Developing the Five Steps of Algorithm Auditing for K-12 Artificial Intelligence Education

PURM Researcher: Evelyn Yu, SEAS 2027 **Mentors: Dr. Yasmin Kafai (GSE), Dr. Lauren Vogelstein, Luis Morales-Navarro, & Dr. Danaë Metaxa**

Introduction

We adapted the 5 steps of algorithm auditing for youth to collaboratively and systematically investigate the biases in AI tools.

Methods & Data

Thematic analysis of youth's ideas while completing the five steps of auditing in relation to AI filter tools.

Context

Summer workshop conducted in 2024 with 12 consenting youth (ages 14-15). Youth designed AI-powered TikTok filters.

Conclusion

Students adapt the steps to serve their goals and bring their personal experiences to the table. The execution of the five steps of auditing are nonlinear, with many steps overlapping and feeding into one another. This leads us to consider a sixth step to the auditing steps: reflection.

Step 1: Hypothesize

Hypothesis Generation Follows Auditing Itself

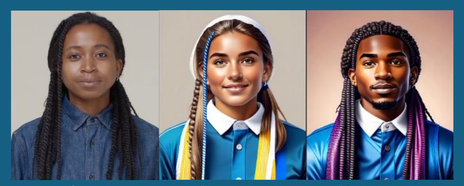
Students formed hypotheses after they had already audited specific filters they had created. Two students hypothesized that the software used had racial biases after testing "tennis player" and "basketball player" on the same input.

> *"When I put tennis player, it made … it made her white, and, like, she stayed a girl, but I did [basketball player], and it made her black, you know."*

> *"Alright, do tennis player [again]? Okay, that's like full racism."*

Another student hypothesized that the AI's training dataset lacked photos of rotated faces after trying to adjust her prompt to create a filter that accounted for these possibilities and failed.

> *"Well, I don't know how we're supposed to fix that… it's just the AI's fault I think it's just the way that the AI was trained."*



Original "Tennis player" "Basketball player"

Step 2: Design Inputs

Students Draw on Personal Biases and Stereotype Conceptions to Design Inputs

Students designed their input dataset on stereotypes that they held, informed by their experiences.

One student identified "chef" and "cooker" as possible inputs because she felt that "chef" is male-coded and "cook" is female-coded, despite both jobs performing similar tasks.

One student mentioned that he thinks of "Black females" when he thinks of "nurses" because his mom and his friends' moms are nurses.

The students associated teachers with white women, corner store workers with "Mexican women", rappers with "Black males" and tech support and scammers (which they grouped together) with "Indians".

Student 1: *"Chef--I think like an Italian man."*
Student 2: *"Mmm, no I think that's just racism."*
Student 1: *"That's the stereotype."*

The students are investigating if they are racist, or just regurgitating a generally understood racist stereotype, much like the AI.

Step 3: Test Inputs

Students Perform Proto-Analysis While Testing

As the students tested inputs, they formed perceptions of what the AI's biases are.

Inputs were tested with faces one at a time (vertically), but students scrolled through individual prompts across different inputs (horizontally) to look for patterns.

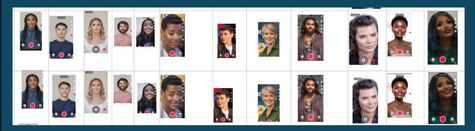
After looking through the tattoo artist prompts:

> Student 1: *"A lot of people do got tattoos… I wonder why they gave [Lupita Nyong'o] a neck tattoo." (Other inputs did not result in neck tattoos specifically.)*

> Student 2: *"That's to represent the shackles of slavery, cuz she's Black."*

Students hypothesizing that certain inputs can amplify the bias and attributed a kind of meaning to the output.

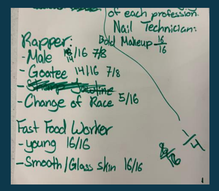
Nail Technician Prompt:



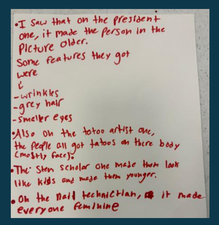
Step 4: Analyze

Qualitative and Quantitative Techniques are Used To Code and Analyze Data

Students investigated what parts of their perception might be inaccurate and looked for more empirical ways to analyze. They saw wrinkles and grey hair as a sign of age, smooth skin as a sign of youth, facial hair as a masculine signal, and makeup as a feminine signal:

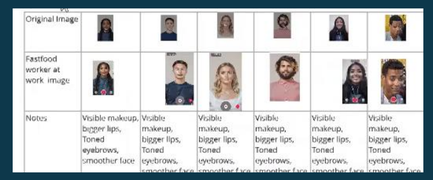


> *"So, let's look at all the judges and see how many look older—by older we're talking about wrinkles and stuff. And grey hair."*



Many students made quantitative observations

While others made qualitative observations:



Step 5: Report

Different Audiences, Different Reporting Styles

Students chose audiences for their audit reports: the creators and users of Effect House, TikTok filter users, and tailored their reporting styles/mediums. Students made slides, graphs, posters. Most made TikToks for their report, connecting with the medium they were auditing.

After being asked who should know about biases in the AI:

> *"Well, I thought, … the people behind Effect House… I feel like it's more effective or useful if you told people so they can try to fix it."*

> *"Users [should know] it can be biased, like, no matter what."*

> *"People … use Effect House to make filters… should know that…AI is going to lean to one specific group when it comes to certain jobs and occupations."*

> *"Teenagers, people who use TikTok… AI is everywhere and kids who are still learning and don't understand could be affected by it."*