can improve the representation of a group (e.g., the secretary case) in the society, moving



(a) PA1's algorithm auditing idea, where users can report and select why a search result may have an issue via a drop down menu, including: offensive, bullying, discriminatory, racism, bigotry.



(b) CA2's algorithm auditing idea, which includes a report process that gets sent to a "diverse Google review team"

Fig. 3. Ideas about surfacing and mitigating bias from CompuFam A with a parent (left) and child (right).

towards an ideal state: "*If we could change the algorithms to be less discriminatory, then we can also influence how people see this job.*" Such nuanced understanding of taking actions toward mitigating AI harm and its impacts was also demonstrated in other groups such as TechView, in which some high school students talked about moving from outdated biases from past times (e.g., stereotypical gender roles) towards a more fair and bright future with women and girls being involved in the forefront of STEM (and not just working as secretaries). This also reflected as the need for *systemic change*, as participants connected personal experiences of marginalization to the examples of AI. For example, using the Google Search: 'why are asian', CA4, a young Black student, demonstrated a strong interest in combating stereotypes by drawing an image depicting supporting more Black children in learning math.

4.2.2 Who should mitigate AI harm? Perspectives on the responsibility of the creators and users. Amongst AI examples that were higher harm and less nuanced, participants believed that the creators (i.e., the technologists and companies) were responsible for fixing it. Some participants also acknowledged that the companies were responsible for creating the bias, as reasoning about bias in the examples was tied back to the human creators. For example, in the follow-up conversation in the Bias Identification activity, SJ7 talked about the Coded Bias example, suggesting that *"clearly someone has taught it to only recognize people with lighter skin, which is not fair."*

In terms of user responsibilities, some participants believed that users had the agency to work around biased AI, mainly in lower-harm cases. When participants thought that an example was fair, a common reason was that the user could have better specified what they wanted from the AI-driven technology. This was the case mostly for more subtle or nuanced examples. For instance, PA1 suggested that the Google Image: 'wedding' example showed *"general"* results and that the user could better describe in the search bar what they wanted to see. Similar arguments were made for 'trees' and 'food' Google Image search results.

4.2.3 Giving users the ability to report potential AI biases. When it came to designing systems for surfacing and mitigating harm, participants repurposed existing design features on apps that

they had previously encountered. Reporting and feedback features were the most common in the crafting activity. For example, SJ2 suggested adding a report option, which would trigger an email to the creator to *"complain if [the AI was] unfair."* Parents also discussed this feedback mechanism as a design option. For example, PA1 created a drop-down menu, with different reasons for why the results of the AI were an issue (Figure 3a).

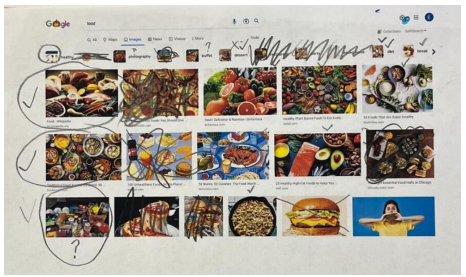
Most participants, however, were not sure what would happen to the feedback once users submitted it. Even parents expressed uncertainty: "*That's above me*" (PB1). To tackle this, a few participants suggested that companies should have dedicated and diverse people to support fairer AI. CA2 ideated a system that could take into account user feedback to *"be sent to Google, and it'll be reviewed by a diverse support team that works on just making sure there's not a lot of bias in Google"* (see Figure 3b).

Many also considered system designs with open-ended text boxes, where users could either report problematic behavior or add their opinions to affect the future behavior of the AI. In other words, participants also ideated how suggestions could be taken into account for what the user *did* want to see, not just avoiding bad outcomes. This led to discussions around providing users with the agency of changing and adjusting algorithmic outputs that we describe below.

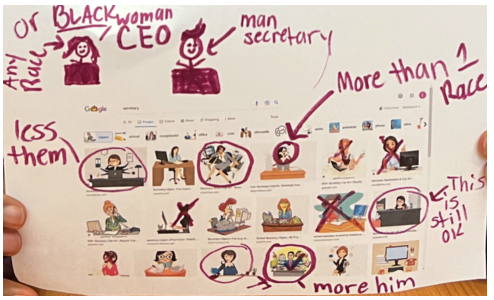
4.2.4 Giving users agency to adjust algorithmic outputs. In the designs, participants demonstrated the need for not only reporting but also mitigating harm by gaining control, typically through user preference options. For example, CB2 made an interface that offers the option of removing algorithmic outputs that *"offended"* her using the Google Image: 'food' printout as a base (see Figure 4a). A high school student suggested a feature in which a user could click a search image to allow for a more specific search. Another high school participant created an interface option with a toggle, where the user could decide if they want to see 'ideal' or current research results, with the Google Image: 'programmer' example printout as a base. Adjusting algorithmic outputs also came up in all of the parents' ideas. PA2 described a system, using the Google Image: 'secretary' clip art search results printout as a base, where the user could change the skin tone in the art, *"like Emojis."* Ideas like this suggest giving users control to curate algorithmic output that is desirable to them.

In their desire to revise algorithmic output, the majority of participants aimed to curate more representative and inclusive outputs. Many participants crossed out and suggested replacing some algorithmic outputs to include more diverse ones. For example, SJ7, in the secretary example, stated that she wanted the ability to revise the results to see *"less of women, more of man! More people of color"* (Figure 4b). Others brought up similar ideas about adding more diverse and inclusive results in other examples, such as *"less fancy weddings, weddings that are more multi-racial and LGBTQ+"* (SJ8) for the Google Image: 'wedding' example, or considering "*disability, women, [...], people of color*" (SJ11) for the Dall-E case. We saw a connection in this need for diversity and inclusion in adjusting algorithmic outputs with participants valuing equality as a measure of AI fairness from the outset of the study.

In addition to diversity and inclusion, we noticed that some participants also wanted to have the ability of adjusting algorithmic outputs to bring more "*realism*" (SJ2) to the surface. For example, when suggesting changes to the Dall-E rich doctor example, SJ2 stated that she would want to see also "*tired*" doctors. In another example, SJ4 stated that the way the Google search results for the 'secretary' are depicted "*is not really realistic*," and suggested more accurate presentations of the occupation. This also was discussed in other forms, such as "*moving around and not just at their desk with phone*" (SJ3), versus the current search results in which most secretaries were sitting behind their desk. The high school and middle school groups also brought up that images in the 'computer programmer' research results contained misrepresentations of people looking like they



(a) CB2's algorithm auditing idea. Participant mentioned, "Click x [on the images], because it makes me feel offended. It makes me feel like, 'I'm hungry maybe I should go get a fast food.' I want healthiness."



(b) SJ7's drawing, showing what she would change about the current AI example, emphasizing diversity.

Fig. 4. Ideas for about surfacing and mitigating bias from CompuFam B (left) ScienceJam (right).

were engaging in 'hacking' with the black screen and green text. They believed that this was not accurate and thus doing harm by spreading misinformation about what programmers actually do.

4.2.5 Giving users awareness about the presence of potential AI harm. While participants asked for adjusting algorithmic outputs to mitigate potential harm, some were also aware that not all problematic AI behavior is easily fixable. Therefore, a few participants in the middle and high school groups designed systems where the AI would have transparency, giving a warning message or caveat for the results. For example, one TechView student's idea showed a large red warning message at the top of the search that suggested that the displayed results could be outdated.

4.3 Challenges and Support Mechanisms for Engaging Youth in Encountering Algorithmic Bias and Harm (RQ3)

While participants had critical insight, and both children and parents saw youth as capable of participating in surfacing harm and raising awareness, there were sometimes challenges for younger participants in engaging fully. This was particularly the case for CompuFam A, the group with the youngest participants. For example, during the individual ideation activity, the younger learners (e.g., CA3, CA4), could not think of approaches, interfaces, or systems for algorithm auditing. However, even among participants who did not think of any design processes for algorithm auditing, there were discussions about bias, fairness, and systematic changes in society (e.g., CA4 suggesting supporting more Black children in learning math).

Throughout the workshops, we noticed that younger children in the CompuFam groups tended to involve their parents, the workshop facilitators, and their older peers more than the other kids (i.e., the middle school and high school students). Specifically, we found that *prior knowledge* and *lived experiences* helped with critical thinking around algorithmic bias, which are areas that parents or older peers can support in.

4.3.1 Peer and parental support in providing prior knowledge and lived experiences. We observed that the most nuanced critiques of AI came from participants who had more prior exposure to STEM and technology across groups. In other words, those who had more computing experience were more likely to hold techno-skeptical views and find it easier to identify harm or bias in AI examples. However, even those who did not have much exposure were able to surface bias in the higher-harm examples. In some group conversations with just youth (in TechView and ScienceJam), often a higher prior-knowledge participant sparked critical thought and conversation, supporting

further understanding and engagement for those with initially lower prior knowledge, ultimately leading to more careful reasoning about harmful AI.

Parents also had awareness around topics that were less familiar to children, e.g., commercialization and economic systems, which they were able to relate to fairness in AI. For example, PA2 voiced concern that companies had *"corporate America[n]"* values, with AI prioritizing advertisements *"like a big ad"* and would not work to empower people, despite it being part of their responsibility.

Parents more often tied back their personal experiences to the topics of AI we discussed in the workshop, as well as connected both personal experience and AI fairness to larger social structures. This contributed to more collaborative sensemaking amongst the other parents and youth in the group. For example, PA2 discussed being *"the only Black person"* in many settings and how that affected her experiences. She went on further to describe that *"there might be a lot of minorities that aren't in … higher up roles, … decision-making and whatnot … [Minorities are not the ones who] control these algorithms basically."* PB1 also described how her views of bias were shaped by *"being a woman in IT."*

4.3.2 Youth-inclusive designs with sensemaking support. When asked if they wanted to be a part of surfacing harmful behavior in AI, participants with higher prior knowledge generally agreed eagerly, while participants with lower prior exposure were more hesitant to see themselves as having a part in algorithm auditing. Despite this difference, when we reached the individual design and crafting activity in the workshop, no one mentioned that youth should not be a part of algorithm auditing, and most either ideated a system or interface intended to be youth-inclusive or engaged in the act of mock algorithm auditing, noting bias that they would surface in the example printouts.

While parents were prompted to make youth-friendly features, no parent made youth-*specific* features, with PB3 describing her technology idea as suitable for *"all ages."* However, some of the designs that adults made included words that may not be understood by some youth, such as *"bigotry"* (PA1). Middle and high school participants also did not make youth-specific features. Only one TechView participant considered a youth-friendly reporting, warning, and feedback interface, in which these features were cat-themed and included a supportive cat character to help with sensemaking and receiving user opinions.

In a conversation with two participants in TechView, they suggested that since youth are using AI unsupervised, they are capable of raising awareness of issues in the AI behavior on their own without necessarily needing support from family. However, one other Techview participant suggested that younger children may have different norms for what is deemed appropriate or harmful. They explained that younger children may consider the word *"crap,"* be a problem, while older users would not consider this as harmful, which could result in unnecessary reporting. The student concluded that there may need to be systems in place to avoid this problem from occurring.

5 DISCUSSION

5.1 Notions Against Youth Participation

In this work, we engage three perceived notions for why youth might not be included in the discussion and design of responsible AI: lack of technical knowledge, lack of ethical knowledge, and need for protection. We found empirical evidence that in all three cases, participants could contribute to algorithm auditing or critical conversation around AI fairness and often had capabilities beyond what we expected. We saw that youth were sensitive and articulate about bias that they saw in AI, in line with what adults are capable of [21].

5.1.1 Technical knowledge. One potential misconception is that *children without technical knowledge cannot understand fair AI and thus cannot contribute.* However, we saw that all participants aged 11

or older could identify bias in AI and contemplated complexities in what fairness looks like in AI. They showed the capacity to participate in algorithm auditing, either by coming up with a system idea for auditing to raise awareness or by participating in giving direct feedback and surfacing bias in the examples provided. At the same time, all participants, regardless of age, were uncertain how their suggestions could exactly be implemented. For example, participants were also not sure or did not think about where potential user feedback went.

We did see that participants with higher prior technical knowledge produced more critical responses in recognizing bias and potential harm in nuanced cases, like the Google Images: 'food' example. These participants also referred to the role of the users in creating biased systems by introducing biased input data to the system, in addition to the role of people who created the AI. Older participants, especially the parents, were more able to articulate and tie their knowledge of larger social systems and lived experiences to the technology.

Overall, even youth without prior technical experience were able to identify bias in AI and were more likely to do so with more blatant biases. Their limited technical understanding was primarily visible in their attribution of AI bias entirely to the developers. We believe that youth can contribute to algorithm auditing even with this limitation, such as in systems aimed at everyday users.

5.1.2 Ethical and moral reasoning. Another potential misconception hindering youths' participation in responsible AI is that *children's ethical and moral reasoning is not sufficiently developed.* However, the moral development literature would suggest that children as young as 11 begin to view situations from the perspectives of others and decide fairness for themselves [53], and even infants are able to identify behavior that they view as unfair [69]. We saw this played out when children described ideas of what was fair and unfair. In line with the research, younger children described simpler ideas of fairness (e.g., equality), while older children or adults were able to describe more complex ideas (e.g., equity) [33, 35, 56, 61]. However, younger participants were able to contribute to discussions using more nuanced ideas of fairness once those ideas were raised by older peers or adults.

Even simpler ideas of fairness, when applied to AI, had value. For example, many participants thought of fairness as equality, which is typically seen as limited by adults. However, in our AI examples, this concept materialized as a commitment to diversity. Using the equality lens, participants advocated for darker-skinned faces to be as well-recognized by computer vision as lighter-skinned faces, among other examples.

Additionally, youth had critical insight beyond what the adult researchers had predicted. For example, some of the ideas related to our control example included concerns about the trees pictured not representing a variety of different climates, as they were all green and lush. Youth could further bring up researcher-debated topics around critical ethical considerations, such as to what extent AI should reflect current vs. ideal states. The insight that parents or older youth might have when they contemplate fairness together with younger youth can help to support some age-related norms around ethics and appropriateness, such as understanding that certain words may be considered inappropriate by younger individuals but not older people or the broader population.

5.1.3 Protection from serious topics. The last potential barrier is that *children may not be able to grapple with serious justice-related topics, such as bias, and need to be protected from them.* We found that children were indeed affected by exposure to bias, sometimes having emotional responses to the AI examples. We saw that the strength of the reaction was related to the severity of the example, e.g., many children who saw the example about Google Search: 'why are asian', whether they were Asian or not, were shocked. Children expressed their opinions in the Bias Identification activity, articulating how it made them feel and why, showing resilience and informed opinions that were supported by their emotions.

Despite some initial distress, our participants felt that they could engage analytically, even if it meant exposure to bias. However, we note that our participants had opportunities to express their feelings and have them validated by the research team. Should children come across these experiences elsewhere, opportunities to express their feelings and opinions are important for informing a desired future. We see algorithm auditing and other analytic processes as possible outlets for negative feelings, rather than having them be a reason to exclude youth participation.

Finally, we found that the majority of parents in our sample were positive toward having their children participate in our activity, despite the exposure to content about bias. This runs counter to prior work which shows that parents believe children, especially girls, need to be protected online [48]. We suspect this may be related to our participant pool. Black families discuss bias with their children early on, since Black children are already experiencing it. These discussions include serious topics, such as authority violence in 'the police talk' many Black parents give their children [72]. Even though they bring up difficult feelings, they are meant to protect their youth. Along similar lines, Black parents may see a discussion about the harms of technology as a way of protecting their children. With this interpretation, we saw that parents felt algorithm auditing was important, because it improved the technologies their children would interact with. In other words, youth engagement in algorithm auditing is a way to create a better future, protecting youth in a different way.

5.2 Empowering Youth Agency and Action around Bias in AI Systems: Design Implications

We next describe the design implications of this work for potential systems and outcomes.

5.2.1 Solitary vs. collaborative designs. We saw that critical conversation with older and younger participants, as well as those with higher and lower prior computing knowledge, were generative of more sophisticated insights and contemplation about ethics and the role of AI. Sensemaking has been shown to be supported in groups when people can discuss together (e.g., [3, 52]). Contrasting with prior literature where groups with only younger youth (fifth and sixth grade) with lower prior computing knowledge sometimes struggled to consider the responsibility that practitioners have in responsible AI [67], we observed that when youth with lower prior discussed in groups with parents and peers with more computational exposure (and thus techno-skepticism), discourse evolved to more critique of AI systems and societal impacts. This work adds onto prior work, which explores how children and parents may learn about AI together, by emphasizing ethics and bias in AI [23]. Future systems may leverage youth learning from family and peers.

5.2.2 Parent endorsement and generalizability. Parent endorsement of systems is important, since they are gatekeepers for youth and technology. Parents' suggestions that their algorithm auditing solutions could be applicable to all suggests their endorsement of their children's engagement with surfacing and mitigating algorithmic harm. Although this is a situation where parents may have been unusually accepting of children's engagement with algorithmic bias, we may have discovered a place where strength lives at the margins – as prior work suggests White families and those with higher socioeconomic backgrounds may be more likely to protect their children from serious topics [47], it isn't clear that such parents would be as willing to expose their youth to contemplating topics of bias.

5.2.3 Lived experiences of marginalized youth. We found that the lived experiences that girls and Black youth, along with their families, experienced in encountering societal biases empowered them in identifying nuanced and complex AI harms. This is in line with recent work on everyday algorithm audits that shows how people search for and understand biased algorithmic behavior

is heavily informed by their prior experiences with bias [21]. This highlights the importance of involving youth with diverse and marginalized backgrounds to enable identification of different types of AI biases and harms.

5.3 Future Directions

As a next step, we can use the insights from this work to prototype algorithm auditing systems with youth to understand how they might interact with them around potentially harmful AI. This could yield data to more directly inform the design of these systems, as well as more information about possible constraints, strengths, and preferences of youth in algorithm auditing.

We also believe we can engage AI developers themselves in the work to gain insight on new ways to incorporate feedback from youth. Creating engagement opportunities between children and the creators of technologies that are popular among youth (e.g., social media platforms, educational websites, and gaming environments) is a critical step toward youth-inclusive responsible AI. This could involve adding features that allow youth users to report suspected algorithmic harms or providing transparency into how the algorithms work.

5.3.1 Facilitate collaborative learning and peer-to-peer support. Taking into account our design implication of solitary vs. collaborative designs, such systems may incorporate elements where youth can discuss together or see opinions of others on potentially biased content. There may also be youth-inclusive pathways toward users working together to engage in algorithm auditing systems. Recent prior literature highlighting the power of collectiveness and collaboration suggests that the ability to make sense together can be vital to harm recognition and mitigation processes [21, 63]. This can be achieved by integrating features, such as shared workspaces, group chats, or discussion forums within the tools they use. Another direction would be creating a community-driven platform where users can collaborate on identifying and addressing algorithmic harms, exchange ideas, and learn from one another. By fostering a community that understands the importance of AI ethics, young people can collaboratively audit algorithms and share their findings with others, contributing to the development of better AI systems. This will not only help young people develop their algorithm auditing skills but also foster a sense of responsibility and ownership for the digital environment they are part of.

Regarding our design implication of parent endorsement and generalizability, we believe it is important to incorporate parents as stakeholders in the design of youth-inclusive systems and hope to explore how families and parents can play a role in supporting children in engaging with fairness in AI. Systems may be parent-collaborative friendly as well, such that elements of systems may also encourage discussion with parents about potential bias, so that youth can process the content at hand in a way that leverages additional insight from parents. For example, we might explore asymmetric designs where youth produce audits and adults validate their data. We might also consider cases where parents may not be supportive of their children's emotional reactions to bias, building in systems for youth to share their feelings as part of the audit process.

5.3.2 Fostering AI literacy and ethical awareness among youth. Lastly, as our work highlights how many children may have a solid grasp on complex topics *and* they are still learning about bias and how to talk about it, youth-inclusive systems could also teach them to think critically about content, i.e., they could learn from the system itself about recognizing and contemplating algorithmic bias. This could be by describing the concept of algorithmic bias, showing them examples of algorithmic bias, or asking them more introspective questions that get them to reflect on what bias might look like to them in certain AI-driven technologies. As youth become more AI-literate in this way, they can better develop the ability to make informed decisions when interacting with AI-driven systems. This understanding supports identification of instances where algorithmic harms may be occurring

in order to take appropriate action, such as reporting concerns, seeking further information, or choosing alternative tools or platforms.

6 CONCLUSION

In this study, we sought to investigate how diverse and marginalized youth, potentially supported by their parents and peers, can perceive biased AI, as well as how they ideate possible future systems that are youth-inclusive to contribute to surfacing harmful AI-driven behavior. We find that youths' critical insight and emboldened opinions should not be ignored, as they can contribute valuable information about potentially harmful behavior in technology. Ultimately, the question is not if, but in what ways can youth as stakeholders have the opportunity to be a part of what fairness looks like in AI?

ACKNOWLEDGMENTS

We thank the learners who participated in this study, the student-led organizations we worked with, and the school that we partnered with. This work is supported by the Jacobs Foundation through the CERES Network and the NSF DRL-1811086.

REFERENCES

[1] Noor Al-Sibai and Jon Christian. 2022. That AI Image Generator Is Spitting Out Some Awfully Racist Stuff. *Futurism* (2022).
[2] Monica Anderson. 2022. Teens, Social Media and Technology 2018. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/
[3] Golnaz Arastoopour Irgens and JaCoya Thompson. 2020. "Would You Rather Have it be Accurate or Diverse?" How Male Middle-School Students Make Sense of Algorithm Bias. (2020).
[4] Catherine Ashcraft, Elizabeth K Eger, and Kimberly A Scott. 2017. Becoming technosocial change agents: Intersectionality and culturally responsive pedagogies as vital resources for increasing girls' participation in computing. *Anthropology & Education Quarterly* 48, 3 (2017), 233–251.
[5] Joan Palmiter Bajorek. 2019. Voice recognition still has significant race and gender biases. *Harvard Business Review* 10 (2019).
[6] Paige Hamby Barbeauld. 2014. Don't Say Gay Bills and the Movement to Keep Discussion of LGBT Issues out of Schools. *JL & Educ.* 43 (2014), 137.
[7] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).
[9] Eduardo Bonilla-Silva. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers.
[10] Diamond Y Bravo, Julia Jefferies, Avriel Epps, and Nancy E Hill. 2019. When things go viral: Youth's discrimination exposure in the world of social media. In *Handbook of children and prejudice*. Springer, 269–287.
[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
[12] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.
[13] Sarah Childress. 2014. danah boyd: The Kids Are All Right. *PBS Frontline* (2014).
[14] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.
[15] Merijke Coenraad. 2022. "That's what techquity is": youth perceptions of technological and algorithmic bias. *Information and Learning Sciences* ahead-of-print (2022).
[16] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[17] Audun Dahl. 2020. Young children use reason, not gut feelings, to decide moral issues. *Psyche* (2020).
[18] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of Data and Analytics*. Auerbach Publications, 296–299.
[19] Louise Derman-Sparks and Julie Olsen Edwards. 2019. Understanding anti-bias education. *YC Young Children* 74, 5 (2019), 6–13.
[20] Michel Désert, Marie Préaux, and Robin Jund. 2009. So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education* 24, 2 (2009), 207–218.
[21] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
[22] Daniella DiPaola, Blakeley H Payne, and Cynthia Breazeal. 2020. Decoding design agendas: an ethical design activity for middle school students. In *Proceedings of the interaction design and children conference*. 1–10.
[23] Stefania Druga, Fee Lia Christoph, and Amy J Ko. 2022. Family as a Third Space for AI Literacies: How do children and parents learn about AI together?. In *CHI Conference on Human Factors in Computing Systems*. 1–17.
[24] Benj Edwards. 2022. New Meta AI demo writes racist and inaccurate scientific literature, gets pulled. *Ars Technica* (2022).
[25] Jan Eichhorn and Johannes Bergh. 2021. Lowering the voting age to 16 in practice: Processes and outcomes compared. *Parliamentary Affairs* 74, 3 (2021), 507–521.
[26] A Epps-Darling. 2020. How the racism baked into technology hurts teens. *The Atlantic* (2020).
[27] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 153–162.
[28] David Hammer and Leema K Berland. 2014. Confusing claims for data: A critique of common practices for presenting qualitative research on learning. *Journal of the Learning Sciences* 23, 1 (2014), 37–46.
[29] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.
[30] Alex Hern. 2020. Twitter apologises for'racist'image-cropping algorithm. *The Guardian (Sept. 2020). https://www. theguardian. com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm* (2020).
[31] Matthew K Hong, Udaya Lakshmi, Kimberly Do, Sampath Prahalad, Thomas Olson, Rosa I Arriaga, and Lauren Wilcox. 2020. Using diaries to probe the illness experiences of adolescent patients and parental caregivers. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–16.
[32] Matthew K Hong, Lauren Wilcox, Daniel Machado, Thomas A Olson, and Stephen F Simoneaux. 2016. Care partnerships: Toward technology to support teens' participation in their health care. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 5337–5349.
[33] Elizabeth Huppert, Jason M Cowell, Yawei Cheng, Carlos Contreras-Ibáñez, Natalia Gomez-Sicard, Maria Luz Gonzalez-Gadea, David Huepe, Agustin Ibanez, Kang Lee, Randa Mahasneh, et al. 2019. The development of children's preferences for equality and equity across 13 individualistic and collectivist cultures. *Developmental science* 22, 2 (2019), e12729.
[34] Anna Jobin. 2013. Google's autocompletion: algorithms, stereotypes and accountability. *sociostrategy. com* (2013).
[35] Jillian J Jordan, Katherine McAuliffe, and Felix Warneken. 2014. Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences* 111, 35 (2014), 12710–12715.
[36] Bo Ju, Olivia Ravenscroft, Evelyn Flores, Denise Nacu, Sheena Erete, and Nichole Pinkard. 2020. Understanding Parents' Perceived Barriers to Engaging Their Children in Out-of-School STEM Programs. In *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, Vol. 1. IEEE, 1–4.
[37] Shalini Kantayya. 2020. Coded Bias.
[38] Phyllis A Katz and Jennifer A Kofkin. 1997. Race, gender, and young children. *Developmental psychopathology: Perspectives on adjustment, risk, and disorder* 21 (1997), 51–74.
[39] 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*. 3819–3828.
[40] Daan Kolkman. 2020. F** k the algorithm?: What the world can learn from the UK's A-level grading fiasco. *Impact of Social Sciences Blog* (2020).
[41] Issie Lapowsky. 2018. Google Autocomplete still makes vile suggestions. *Wired. URL: GoogleAutocompleteStillMakesVile-Suggestions* (2018).

[42] Susan Leavy. 2018. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*. 14–16.

[43] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing Middle School Students' AI Literacy. In *Proceedings of the 52nd ACM technical symposium on computer science education*. 191–197.

[44] Steve Lohr. 2018. Facial recognition is accurate, if you're a white guy. In *Ethics of Data and Analytics*. Auerbach Publications, 143–147.

[45] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[46] Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. 2021. Using Explainability to Help Children Understand Gender Bias in AI. In *Interaction Design and Children*. 87–99.

[47] Margaret K Nelson. 2010. *Parenting out of control: Anxious parents in uncertain times*. NYU Press.

[48] Natascha Notten and Peter Nikken. 2016. Boys and girls taking risks online: A gendered perspective on social context and adolescents' risky online behavior. *New Media & Society* 18, 6 (2016), 966–988.

[49] Ziad Obermeyer, Brian W. Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (2019), 447 – 453.

[50] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[51] Kayur Dushyant Patel. 2012. *Lowering the barrier to applying machine learning*. University of Washington.

[52] Sharoda A Paul and Madhu C Reddy. 2010. Understanding together: sensemaking in collaborative information seeking. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 321–330.

[53] Jean Piaget. 1965. The stages of the intellectual development of the child. *Educational psychology in context: Readings for future teachers* 63, 4 (1965), 98–106.

[54] Jennifer Pierre, Roderic Crooks, Morgan Currie, Britt Paris, and Irene Pasquetto. 2021. Getting Ourselves Together: Data-centered participatory design research & epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[55] Kelsey Piper. 2020. Young people have a stake in our future. Let them vote. *Vox* (2020).

[56] Michael T Rizzo, Laura Elenbaas, Shelby Cooley, and Melanie Killen. 2016. Children's recognition of fairness and others' welfare in a resource allocation task: Age related changes. *Developmental psychology* 52, 8 (2016), 1307.

[57] Daniela K Rosner, Saba Kawas, Wenqi Li, Nicole Tilly, and Yi-Chen Sung. 2016. Out of time, out of place: Reflections on design workshops as a research method. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1131–1141.

[58] 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), 4349–4357.

[59] P Scanlon. 2020. Voice assistants don't work for kids: The problem with speech recognition in the classroom. *TechCrunch* (2020).

[60] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 1 – 27.

[61] Marco FH Schmidt, Margarita Svetlova, Jana Johe, and Michael Tomasello. 2016. Children's developing understanding of legitimate reasons for allocating resources unequally. *Cognitive Development* 37 (2016), 42–52.

[62] Deborah Seehorn and Lissa Clayborn. 2017. CSTA K-12 CS standards for all. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. 730–730.

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

[64] Zoe Skinner, Stacey Brown, and Greg Walsh. 2020. Children of Color's Perceptions of Fairness in AI: An Exploration of Equitable and Inclusive Co-Design. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[65] Judith G Smetana, Marc Jambon, Clare Conry-Murray, and Melissa L Sturge-Apple. 2012. Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental psychology* 48, 4 (2012), 1144.

[66] Jaemarie Solyst, Alexis Axon, Angela E.B. Stewart, Motahhare Eslami, and Amy Ogan. 2022. Investigating Girls' Perspectives and Knowledge Gaps on Ethics and Fairness in Artificial Intelligence in a Lightweight Workshop. *International Society of the Learning Sciences (ISLS)* (2022).

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

[68] Jaemarie Solyst, Laura Yao, Alexis Axon, and Amy Ogan. 2022. " It is the Future": Exploring Parent Perspectives of CS Education. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 1*. 258–264.
[69] Jessica A Sommerville. 2018. Infants' understanding of distributive fairness as a test case for identifying the extents and limits of infants' sociomoral cognition and behavior. *Child Development Perspectives* 12, 3 (2018), 141–145.
[70] United States. [n. d.]. Right to Vote at Age 18. *US Constitution, Amendment 26* ([n. d.]).
[71] Lauren Strapagiel. 2020. This researcher's observation shows the uncomfortable bias of TikTok's algorithm. https://www.buzzfeednews.com/article/laurenstrapagiel/tiktok-algorithim-racial-bias
[72] Angie Thomas. 2017. *The hate u give*. Gyldendal A/S.
[73] Greta Thunberg. 2019. *No one is too small to make a difference*. Penguin.
[74] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9795–9799.
[75] Mike Van Duuren, Barbara Dossett, and Dawn Robinson. 1998. Gauging children's understanding of artificially intelligent objects: a presentation of "counterfactuals". *International Journal of Behavioral Development* 22, 4 (1998), 871–889.
[76] Rosalie Waelen and Michał Wieczorek. 2022. The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition. *Philosophy & Technology* 35 (2022), 1–17.
[77] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2022. Informing Age-Appropriate AI: Examining Principles and Practices of AI for Children. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–29.
[78] Jennifer Wang and Sepehr Hejazi Moghadam. 2017. Diversity barriers in K-12 computer science education: Structural and social. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*. 615–620.
[79] David Wendler, Jonathan E Rackoff, Ezekiel J Emanuel, and Christine Grady. 2002. The ethics of paying for children's participation in research. *The Journal of pediatrics* 141, 2 (2002), 166–171.
[80] Randi Williams, Stephen P Kaputsos, and Cynthia Breazeal. 2021. Teacher Perspectives on How To Train Your Robot: A Middle School AI and Ethics Curriculum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15678–15686.
[81] Oren Zuckerman and Mitchel Resnick. 2005. Children's Misconceptions as Barriers to the Learning of Systems Concepts. (2005).

Received January 2023; revised April 2023; accepted May 2023