RAD: A Framework to Support Youth in Critiquing AI

Jaemarie Solyst
jsolyst@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Emily Amspoker
eamspoke@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Ellia Yang
elliay@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Motahhare Eslami
meslami@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jessica Hammer
hammerj@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Amy Ogan
aeo@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Artificial intelligence (AI) is ubiquitous in K-12 youths' everyday lives. However, it has become increasingly well-documented that AI can cause harm by reflecting and amplifying societal biases. While many youth are not currently empowered to engage in broader responsible AI discourse and processes, there is great potential. Foundational to engaging in critical conversations is ability to critique AI. We present the RAD framework, designed to scaffold critique of AI in three steps: Recognize (harms of AI), Analyze (societal aspects of AI harms), and Deliberate (what more responsible AI could be). We ran a workshop study with racially diverse middle school girls (N = 21) to investigate its effectiveness. We found that through being scaffolded with the framework, the youth could articulate biases that they saw in an AI scenario and consider how biases may impact different stakeholders. They then could contemplate how different stakeholders had varying amounts of power in the AI scenario and what that meant in terms of creating more responsible AI systems and processes. After participating in the study, the youth felt more strongly about voicing their opinions about AI with others. The RAD framework and activities work toward emboldening youths' engagement in critical discourse about AI.

CCS CONCEPTS

• **Social and professional topics** → **K-12 education**; **Computer science education**; **Socio-technical systems**.

KEYWORDS

AI literacy, socio-technical, algorithmic justice, K-12, youth

ACM Reference Format:
Jaemarie Solyst, Emily Amspoker, Ellia Yang, Motahhare Eslami, Jessica Hammer, and Amy Ogan. 2025. RAD: A Framework to Support Youth in Critiquing AI. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE TS 2025), February 26-March 1, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3641554.3701966

SIGCSE TS 2025, February 26-March 1, 2025, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0531-1/25/02
https://doi.org/10.1145/3641554.3701966

1 INTRODUCTION

AI is ubiquitous in children's lives, from entertainment to school-related activities. Ninety-five percent of teens have smartphones [1] with AI-embedded apps and services. However, AI has been increasingly well-documented to cause harm to young users. This is especially the case for *marginalized* youth, including Black youth and girls, where AI regularly facilitates exposure to racism and gender injustice [12]. Similarly, AI has been used to make high stakes decisions for youth; for example, the UK deployed an algorithm, which predicted A-level grades for students. However, the AI was biased against lower-income students, negatively impacting them in their trajectory to university. Youth and their families took to the streets and protested the AI, ultimately resulting in the UK stopping use of the algorithm due to their activism [19]. These examples illustrate the necessity of youths' empowerment in the age of AI.

A core skill for youth empowerment is the ability to critique AI. Prior work in culturally responsive computing pedagogy emphasizes the importance of skills that support youths' ability to "advocate for technical changes that could remake the world" [2]. Understood as a socio-technical concept, we consider a critique as a detailed analysis with recommendations for improvement. Yet, critique is a multi-step skill, which needs to be taught and scaffolded.

To address this, we designed the RAD framework, which stands for: *Recognize* AI harms, *Analyze* why harms exist and how they impact stakeholders, and *Deliberate* more responsible AI. We instantiated the framework with activities to apply the steps of RAD. We then ran educational research workshops with racially diverse middle school-aged girls to investigate the effectiveness of the framework, which we present as a case study. This work explores: How does the RAD framework...

- Support youths' ability to critique AI?
- Impact attitudes toward AI?
- Impact beliefs about voicing their opinions on AI?

We find that the framework's instantiated activities successfully scaffolded the girls in developing detailed critiques and contemplating socio-technical aspects of AI. For most learners, engaging in critique of AI led to more awareness of AI's limitations but did not negatively impact optimism for its good uses. After the workshop, the girls felt more emboldened to share their opinions about AI with others, suggesting that this framework both supports youths' learning and empowerment to engage in critical AI discourse.

2 RELATED WORK

2.1 Youths' Potential to Participate

As regular users of AI, youth are stakeholders and exposed to the harms of AI. Despite this, often youth perspectives are not well-included, if at all, in the design and upkeep of AI. Prior work addresses three notions holding back youth (minors under the age of 18) from being viewed as full-fledged stakeholders who can engage with responsible AI, which are (1) youth do not know enough about AI, (2) youth do not have developed enough moral reasoning, and (3) youth need to be protected from engaging with serious topics like AI bias [34]. Contrasting these notions, this work finds that youth as young as 11 years old could often identify and discuss bias that they saw in common examples of AI, including those who did not know much about the technical aspects of AI. They even contemplated adult-debated ethical nuances, such as if AI should reflect the current state of society or a more ideal future. Salac et al. also found that diverse youth bring perspectives grounded in their unique identities and lived experiences when making sense of AI ethics [27]. Further, work suggests that youth are interested in having a say in the design of AI [32], as well as engaging in AI fairness processes, such as user-engaged algorithm auditing [21, 34]. Taken together, prior studies show that youth have important insights and potential to participate in critical discourse to help define the future of responsible AI. Key to engaging in this discourse is the right scaffolding and their preparation to do so.

2.2 Emboldening Socio-technical AI Literacy

AI literacy can help prepare youth to actively engage with the creation and upkeep of algorithms that impact them. Specifically, AI literacy has been defined as not only inclusive of the technical aspects of how AI functions but also the socio-technical (e.g., ethical and societal) aspects of AI [20, 36]. In addition to research focusing primarily on technical aspects about programming AI (e.g., [15, 17]), emerging approaches also scaffold reasoning about these ethical and social aspects. For example, prior work has explored how middle school youth could be supported in thinking about AI stakeholders [39], as well as about biased data in machine learning [30, 31].

Computing education more broadly has had growing emphasis on critical aspects of computing technologies (e.g., [18, 22]). Examples of efforts focusing on K-12 youth from marginalized backgrounds include culturally responsive computing [2, 28], techquity [7, 8], and critically conscious computing [13]. Across these efforts, learners are encouraged to develop *critical consciousness*, which refers to the ability to realize and make sense of power dynamics and injustices in the context of society in order to explore ways of remediation [14]. Critical literacy includes not only analyzing but also suggesting improvements [38]. In the context of this work, we consider critique of AI as a socio-technical analysis with recommendations for a more responsible future with AI. In practice, critiquing AI can be a complex process, which requires scaffolding.

3 RAD FRAMEWORK

Inspired by literature in AI ethics education and critical pedagogy, we created the RAD framework, standing for *Recognize*, *Analyze*, and *Deliberate*, to explicitly support socio-technical critique of AI.

Recognize refers to awareness of potential AI harms. Unless prompted, youth may overtrust AI [33] or even trust AI over humans to make unbiased high-stakes decisions [32]. However, once prompted and shown examples of bias in AI, youth can be capable of recognizing and describing AI bias, often in great detail [34]. *Recognize* is meant to educate and increase sensitivity to AI harms.

Analyze refers to critical thinking about socio-technical aspects of AI harms, specifically contemplating *why* AI harms exist and *how* they impact people. Building on youths' ability to articulate bias once they see it [34], critical analysis of AI takes inspiration from culturally responsive computing [2, 28] and critical consciousness [13, 14] to support youth in deconstructing power dynamics and stakeholders of AI in a broader societal context.

Deliberate refers to consideration of how the AI may be more responsible, which takes into account work on critical literacy suggesting that part of critique is giving recommendations for future improvement [38]. Different from *Analysis*, which is about deconstructing the *current* AI scenario at hand, Deliberation is *looking ahead* to consider what would a more ethical future with AI look like (outcomes), and what is needed to get there (processes)?

4 METHODS

4.1 Participants and Recruitment

We ran an IRB-approved educational workshop study with two groups of racially diverse middle school girls (N = 21) – "AlgoSpark" (N = 4) and "CompuStars" (N = 17), anonymized names. Workshops are a well-documented method in human-computer interaction and help shift power away from researchers toward participants [25] and have been used in prior work to support not only research efforts but also offer youth participants an educational experience [30]. Sessions were roughly 90 minutes long and took place in a mid-sized city in the United States on the East Coast. All learners had heard of AI coming into the study.

In terms of choices for demographics, we worked with middle schoolers, since it is a time when youth begin to use more AI-driven technology and is a prime age for STEM identity development [26]. We worked with girls and girls of color, since women and people of color are underrepresented as AI creators [35], resulting in facing heightened harms from AI (e.g., voice recognition not working as well for feminine voices and accents [4]). This amplifies a need for the ability to critique AI. We worked with two groups centering girls' empowerment in computing, AlgoSpark and CompuStars.

AlgoSpark was an after-school program for girls at a predominantly Black school in a lower income community (qualified for free lunches). This study was the first session of three computing-related sessions in the week-long AlgoSpark program. The Black girls (aged 10 - 12, mean = 11) were recruited in collaboration with

the school's programming staff, who sent out messages to families to partake in AlgoSpark. Participants were compensated an IRB-approved rate of $25 for their participation in the session.

CompuStars was an after-school program for middle school girls, which took place weekly at a research university for the duration of a semester. The program was aimed at exposing learners to technical topics. Our study took place in one of these sessions toward the end of the program. The girls (aged 11 - 14 years old, mean = 12.3) were racially diverse (White = 11, Black = 3, Asian = 2, Latinx Hispanic = 1). Research participation was not compensated, since the program organizers and staff diversity coordinators did not see payment as aligned with their goals.

4.2 Educational Workshop Case Study

The workshop content consisted of first an introduction to AI and fairness, and then going through each step of the RAD framework: (1) Recognition of AI harms, (2) Analysis of AI harms, stakeholders, and power, and (3) Deliberation of fairer AI processes and outcomes.

Introduction to AI and fairness. We first asked what types of AI the learners had in their lives and then discussed common AI examples, including face and voice recognition, search algorithms, and text-to-image generative AI (genAI). We then gave a high level definition of AI. To set the tone for the session's focus on ethics, we asked learners to reflect on what fairness meant to them and discussed a few examples of (un)fairness that had recently come up in their lives. Equality versus equity was emphasized by discussing a common illustration depicting people of differing heights standing on crates to view a baseball game over a fence.

AI scenario. We focused on genAI in the scenario for learners to apply the steps of RAD due to its increasing prominence in children's lives [24], newly documented biases and harms (e.g., [3, 23]), and youths' potential overtrust of genAI [33]. There is a growing need for youth to be sensitive to the limitations of genAI. Recent work also found that text-to-image genAI is particularly useful to scaffold learning about AI bias more broadly [37]. The AI scenario (see Figure 1) was a short story about a girl using a text-to-image genAI 'PictureMake,' which had biased outputs.

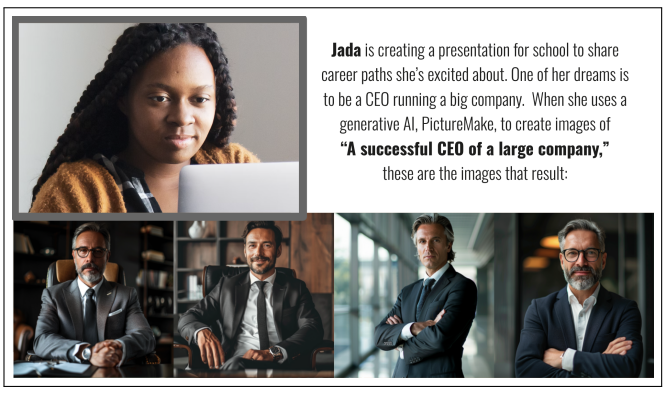

Figure 1: The main AI scenario, showing biased output from a text-to-image generative AI model.[1]

[1]Image of girl from Unsplash.com/@soy_danielthomas

Recognize. Following the AI scenario, learners were prompted to write down on sticky notes if they saw any potential unfairness in the scenario (Figure 1), if they thought it was fair, or if they were uncertain. Learners worked in small groups of 2-5 girls, putting their notes for potential unfairness on a piece of big paper on the tables they sat at together. After their ideas began to slow down, we further scaffolded recognition by defining 'bias' (a preference for someone or something, usually but not always resulting in unfairness) and bias in AI. To illustrate this, we covered examples of two types of representational bias in image-based AI: Stereotyping (certain people represented in a specific way or associated with a specific characteristic) and Erasure (certain people not represented) [29]. For example, stereotyping was demonstrated by sharing genAI outputs for "guy with a bow" (showing archery bows) beside "girl with a bow" (showing hair bows). Erasure was demonstrated with genAI outputs for "a wedding" (showing primarily White American weddings with a man groom and a woman bride, erasing representation of other types of weddings). Learners then were prompted to ideate more about potential unfairness in the scenario.

Analyze. To support socio-technical analysis of AI [14, 28], we defined key concepts: "stakeholder" (a person or group of people that have interest in, helped to create, or impacted by the AI) and "power" (the ability to control, change, or do something). We then asked the learners to consider what stakeholders were involved in the scenario and what power they had in creating the scenario at hand. First, learners guessed stakeholders, and then we gave them a list of five: schools, students, parents, government, and PictureMake developers. To scaffold contemplation of stakeholders' power, we designed a worksheet, where learners mapped out stakeholders in the AI scenario from less powerful to more powerful (Figure 3). In their small groups, learners were prompted to contemplate individually and then discuss stakeholders that had been harmed by the AI bias and why this harm may have come about.

Deliberate. After socio-technical analysis, youth were prompted to deliberate individually and then discuss together: (1) What would a fairer outcome be? (2) To achieve the fairer idea, could stakeholders with more power share their power or collaborate with harmed stakeholders? If so, how? (3) Is there anything that the harmed stakeholders can do on their own? If so, what? And lastly, (4) are there any policies that could help achieve more fairness? If so, what?

Pre and post surveys. To understand how the framework may have imapcted learners' attitudes toward AI and beliefs about voicing their opinions on AI, we administered pre and post surveys. Learners were asked how much they agreed with the following statements on a five-point Likert scale (strongly disagree to strongly agree): "There are many good uses for artificial intelligence," "Artificial intelligence could harm people if it is not used the right way," and "I have important opinions to share about artificial intelligence." The last question on opinions was created in reference to the cognitive autonomy self-evaluation inventory for youth [5]. Due to missing data, we excluded two learners' survey responses for all questions and one learners' survey response for just the opinions question. On the pre-survey, learners also reported their age, race, and prior experience with computing or robotics education. We did not run statistical analyses on the survey responses due to the N-size but rather used this data to understand trends and supplement our focus on qualitative data.

4.3 Data Capture and Analysis

In each session, there was one researcher that led the session and at least one research notetaker for every four learners per session. Notetakers captured observations of the sessions, transcribed what learners said in conversations, and took pictures of learners' artifacts. To analyze the data, we conducted consensus-based [16], qualitative thematic inductive analysis [6] across data sources. The first three authors of the paper conducted close analysis of the data, coming up with subthemes associated with the steps of the RAD framework. Analysis was iteratively checked between the first three authors, and the last three authors continually gave feedback.

A main limitation of this study is that for our participant population, learners opted in, which means that they had some interest in AI or more knowledge about computing prior to the study.

Positionality. Our team was all women, which impacted how we may have been able to relate to our participant population as girls. We come from diverse racial (Asian, Middle Eastern, White, White and Asian) and cultural (American, Iranian, Jewish) backgrounds, which helped us bring unique perspectives around inclusion to this work. We did not have any team members from the community in which AlgoSpark took place, although we have had the great fortune to learn from them in our partnership over the years. Our academic fields included learning science, responsible AI, human-computer interaction, and games, which impacted the perspectives we brought to the analyses and study design.

5 FINDINGS

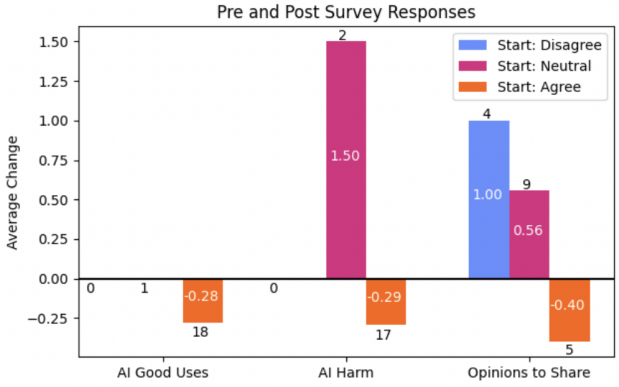

Figure 2: Bars show average changes in pre to post survey scores, given where learners started initially. From the five-point Likert scale, Strongly Disagree and Disagree were combined; Strongly Agree and Agree were combined.

5.1 Enhancing Youth Perspectives on AI

We saw that the girls had some shifts in their pre and post survey responses (Figure 2). They generally continued to believe that AI had good uses after critiquing it; more youth were aware of its harms; and many had an increase in feeling that they had important opinions to contribute about AI. Additionally, we saw that youth engaged in collective sensemaking, by building on and thoughtfully discussing varying perspectives, which led to more developed understanding and critique of AI.

5.1.1 Advancing Youth Knowledge of AI Harms and Benefits. Learners who at first felt neutral about the notion that *"AI could harm people if it isn't used the right way"* finished the workshop having on average a 1.50 increase in their score, suggesting that they left more aware of AI's limitations. However, despite engaging in critique of AI, most learners did not lower their belief that AI had many good uses. In other words, they still had optimism for AI's applications, along with heightened awareness of its limitations. Out of three learners who initially agreed with AI's potential good uses, two decreased their Likert-scale answers by one point, and only one decreased their answer by four points. Out of four learners who initially agreed with the AI's potential to harm, three learners decreased their answers by one point, and one decreased their answer by three points (the same who had decreased their rating drastically for AI's good uses). We believe that trends in Figure 2 show that learners had more well-rounded and informed opinions about the nuances of AI; e.g., one learner explained that they decreased their score on the AI's potential good uses because, *"There are many good uses [for AI]. There are some bad, but there are still many good."*

5.1.2 Boosting Confidence in AI Opinions. We also noted an upward trend in learners' confidence in feeling that they had important opinions to share about AI. Girls who were not as confident in their opinions about AI increased their score in the post survey. This was especially the case for learners who began with disagreeing, increasing their scores on average by 1 whole point. One of these learners wrote in their post survey, *"I have learned new ideas to share with people to help inform them about AI."* Only one participant in the study lowered their Likert scale response on the Opinions to Share question, which decreased by two points.

5.1.3 Collective Sensemaking. We observed the girls engaging in collective sensemaking throughout, building on each others' ideas in each step of RAD. For example, they collectively deliberated together. One AlgoSpark learner suggested that a more fair AI output would have greater racial representation. Another learner added on that developers should *"at least put [a message saying] 'this might not be so right, so be aware.'"* The first learner agreed with this idea, confirming, *"like a warning."* Together, the learners came up with multiple solutions that were more dynamic than one alone. Collective sensemaking also included learners disagreeing with one another in their critical discourse. When discussing the genAI Wedding example of erasure, one learner suggested that the output was fine in terms of representing primarily heterosexual couples because marriage *"should be men and women according to the Bible."* Another learner countered, by sharing how she had *"two uncles. They're married."* This example of discourse shows youth engaging in complex topics related to AI ethics, which can be enhanced by realizing differing perspectives of fellow community members.

In the next sections, we describe results from each aspect of RAD, as well as speculate how scaffolding could be improved.

5.2 Recognition

When prompted in the recognition activity if they recognized any unfairness, most learners were quick to see that there was a lack of representation for the AI scenario (Figure 1). For example, the girls brought up that *"everyone [was] White, and there [was] not even one Black person," "mostly old guys,"* and *"no women."* Learners also considered why this bias was harmful, suggesting that the outputs could limit users by reinforcing false beliefs, such as *"girls can't be CEOs."* A few learners suggested that there was an argument for the bias in the outputs being fair, since the outputs reflected the state of reality, i.e., that most CEOs are indeed older White men.

Overall, we saw that prompting youth to recognize unfairness led to them quickly noticing the possible harms of AI bias. Formally defining bias and giving examples of bias further gave them the vocabulary to talk about it. Stereotypes were a concept that youth were initially more familiar with compared to erasure. We saw that learners could use their new understanding of erasure to generate new observations of groups not being represented by AI output. For instance, one group of learners identified that disability was not represented (i.e., erased) in the image outputs, noting that the images *"erase[d] ... disabilities."* After scaffolding, CompuStars learners generated twenty more sticky notes (totaling 100), and AlgoSpark learners generated 8 more sticky notes (totaling 23) .

Possible improvements. One thing we think would enhance this activity is adding prompts to consider differing harm levels. Some youth recognized bias through pattern recognition (i.e., things that were the same across the genAI outputs like noticing that the CEOs were *"not smiling"*). Yet, not all bias had equal amounts of harm, e.g., similar facial expression across CEOs vs. a lack of racial and gender representation. Drawing their attention to this could further help with recognizing and paying attention to the most prominent harms stemming from AI bias.

5.3 Analysis

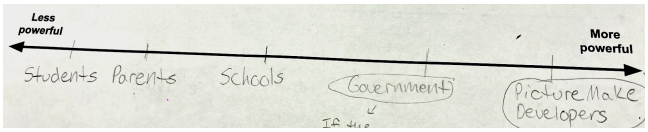

Figure 3: A learner's ranking of stakeholders' power.

Power and stakeholders. When it came to the analysis activities, none of the learners in the study had heard of the term 'stakeholder' prior to us defining it, but 'power' was a concept that was quite familiar. Youth in both groups were quick to define power as *"control,"* in line with the definition we gave them following. After these concepts, in the power worksheet, we saw that most youth ordered the stakeholders in a way that we, the research group, found compelling and might have ordered the stakeholders ourselves, such as in Figure 3. Sixteen out of the 21 learners had similar stakeholder rankings, such that some ordering of Students, Parents, and Schools were on the left most side of the line, and some ordering of Government and PictureMake Developers were on the right most side of the line. This suggests awareness of how stakeholders with governing abilities or direct control over the AI often have power greater with AI. Most learners identified the stakeholders with the least power as being harmed, although many reasoned that harmful AI could negatively impact all stakeholders.

Reasoning about why AI had harmful bias. In considering *why* there was bias, some youth reasoned about representation in data, e.g., that *"there [were] not as many women as CEOs for AI to take examples from."* The CompuStars girls had learned about the concept of data beforehand during the program, so we observed them apply this general understanding of how AI was trained on data. Some learners were also able to perceive how AI could reflect those who have a say in its creation. One learner reasoned that the outputs would not show only people with light skin if *"people from all over the world"* were involved in making the AI—there would be more *"people who looked different because they'd have a say."*

Possible improvements. The power worksheet activity was designed to be simple for the limited time of the workshops and to spur critical thought and conversation about stakeholders' power. Future iterations exploring these concepts could allow for more complex exploration of stakeholders and power dynamics (e.g., beyond a 2-dimensional representation). We also recommend addressing a misconception that we saw come up with a few learners, which was that the *"technology"* or *"the AI"* itself was a stakeholder.

5.4 Deliberation

Defining fairness. When the youth deliberated what a fairer outcome would be, they were quick to bring up a need for diversity, aligned with prior work [32, 34]. They called for *"more race in the technology, more women in the pictures."* One suggested that the developers should fix the AI by having *"more diverse"* outputs.

Stakeholders sharing power. Learners were thoughtful when they considered ways of sharing and reclaiming power in reaching fairer AI outcomes. For sharing power, often learners suggested that harmed stakeholders could voice their perspectives to more powerful stakeholders through feedback processes (e.g., users *"leave comments for the developers to improve"* or use a *"suggestion box for improvements"*) and co-design (e.g., *"get some of the developers to talk to schools"* and *"company developers meet with students"*). Although, there was some cynicism about stakeholders with more power—one learner asked: *"Does the government pay attention to your issues, or do they just want your money?"*

Harmed stakeholders' self-empowerment. The girls also ideated how harmed stakeholders could self-empower or resist AI harms on both individual and community-based scales. Individually, youth suggested that negatively impacted stakeholders could reject AI by *"not using it,"* banning its use locally (e.g., in schools), or break it by *"hack[ing] it."* One learner suggested that harmed users could leverage the fact that AI behavior was *"not a stagnant thing."* She continued to reason that through harmed stakeholders' intentional use of the AI to provide systems with new data and *"different ideas, ... [it] would be trained differently, as AI is being continuously trained, hence it can be [improved]."* Another learner brought up that some stakeholders may be capable of creating suitable alternative systems, such that *"negatively impacted stakeholders could take their experience and ... make [an AI] better ... on their own."*

Ways of self-empowerment that did not include hands-on work with AI included spreading awareness (e.g., *"make a TikTok video"*

about the negative impacts of the technology) and resistance on a larger scale. The learners suggested a number of ways that communities could *"organize themselves"* in order to *"threaten the company,"* e.g., through *"a bunch of different strikes."* A learner elaborated that *"if a bunch of people get together, people high up have to do something because they're losing money,"* relating this notion to news they had heard about Amazon warehouse workers striking.

Policy ideation. The girls considered a variety of responsible AI policies, related to fairer outcomes and processes. For outcomes, a common policy was to mandate diversity in the output of AI models. Learners clarified that this diversity must be related to identity, with one noting that diverse text-to-image outputs *"can't [just] be if someone is wearing a watch or not, has to be gender or race or something."* Another learner recommended that certain AI shouldn't be used at all until it meets specific standards, mentioning that some AI *"could be helpful eventually, but until there is a way that everyone is safe and happy, then [they] shouldn't be used."*

In terms of processes to achieve more fair outcomes, some learners thought that training data should be regulated, noting that companies should be required to *"try not to use similar images so that the AI can develop more diversity."* Others suggested that the government should oversee how companies solicit and use feedback. For example, one learner suggested that companies should be required to *"have a variety of people involved and check with the people who use the website once a month and ask for suggestions, then take the ones that are most beneficial."* Learners also noted that the government could protect harmed stakeholders by requiring companies to be accountable and *"pay the fees"* of negative AI impacts. Learners sometimes considered policy beyond AI itself, explaining that policies should first focus on enhancing fairness in societies, such as by increasing diversity, e.g., *"employ[ing] more girls in the FBI, hav[ing] more girls do archery to increase representation."*

Possible improvements. Upon reflection, we saw opportunity to explicitly scaffold thinking about under what circumstances AI should be used, if at all—in other words, support youths' consideration of ways to evaluate the use of AI in different contexts. Additionally, when prompting learners to consider how stakeholders could share power, often they suggested that stakeholders with more power, specifically developers, could pass their power 'down' to harmed stakeholders. This was likely because of the prompt we used. However, what could be more effective is prompting consideration of *co-liberation* [10], such that stakeholders benefit from each others' empowerment. We did not want to instill a belief that companies will be accountable and empower users, due to societal and economic forces (like capitalism) at play.

6 DISCUSSION

Overall, we saw that going through the RAD framework steps supported the learners in engaging in critical understanding and discourse about AI harms. After the workshop, the girls were more confident that they had important opinions to share about AI. Youth having cognitive autonomy in voicing their opinions [5] supports their empowerment to engage in critical discourse as stakeholders and potential agents of change in both local ways (e.g., discussing and informing their communities) and broader ways (e.g., helping to create policies). Contemplation of AI ethics was bolstered by learners' collective sensemaking with one another, as they built on each others' ideas with their own perspectives.

While previous work has introduced the concept of stakeholders in AI education [11], *power* was not explicitly emphasized with stakeholders in prior work scaffolding socio-technical AI literacy, despite it being a key concept in critical understanding of AI [2, 14, 28]. The framework activities provide opportunity to engage with the interaction of the two concepts of stakeholders and power.

For example, youth ideated processes for stakeholders to share power through co-design. They also were able to ideate numerous complex ways that harmed stakeholders without direct control over the algorithms or governing power could build and leverage their power, both individually and collectively. These approaches build on previous work that taxonomizes ways in which adults have self-empowered in the face of algorithmic injustices [9]. For example, youth suggested that people could discontinue use or even break the AI as a way to empower, aligned with this work's definition of 'refusing legitimate engagement with harmful algorithmic systems,' or that they might spread awareness using TikTok and organizing with others to resist AI, aligned with this work's definition of 'communicating algorithmic harm with others' [9]. Youths' ideas were realistic and mirrored responses that adults and communities have taken against AI harms, suggesting that youth can be a part of strategizing individual and community-led efforts toward empowerment with AI.

There is opportunity for work to explore supporting youths' engagement with AI policy. This study suggests that young people, even as young as 10 years old, could help ideate thoughtful AI policies. Building on [13], we saw learners recognize that harms from AI often stem from societal inequities, indicating that achieving fairness involves both AI and broader societal justice. With proper guidance, youth can effectively engage in AI policy discourse.

Leveraging the RAD framework can enable youth to engage in critical conversations about AI design, upkeep, and policy in the real world. Some activities may be adapted to support other AI systems for youth to weigh in on. We offer the framework to guide the overall steps of critiquing AI and the instantiated activities as a concrete example to support application. While our study was intentionally conducted with girls (due to their underrepresentation as AI creators but heightened harms), we believe that this framework could be effective with youth of varying identities in middle school and older. Critiquing AI is a precursor to further activism, like (re)designing and building AI systems more responsibly.

Ultimately, this work fosters diverse youths' empowerment in the age of AI by supporting them in recognizing potential harms of AI, analyzing the dynamics of power among stakeholders, and deliberating a more fair future of AI. This empowerment could support youth in making informed decisions in their interactions with AI, or engaging in larger scale efforts, such as becoming community activists, poised to advocate for algorithmic justice.

ACKNOWLEDGMENTS

We thank the child participants and community partners we worked with to make this research possible. We additionally thank Yi Luo for her help with data collection. This work was supported by the Jacobs Foundation CERES Network and the NSF DRL-1811086.

REFERENCES

[1] Monica Anderson. 2022. Teens, Social Media and Technology 2018. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

[2] 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.

[3] Alexander Baines, Lidia Gruia, Gail Collyer-Hoar, and Elisa Rubegni. 2024. Playgrounds and Prejudices: Exploring Biases in Generative AI For Children.. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 839–843.

[4] Joan Palmiter Bajorek. 2019. Voice recognition still has significant race and gender biases. *Harvard Business Review* 10 (2019), 1–4.

[5] Troy E Beckert. 2007. Cognitive autonomy and self-evaluation in adolescence: A conceptual investigation and instrument development. *North American Journal of Psychology* 9, 3 (2007).

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

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

[8] Merijke Coenraad and David Weintrop. 2024. Talking Techquity: Teaching the Equity and Social Justice Impacts of Computing in Middle School Classrooms. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. 227–233.

[9] Alicia DeVrio, Motahhare Eslami, and Kenneth Holstein. 2024. Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1093–1106.

[10] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism.* MIT press.

[11] 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.

[12] Avriel Epps-Darling. 2020. How the racism baked into technology hurts teens. *The Atlantic* 24 (2020).

[13] Jayne Everson, F Megumi Kivuva, and Amy J Ko. 2022. " A Key to Reducing Inequities in Like, AI, is by Reducing Inequities Everywhere First" Emerging Critical Consciousness in a Co-Constructed Secondary CS Classroom. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education-Volume 1*. 209–215.

[14] Paulo Freire. 2020. Pedagogy of the oppressed. In *Toward a sociology of education*. Routledge, 374–386.

[15] D Ganesh, M Sunil Kumar, P Venkateswarlu Reddy, S Kavitha, and D Sudarsana Murthy. 2022. Implementation of AI Pop Bots and its allied Applications for Designing Efficient Curriculum in Early Childhood Education. *International Journal of Early Childhood Special Education* 14, 3 (2022).

[16] 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.

[17] Ken Kahn, Rani Megasari, Erna Piantari, and Enjun Junaeti. 2018. AI programming by children using snap! Block programming in a developing country. In *Thirteenth European conference on technology enhanced learning*, Vol. 11082. Springer.

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

[19] 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).

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

[21] Luis Morales-Navarro, Yasmin Kafai, Vedya Konda, and Danaë Metaxa. 2024. Youth as Peer Auditors: Engaging Teenagers with Algorithm Auditing of Machine Learning Applications. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 560–573.

[22] Luis Morales-Navarro and Yasmin B Kafai. 2023. Conceptualizing approaches to critical computing education: Inquiry, design, and reimagination. In *Past, Present and Future of Computing Education Research: A Global Perspective*. Springer, 521–538.

[23] Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. 2023. AI's regimes of representation: A community-centered study of text-to-image models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 506–517.

[24] Mitchel Resnick. 2024. Generative AI and creative learning: Concerns, opportunities, and choices. (2024).

[25] 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.

[26] Philip M Sadler, Gerhard Sonnert, Zahra Hazari, and Robert Tai. 2012. Stability and volatility of STEM career interest in high school: A gender study. *Science education* 96, 3 (2012), 411–427.

[27] Jean Salac, Alannah Oleson, Lena Armstrong, Audrey Le Meur, and Amy J Ko. 2023. Funds of Knowledge used by Adolescents of Color in Scaffolded Sensemaking around Algorithmic Fairness. In *Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1*. 191–205.

[28] Kimberly A Scott, Kimberly M Sheridan, and Kevin Clark. 2015. Culturally responsive computing: A theory revisited. *Learning, media and technology* 40, 4 (2015), 412–436.

[29] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[30] Jaemarie Solyst, Alexis Axon, Angela EB Stewart, Motahhare Eslami, and Amy Ogan. 2023. Investigating girls' perspectives and knowledge gaps on ethics and fairness in Artificial Intelligence in a Lightweight workshop. *arXiv preprint arXiv:2302.13947* (2023).

[31] Jaemarie Solyst, Jennifer Kim, Amy Ogan, and Jessica Hammer. 2022. Data Detectives: A Tabletop Card Game about Training Data. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 2*. 632–632.

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

[33] Jaemarie Solyst, Ellia Yang, Shixian Xie, Jessica Hammer, Amy Ogan, and Motahhare Eslami. 2024. Children's Overtrust and Shifting Perspectives of Generative AI. *arXiv preprint arXiv:2404.14511* (2024).

[34] Jaemarie Solyst, Ellia Yang, Shixian Xie, Amy Ogan, Jessica Hammer, and Motahhare Eslami. 2023. The Potential of Diverse Youth as Stakeholders in Identifying and Mitigating Algorithmic Bias for a Future of Fairer AI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–27.

[35] Konstantinos Stathoulopoulos and Juan C Mateos-Garcia. 2019. Gender diversity in AI research. *Available at SSRN 3428240* (2019).

[36] 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.

[37] Henriikka Vartiainen, Juho Kahila, Matti Tedre, Sonsoles López-Pernas, and Nicolas Pope. 2024. Enhancing children's understanding of algorithmic biases in and with text-to-image generative AI. *new media & society* (2024), 14614448241252820.

[38] Vivian Maria Vasquez, Hilary Janks, and Barbara Comber. 2019. Critical literacy as a way of being and doing. *Language arts* 96, 5 (2019), 300–311.

[39] Helen Zhang, Irene Lee, Safinah Ali, Daniella DiPaola, Yihong Cheng, and Cynthia Breazeal. 2023. Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education* 33, 2 (2023), 290–324.