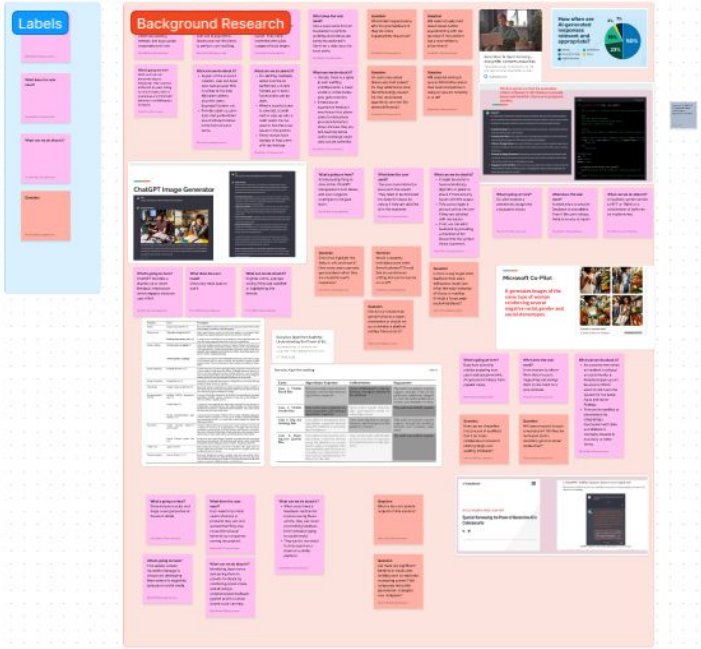
Reframe & Define:
Auditing AI Generated Algorithmic Bias

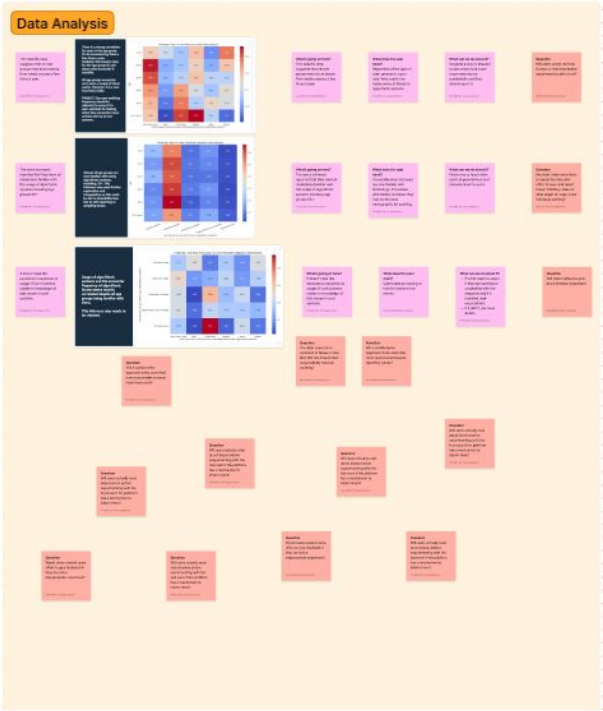
Assumptions, Reversals and New Interventions

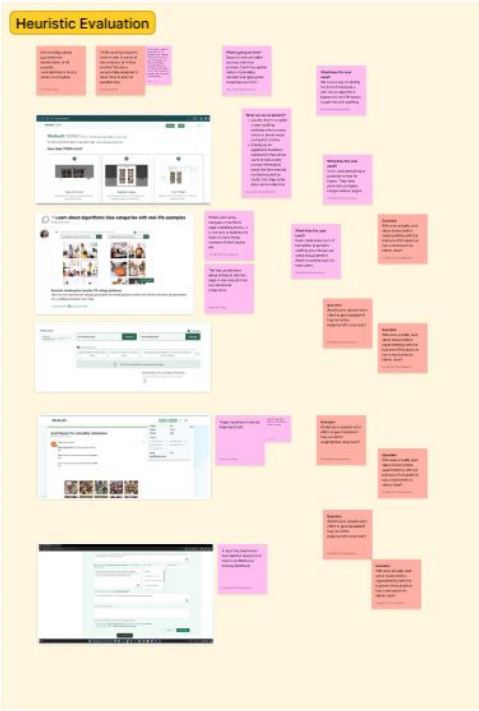
Team Nasa GenAI Vanguard

01

Synthesize by Walking the Wall

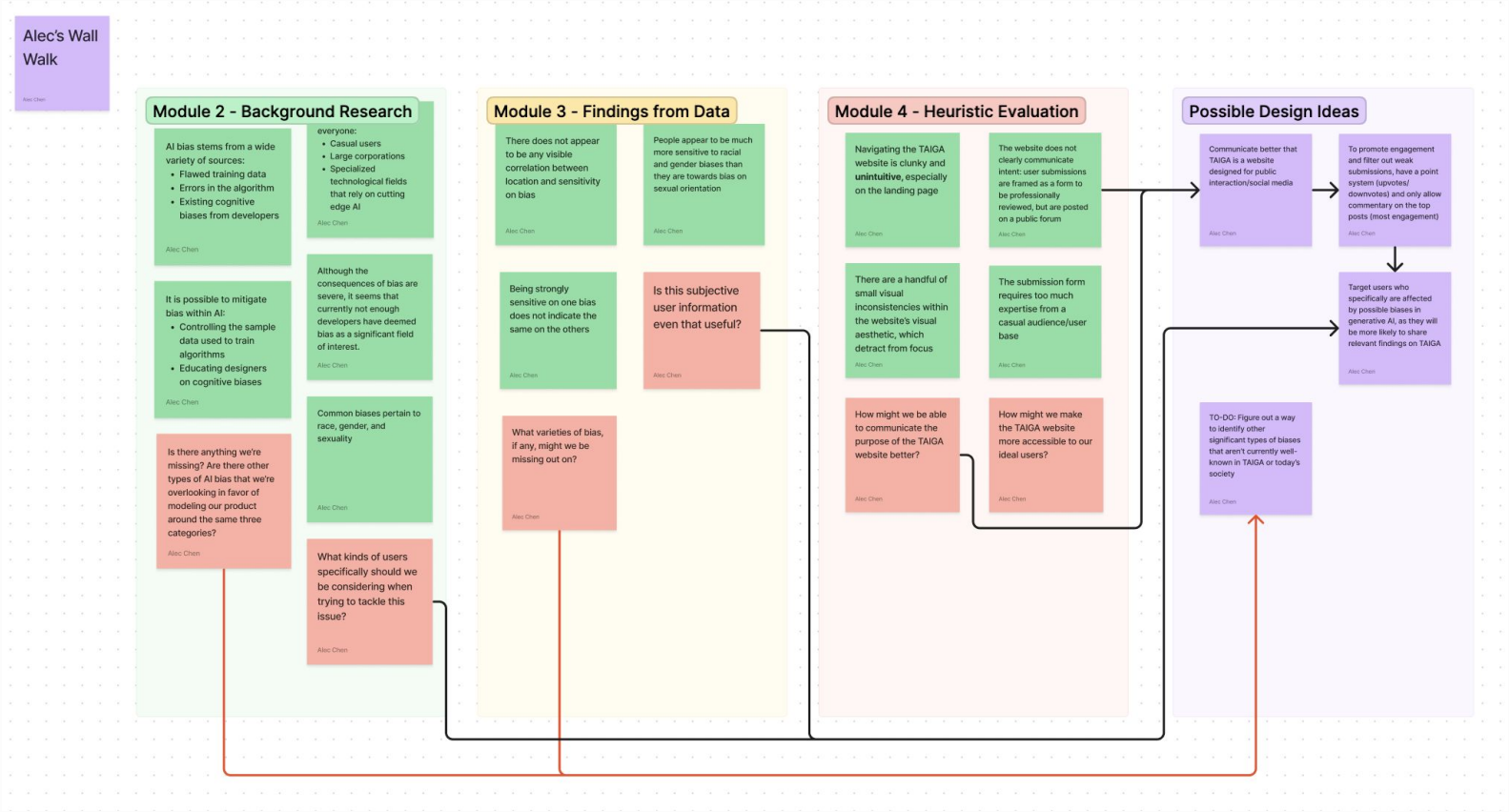




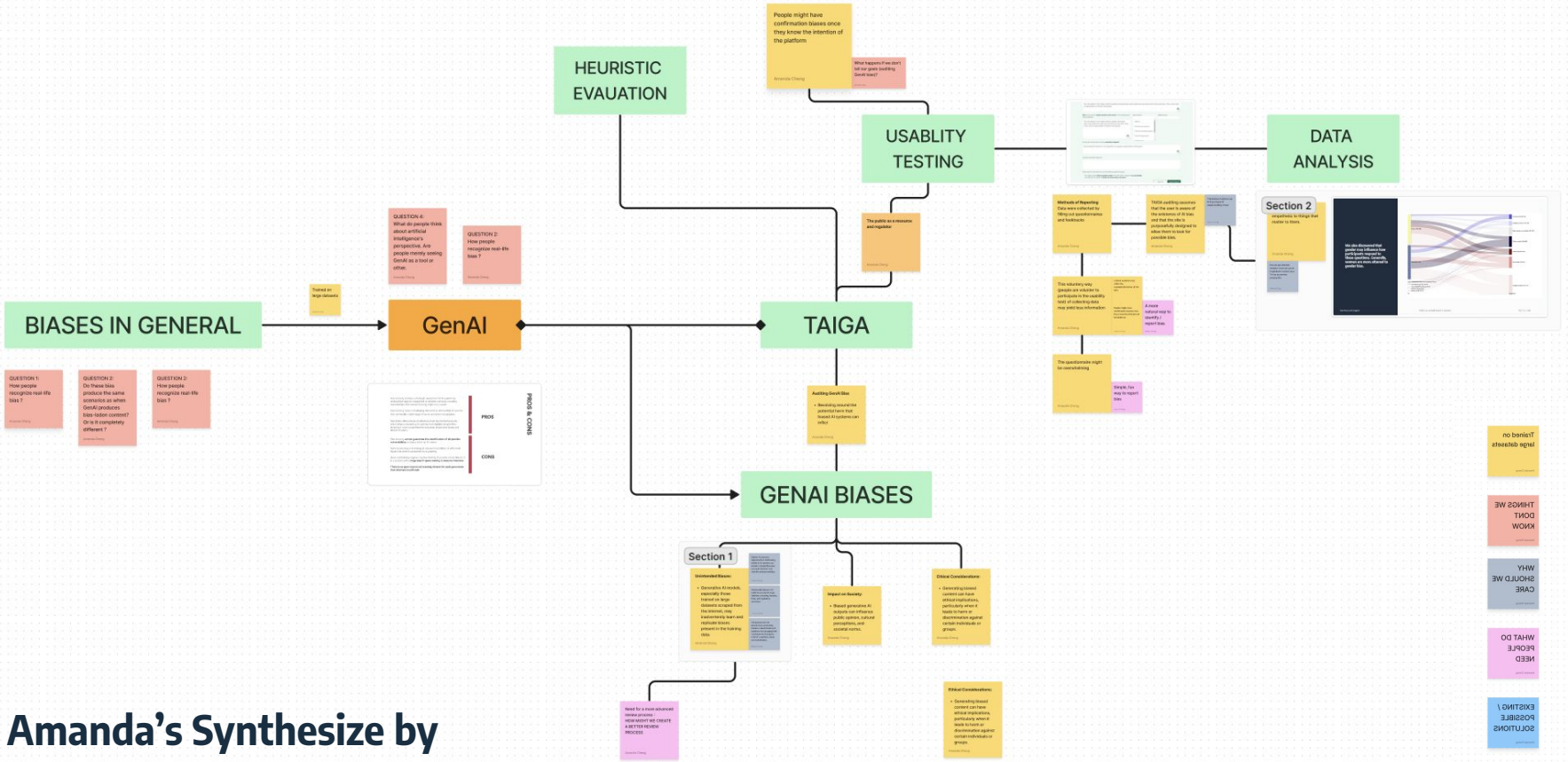


[Link to the Figma Board]

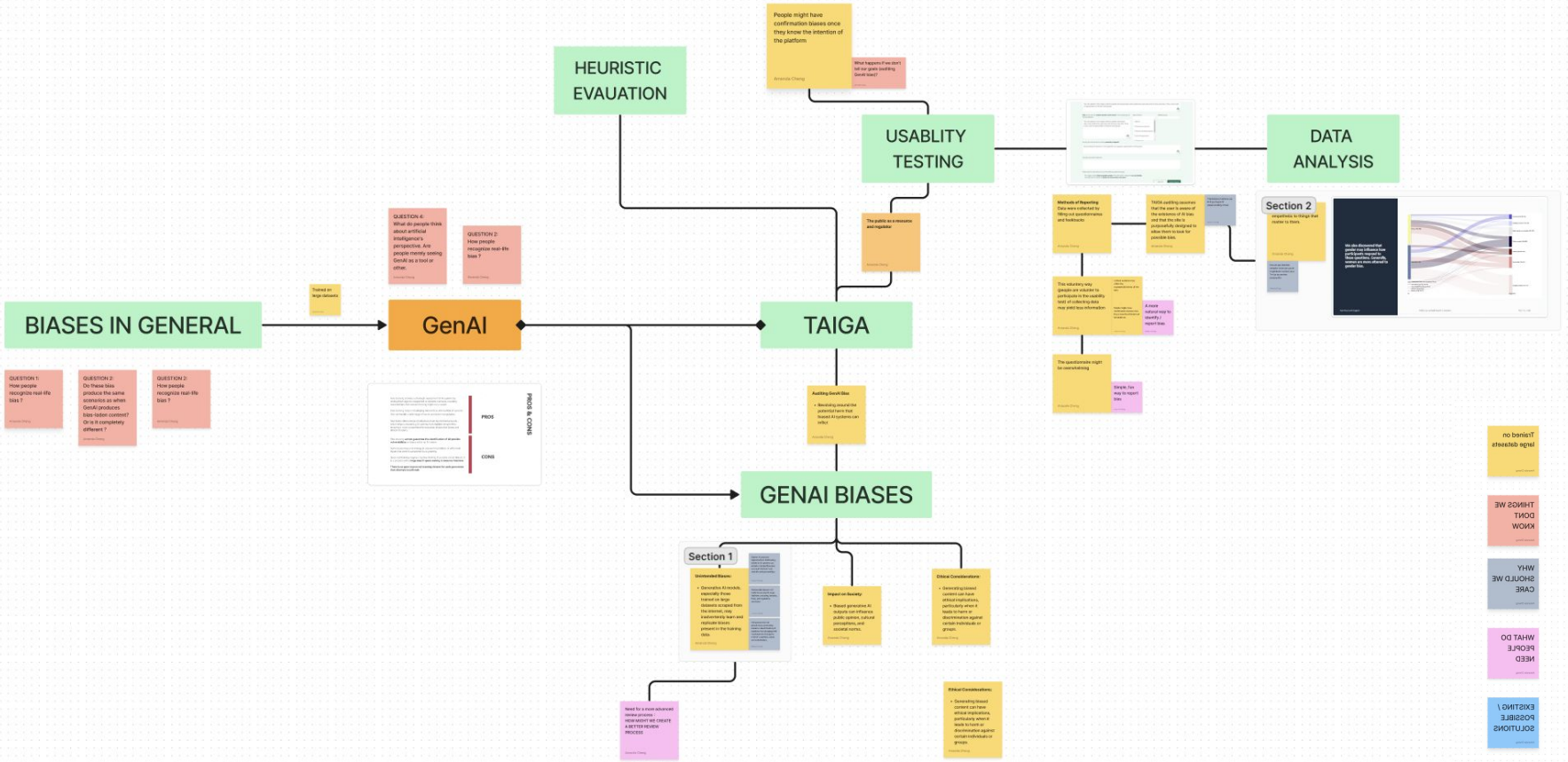
Nivedhitha's Synthesize by Walking the Wall



Alec's Synthesize by Walking the Wall

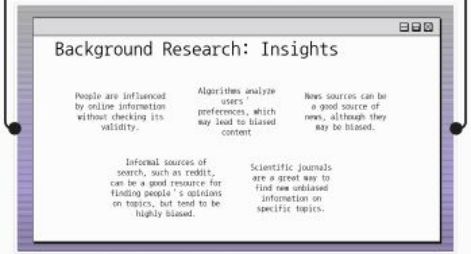


Amanda's Synthesize by Walking the Wall


HEURISTIC EVAUATION
USABLITY TESTING
DATA ANALYSIS
BIASES IN GENERAL
GenAI
TAIGA
GENAI BIASES
Section 1
Section 2
PROS
CONS
PROS & CONS

Sofia's Synthesize by Walking the Wall

Background Research



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

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 unharmful?

sofia

Heuristic Evaluation

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 distiguish 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 ?

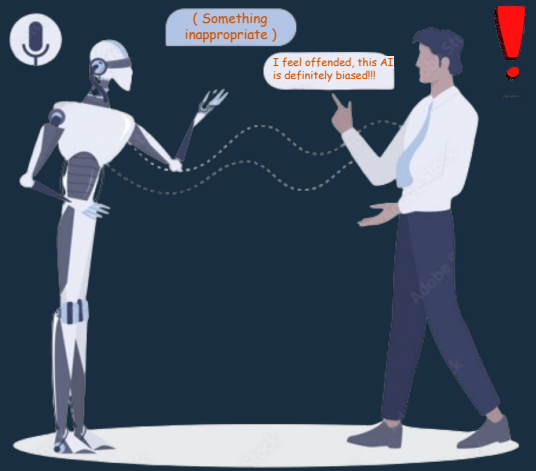
- Our team opted for the reverse assumption method to tackle the reframing activity as it allowed us to challenge and step beyond the boundaries of our established background context through various assignments already performed in this course.
- We realized that we all have different backgrounds in 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.

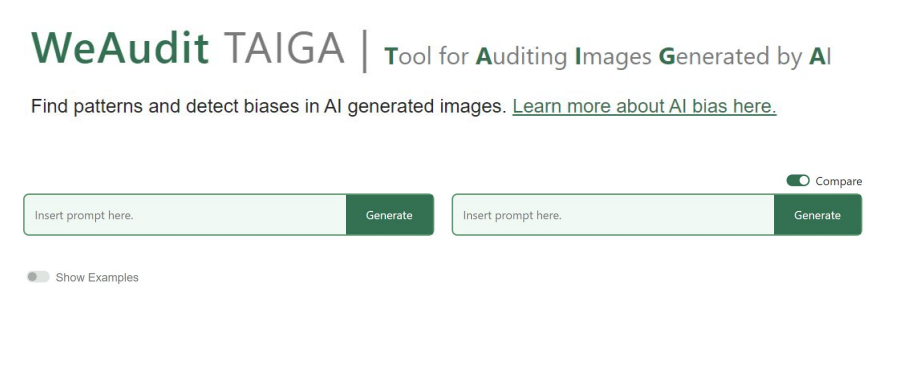
Discussed Reversals

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



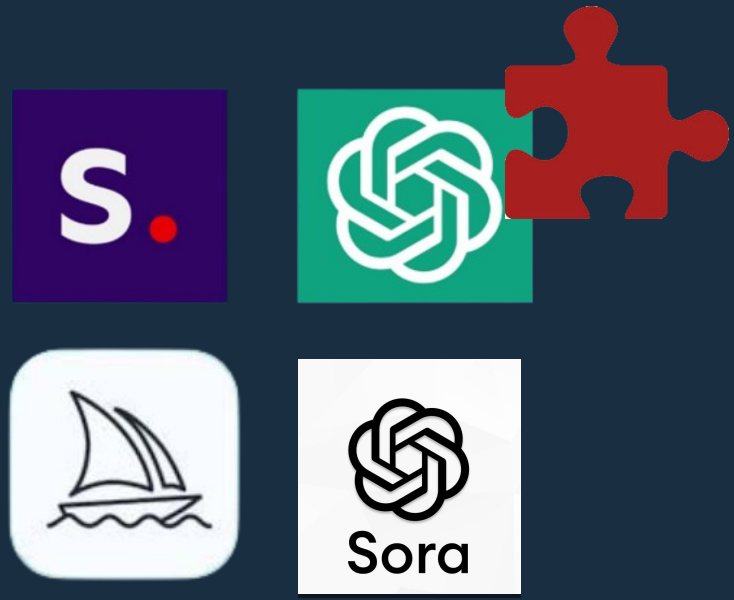
One Assumption we Explored

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



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.

03

Project Redefine

Link to Miro Board for the Redefine:
https://miro.com/app/board/uXjVNh7fuQA=/

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