

# Reframe & Define: Auditing AI Generated Algorithmic Bias



Assumptions, Reversals and New Interventions

Team Nasa GenAI Vanguard

01

# Synthesize by Walking the Wall



Alec Chen

### Module 2 - Background Research

AI bias stems from a wide variety of sources:

- Flawed training data
- Errors in the algorithm
- Existing cognitive biases from developers

Alec Chen

everyone:

- Casual users
- Large corporations
- Specialized technological fields that rely on cutting edge AI

Alec Chen

It is possible to mitigate bias within AI:

- Controlling the sample data used to train algorithms
- Educating designers on cognitive biases

Alec Chen

Is there anything we're missing? Are there other types of AI bias that we're overlooking in favor of modeling our product around the same three categories?

Alec Chen

Although the consequences of bias are severe, it seems that currently not enough developers have deemed bias as a significant field of interest.

Alec Chen

Common biases pertain to race, gender, and sexuality

Alec Chen

What kinds of users specifically should we be considering when trying to tackle this issue?

Alec Chen

### Module 3 - Findings from Data

There does not appear to be any visible correlation between location and sensitivity on bias

Alec Chen

People appear to be much more sensitive to racial and gender biases than they are towards bias on sexual orientation

Alec Chen

Being strongly sensitive on one bias does not indicate the same on the others

Alec Chen

Is this subjective user information even that useful?

Alec Chen

What varieties of bias, if any, might we be missing out on?

Alec Chen

### Module 4 - Heuristic Evaluation

Navigating the TAIGA website is clunky and unintuitive, especially on the landing page

Alec Chen

The website does not clearly communicate intent: user submissions are framed as a form to be professionally reviewed, but are posted on a public forum

Alec Chen

There are a handful of small visual inconsistencies within the website's visual aesthetic, which detract from focus

Alec Chen

The submission form requires too much expertise from a casual audience/user base

Alec Chen

How might we be able to communicate the purpose of the TAIGA website better?

Alec Chen

How might we make the TAIGA website more accessible to our ideal users?

Alec Chen

### Possible Design Ideas

Communicate better that TAIGA is a website designed for public interaction/social media

Alec Chen

To promote engagement and filter out weak submissions, have a point system (upvotes/downvotes) and only allow commentary on the top posts (most engagement)

Alec Chen

Target users who specifically are affected by possible biases in generative AI, as they will be more likely to share relevant findings on TAIGA

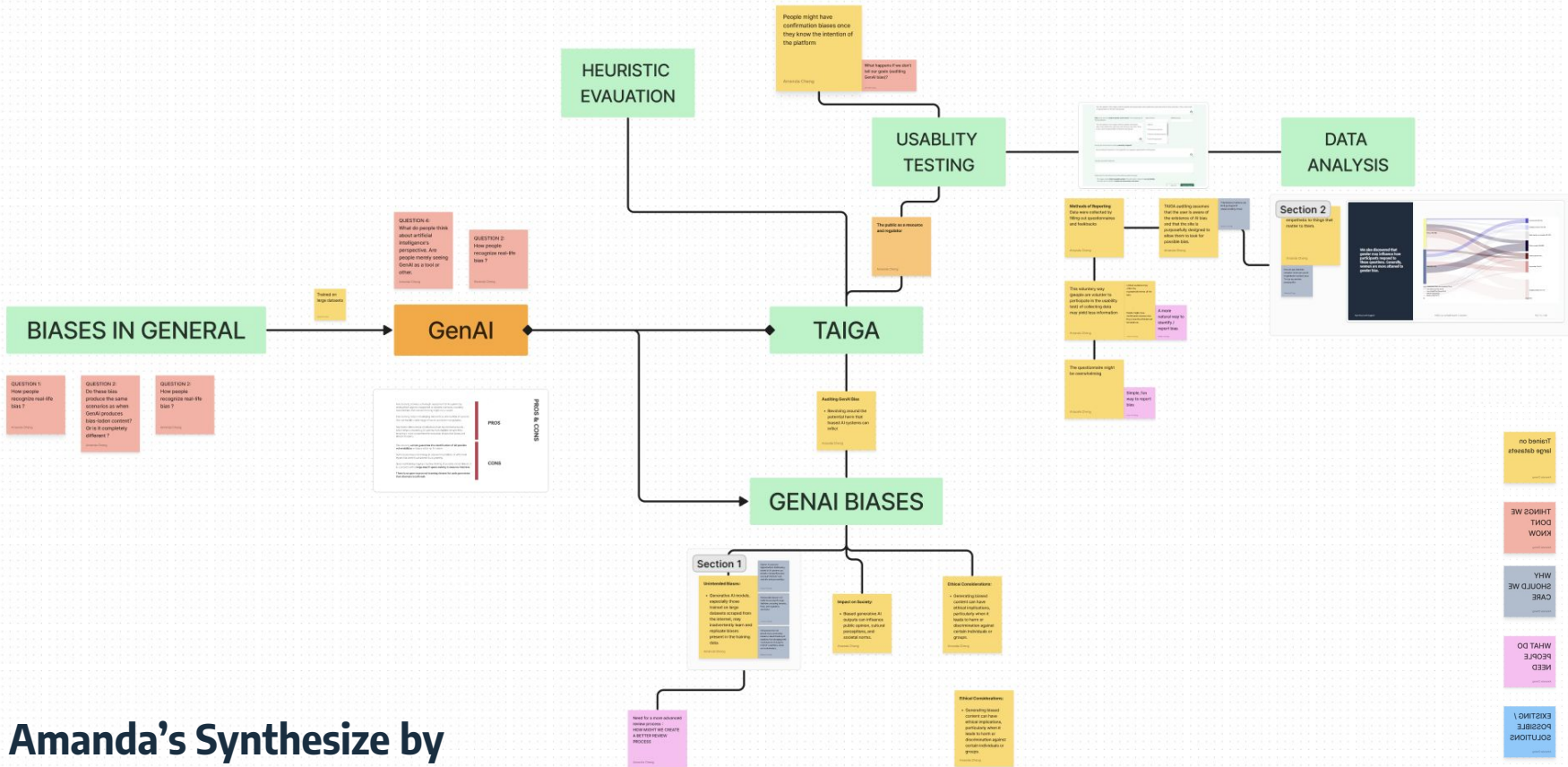
Alec Chen

TO-DO: figure out a way to identify other significant types of biases that aren't currently well-known in TAIGA or today's society

Alec Chen

# Alec's Synthesize by Walking the Wall

# Amanda's Synthesize by Walking the Wall



no things to think we know why care what do need existing

# BIASES IN GENERAL

- QUESTION 1: How do people recognize their life bias?
- QUESTION 2: Do these bias produce the same outcomes as when GenAI produces a given input content? Or is it complementary?
- QUESTION 3: How do people recognize their life bias?

Related to life content

# GenAI



- QUESTION 4: What do people think about artificial intelligence's perspective and people mainly using GenAI as a tool or other.
- QUESTION 5: How people recognize their life bias?

# HEURISTIC EVALUATION

# TAIGA

Adding Search Bias

# GENAI BIASES

**Section 1**

**Methodological Biases:**

- Interviewing methods, especially those based on open-ended questions, are susceptible to social desirability bias and other biases. Additionally, many bias reduction techniques are not applicable to the training data.

**Need for a more advanced methodology that leverages AI to reduce human biases**

Impact on Society

Diverse Considerations

Diverse Considerations

# USABILITY TESTING

People might have confirmation biases since they know the objectives of the platform

What happens if we don't get our users thinking about need?

The platform is a complex management



Methods of Reporting

Methods of Reporting

The solution only

A user interface used to report issues

The questionnaire might be overabundant

Simple, less may be better

# DATA ANALYSIS

Section 2



no trials against

THINGS WE DON'T KNOW

WHY CARE ABOUT IT

WHAT DO PEOPLE NEED

EXISTING POSSIBLE SOLUTIONS



# Sofia's Synthesize by Walking the Wall

## Background Research

### Background Research: Digital Experiential

**Algorithm Biases in Social Media**  
Algorithms are biased by how long users stay on a page, what content they like, what their heritage, and by demographics of the user base. Geographic location, usage of phone, etc.

**Spanish People Watching Futbol**  
Non-Spanish people tend to believe that all Spanish love soccer & to order misadventures, stereotypes, and their ideas and actions based on said nationality. This is an example of algorithm bias.

### Background Research: Digital Informational Search

**Informal: Reddit**  
Subreddits that host the opinions of certain people which can cause extreme bias but do not contain quality advice although the information may be accurate or true.

**Academic: Google Scholar**  
Google Scholar is great resource for less biased information, since it usually has academic-style papers. However, some authors may have biases that they do not express in their papers.

### Background Research: Insights

People are influenced by online information without checking its validity.

Algorithms analyze users' preferences, which may lead to biased content.

News sources can be a good source of news, although they may be biased.

Informal sources of search, such as reddit, can be a good resource for finding people's opinions on topics, but tend to be highly biased.

Scientific journals are a great way to find new unbiased information on specific topics.

## Findings From Data

There is a strong correlation for users in the age group 25-34 encountering them a few times a year. Similarly, this trend is seen for the age group 65 and above who encounter it monthly. All age groups encounter such news a couple of times yearly. However, it is a rare occurrence daily.

sofia

People appear to be more sensitive to racial and gender biases than they are towards bias on sexual orientation

sofia

Usage of algorithmic systems and the encounter frequency of algorithmic biases seems weakly correlated despite all age groups being familiar with them

sofia

### Gender-Bias Sensitivity

### Example: Opinions on "Totally Unharmful"

### Hypotheses and Questions

1. Is there a correlation between geographic location and sensitivity to algorithmic bias?
2. How do demographic factors of users influence perceptions of bias and discrimination in algorithmic systems?
3. What themes emerge from textual responses about why individuals find certain algorithmic outputs harmful or unhelpful?

sofia

## Heuristic Evaluation

Visibility of System Status

Provides clear feedback on the audit process, displaying progress indicators and status updates throughout the workflow. Users can track the status of audits and understand where they are in the process.

Employs a clean and minimalist design approach, focusing on essential elements and decluttering the interface. The platform utilizes whitespace effectively to improve visual hierarchy and readability, making it easier for users to navigate and focus on critical audit information. However, there may be areas where the design could be further optimized to reduce visual noise and enhance user engagement. For instance, refining the layout of dashboard widgets or employing subtle animations to guide users through complex workflows could contribute to a more polished and visually appealing user experience.

sofia

## Aesthetic: Pros and Cons

Very minimalistic design that is easy to use. There are several icons that are displayed for user feedback on the generated responses by hovering over the generated text or images. For example, there is a thumbs up and down icon for users to distinguish and label good and bad responses. There is also a copy and thumbs down for the text generated. Not sure why there is not a thumbs up button for that and it isn't immediately apparent why there is a replication of this mechanism.

sofia

# 02

## Assumptions and Reverse

Link to Miro Board for all the assumptions

<https://miro.com/app/board/uXjVNh7fuQA=>

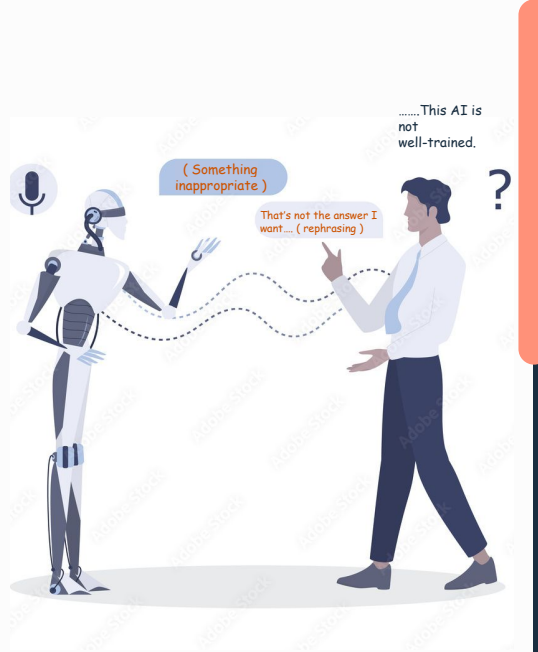


# Why Reverse Assumption ?

- Our team opted for the reverse assumption method to tackle the reframing activity as it allowed us to challenge and step beyond the boundaries of our established background context through various assignments already performed in this course.
- We realized that we all have different backgrounds in dealing with generative AI systems (Psychology, Computer Science, Architecture, Information Systems). This could lead to potential inherent biases, and preconceived notions that could constrain our creativity.
- We made a list of these assumptions using sticky notes in a separate section.
- We then inverted each assumption to generate a list of ideas we can expand on.
- While we generated several routes for solutions to the project, some are highly infeasible.
- However, by intentionally reversing even basic assumptions, we were able to explore new possibilities that were previously unknown or obscured by reality or unexamined assumptions.

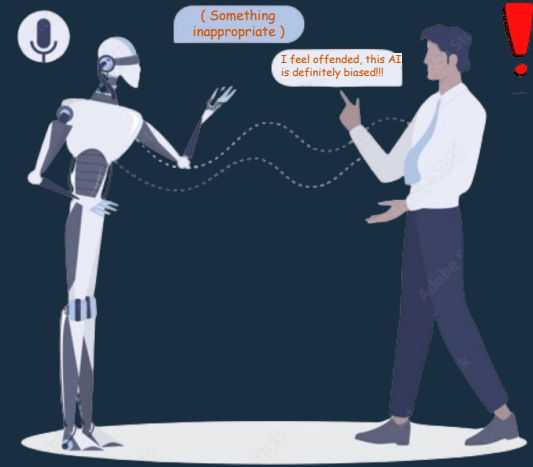
## One Assumption we Explored

Individuals' views on bias originating from human sources differ markedly from their perspectives on bias produced by AI.



## Discussed Reversals

Individuals' perspectives on bias produced by AI are the same as bias originating from human sources.



## One Assumption we Explored

It would be more efficient to provide a dedicated platform for people to audit GenAI's bias

### WeAudit TAIGA | Tool for Auditing Images Generated by AI

Find patterns and detect biases in AI generated images. [Learn more about AI bias here.](#)

Compare

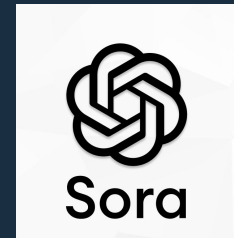
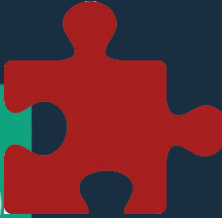
Insert prompt here.

Insert prompt here.

Show Examples

## Discussed Reversals

It would be more efficient to provide a plugin that compatible for different platforms to review GenAI deviations.



## One Assumption we Explored

A better way to recognize GenAI bias is through single user auditing feedback.

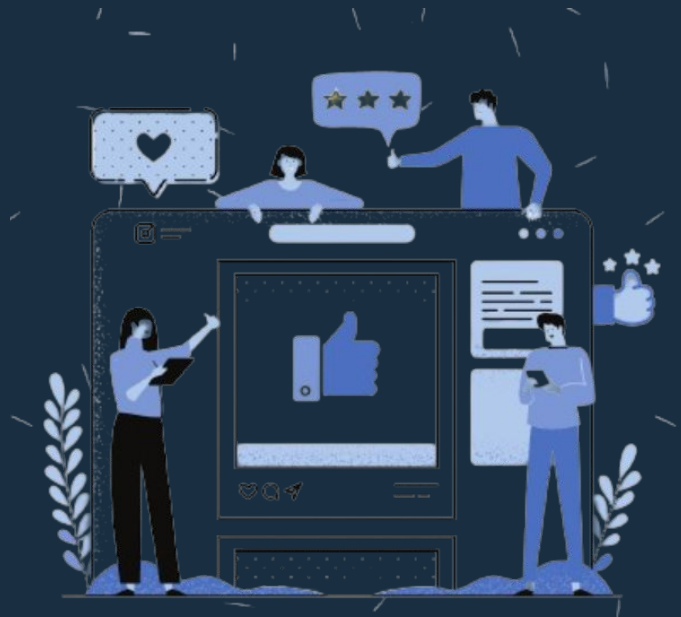


Link to Miro Board for all the assumptions explored:

<https://miro.com/app/board/uXjVNH7fuQA=>

## Discussed Reversals

A better way to recognize GenAI bias is through a collaborative platform.





# Takeaways

- We documented every stage of the process directly on the miro board.
- A comprehensive list of takeaways is also written there.
- Our primary takeaways included:
  - Taking advantage of spike in user activity on generative AI systems to draw specific and large-scale user audited feedback - this spike is usually observed when lots of people red-team generative AI platforms when a particularly malicious behavior is trending on social media.
  - We also discussed setting up a mechanism that incentivized users to report biases by counting the number of up and down votes on their report.

The Miro board has more details:

<https://miro.com/app/board/uXjVNH7fuQA=/>



The Miro board has more details:

<https://miro.com/app/board/uXjvNh7fuQA=>

# Defining the Project #1

- **How Might We Statement:** How might we develop a user-friendly system that enables and motivates everyday users to identify and report emergent biases in generative AI systems?
- **Supported Activities/Tasks:** The project will support the development of educational materials to increase bias awareness among users, the creation of intuitive reporting mechanisms, and the facilitation of user discussions and feedback on bias incidents and a post-mortem red-teaming facilitated by taking advantage user activity spikes during trending social media posts or news articles targeting malicious behavior.
- **Impacted Roles and People:** This project will affect everyday users of generative AI systems, developers, and researchers who are involved in AI bias mitigation, as well as community moderators and educators.
- **Context:** The project will operate in digital spaces where AI systems are interacted with, such as web platforms and apps, and in social spaces that could include online forums or user groups.
- **Tools/Platforms:** We plan to utilize and possibly extend the capabilities of TAIGA, ChatGPT and Co-pilot, incorporating interactive educational modules, feedback tools, and community discussion features to empower users to detect and report biases as they interact with generative AI systems.

The Miro board has more details:

<https://miro.com/app/board/uXjVNH7fuQA=>

# Defining the Project #2

- **How Might We Statement:**
  - How can we evoke strong emotions in users about bias so that they are emotionally more attached to the issue and therefore more motivated to report bias?
  - How do we make people feel that GenAI bias is just as important as real-live bias?
  - How do we stop users perception of thinking that AI-generated content is objectified and disconnected from reality?
- **Context:**
  - By blurring the boundaries between human bias and GenAI bias at the user end, AI-generated content is presented as if it were a "human" point of view to generate discussion.
- **Tools/Platforms:**
  - Social media platforms like Instagram, facebook, twitter

The Miro board has more details:

<https://miro.com/app/board/uXjVNh7fuQA=>

# Contributions

- Nivedhitha Dhanasekaran
  - Individual Wall-Walk (Slide 3)
  - Miro Board (Slide 6)
  - Why Reverse Assumption? (Slide 7)
  - Takeaways (Slide 12)
  - Redefining the Project (Slide 13)
- Amanda Cheng
  - Template (All slides)
  - Individual Wall-Walk (Slide 5)
  - Miro Board (Slide 6)
  - Specific Cases (Slides 8-10)
- Alec Chen
  - Individual Wall-Walk (Slide 4)
- Sofia
  - Individual Wall-Walk