



RESEARCH REPORT



Module 6: Deliverable I: Research Report (Team)

UCRE Spring 2024 Team D1 - Nasa GenAI Vanguard
Nivedhitha Dhanasekaran, Amanda Cheng, Alec Chen, Sofia Reyes Franco

02 APRIL 2024

Table of contents

01

**Executive
Summary**

02

**Research
Method Review**

03

**Insights with
Supporting
Evidence**

04

**Interpretation
Notes**

05

**Affinity
Clustering**

06

Modeling

Research Goals & Questions

How Might We:

- How can we make it more natural for users to report AI-generated bias?
- How can we use current user interactions to unexpected responses to motivate user auditing and reporting?

Related Questions driving our Contextual Inquiry:

- What forms of guidance and feedback are most effective for supporting users in detecting and reporting biases?
- How can we effectively educate users about the nature and presence of algorithmic biases within generative AI systems?
- What design elements in the user interface can prompt users to critically reflect on the responses they receive from generative AI systems?
- How can feedback mechanisms be integrated into generative AI platforms to facilitate easy reporting of detected biases by users?
- How can community-driven platforms enhance everyday users' detection and reporting of algorithmic biases?

How Might We...

How might we make it more natural for users to report AI-generated bias?

To address the challenge of user unawareness and engagement with AI-generated biases, our initiative seeks to simplify the process for users to identify and report these biases. Recognizing that users often overlook or don't critically evaluate AI biases, which affects the fairness and accuracy of AI results, we aim to enhance user awareness and interaction. We aimed to design intuitive user interfaces that encourage reflection and incorporating easy-to-use feedback mechanisms, and to utilize common user behaviors, to motivate more consistent auditing and reporting of biases.

How can we use current user interactions to unexpected responses to motivate user auditing and reporting?

Furthermore, the interest in using current user reactions to unexpected responses as a catalyst for auditing and reporting behavior highlights an innovative approach to enhancing user participation in quality control of AI outputs. Recognizing that users typically opt for re-prompting when faced with unsatisfactory AI responses, the goal is to integrate design approaches that make feedback provision a seamless and intuitive part of the user experience.

Executive Summary



Executive Summary

Research Objectives

Our primary objective is to foster a more intuitive environment for users to **recognize and report biases in generative AI (GenAI) systems**. This initiative recognizes the pivotal role of user **awareness and proactive engagement** in mitigating algorithmic biases, crucial for the fairness and accuracy of AI outputs.

Research Methods

Our team conducted **contextual interviews** and **direct storytelling** sessions with 5 participants, engaging them in tasks they typically perform using generative AI (GenAI), alongside presenting them with hypothetical scenarios for their reaction.

Research Synthesis

Through subsequent interpretation sessions for each interview, we utilized methods like **Affinity Diagramming**, **Empathy Maps**, and **User Journey Mapping** for an in-depth research synthesis. This process unveiled significant insights into user interaction and perception towards GenAI.

High-level Insights

Insights #1:

Users do not prioritize identifying biases in GenAI outputs, as their primary focus is on leveraging AI to support and everyday tasks.

Insights #2:

Current reporting mechanism is unnatural and doesn't fit into the natural workflow of users as they typically resort to re-prompting as an immediate solution to unexpected or unsatisfactory GenAI responses, sometimes even before the generation process is complete by interrupting the flow instead of looking for features to report this behavior.

Insights #3:

User apprehensions about anonymity and privacy loom large when reporting biases underscoring a critical barrier to transparency and accountability in addressing GenAI biases.

Insights #4:

User sensitivity to biases in real-life have little influence on reporting behavior since algorithmic biases don't stand out in the same way by eliciting a negative emotional response unless it is very obvious. Users need to be prompted or reminded to look for them in the responses - the more natural interpretation of results to look for how much the response matches their expectations.

Insights #5:

The reminder strategy and effort required to provide feedback through UI/UX elements on different GenAI tools determines likelihood of getting feedback from the users.

Research Method Review



Research Method Review

Rationale: Given our circumstances (i.e. time and resource constraints), a semi-structured interview worked best for us as it allowed us flexibility for scheduling the interview, allowing for different modalities, as well as conducting it, where both parties could influence the interview and follow-up questions could be asked to better fit/capture the participant's thought processes.

Summary of Research: Our research focused on evaluating how everyday users of Generative AI systems, such as ChatGPT and Stable Diffusion, perceive and can be empowered to detect harmful algorithmic behaviors. This involved collecting insights on user experiences, their ability to identify biases, and their willingness to participate in an everyday algorithm auditing process.

Participants: We tried to interview a diverse group of participants, including those with backgrounds in Chemical Engineering, Robotics, Credit Risk Analysis, Architecture, Fine Art and Electrical and Computer Engineering, hailing from geographical locations such as California, China, and Thailand. Our participants were selected such that they portray basic familiarity with AI technologies, ensuring they could provide informed insights into their interactions with these systems.

Research Process, Goals & Insights: The interviews were conducted on **Zoom** where participants described **diary studies** where participants logged their interactions with AI systems over a week. Our goal was to observe daily interactions of users with their choice of generative AI tools to elicit design ideas that fit into the natural workflow employed by users, and understand how to design AI interfaces that prompt critical reflection and easy reporting of biases. **We observed that users often do not spend much time analyzing AI-generated responses but rather opt to re-prompt the system. We also observed that regardless of user sensitivity to real-life bias, they do not prioritize thinking about biases in algorithmic biases as they are more concerned with obtaining responses to their prompts. These insights will guide our subsequent design formulations into how natural user behaviors could be leveraged to encourage the auditing and reporting of biases.**

Research Method Review

Rationale: Given our circumstances (i.e. time and resource constraints), a semi-structured interview worked best for us as it allowed us flexibility for scheduling the interview, allowing for different modalities, as well as conducting it, where both parties could influence the interview and follow-up questions could be asked to better fit/capture the participant's thought processes.

Summary of Research: Our research focused on evaluating how everyday users of Generative AI systems, such as ChatGPT and Stable Diffusion, perceive and can be empowered to detect harmful algorithmic behaviors. This involved collecting insights on user experiences, their ability to identify biases, and their willingness to participate in an everyday algorithm auditing process.

Participants: We tried to interview a diverse group of participants, including those with backgrounds in Chemical Engineering, Robotics, Credit Risk Analysis, Architecture, Fine Art and Electrical and Computer Engineering, hailing from geographical locations such as California, China, and Thailand. Our participants were selected such that they portray basic familiarity with AI technologies, ensuring they could provide informed insights into their interactions with these systems.

Multiple Professional Backgrounds

- Architecture
- Chemical Engineering
- Credit Risk Analysis
- Electrical and Computer Engineering
- Fine Art
- Robotics

Varied Geographical Locations

- United States
 - California
 - New York City
 - Texas
- China
- Thailand

Diverse Demographic / Ethnicity

Insights



Insight #1

Users do not prioritize identifying biases in GenAI outputs, as their primary focus is on leveraging AI to support everyday tasks.

The urgency with which users seek AI's assistance for efficiency in completing tasks overshadows their concern for biases. People often gravitate towards AI for its rapid, seemingly objective outputs. This inclination is rooted in fundamental human behavior that values immediate performance and results. As such, the efficiency and practical benefits provided by AI tend to take precedence over the scrutiny of potential biases, which may seem less tangible or immediate in the context of day-to-day utility.

Insight #2

Current reporting mechanism is unnatural and doesn't fit into the natural workflow of users as they typically resort to re-prompting as an immediate solution to unexpected or unsatisfactory GenAI responses, sometimes even before the generation process is complete by interrupting the flow instead of looking for features to report this behavior.

User Tendency to Re-prompt vs. Report

- **Re-prompting:** When encountering an issue or bias in GenAI outputs, users are more inclined to adjust their queries or prompts in hopes of receiving a better response, rather than using formal channels to report the problem. This indicates a preference for direct and immediate solutions.
- **Reporting:** Formal reporting mechanisms, although available, are used less frequently. This could be due to the perceived effort involved in reporting, the interruption it causes in the task flow, or skepticism about the effectiveness of reporting in yielding a timely resolution.

Misalignment with User Workflow

The insight suggests that existing mechanisms for reporting problems with GenAI outputs are not well integrated into the user's natural workflow. Ideally, addressing an issue should feel like a seamless part of using the AI, but if the process to report is cumbersome or interrupts the task flow, users are likely to avoid it.

Insight #3

User apprehensions about anonymity and privacy loom large when reporting biases underscoring a critical barrier to transparency and accountability in addressing GenAI biases.

Users' hesitation to discuss personal matters with an AI can be rooted in the perception that AI lacks the emotional empathy and understanding that a trusted friend or confidant possesses. It's not just about data privacy, but also the quality of interaction and the type of response expected. People naturally seek empathy and a nuanced understanding when sharing sensitive information, which they may feel an AI system cannot provide. There is a belief that AI might offer neutral or 'correct' responses, but these can be devoid of the human warmth and genuine concern that personal interactions afford. This apprehension can prevent users from fully engaging with AI in scenarios that require a degree of vulnerability or emotional complexity.

User sensitivity to biases in real-life have little influence on reporting behavior since algorithmic biases don't stand out in the same way by eliciting a negative emotional response unless it is very obvious. Users need to be prompted or reminded to look for them in the responses - the more natural interpretation of results to look for how much the response matches their expectations.

User Sensitivity to Biases in Real Life vs. Algorithmic Biases

- **Real-life biases:** In everyday situations, individuals might be more attuned to biases because they can directly relate these biases to social, cultural, or personal experiences. These biases often elicit a strong emotional response because they can affect one's sense of fairness, identity, or beliefs.
- **Algorithmic biases:** When interacting with algorithms (such as those powering search engines, recommendation systems, or AI chatbots), users may not immediately recognize biases. This is partly because algorithmic outputs are perceived as neutral or objective calculations. Unless the bias is glaringly obvious, it might not trigger the same emotional response as real-life biases.

Influence on Reporting Behavior

The insight suggests that the subtlety of algorithmic biases means they are less likely to influence user reporting behavior. Since these biases don't elicit a strong emotional response, users may not feel compelled to report or criticize them unless prompted to do so. This could be due to a lack of awareness about the existence of algorithmic biases or an assumption that algorithmic decisions are inherently objective.

Insight #5

The reminder strategy and effort required to provide feedback through UI/UX elements on different GenAI tools determines likelihood of getting feedback from the users.

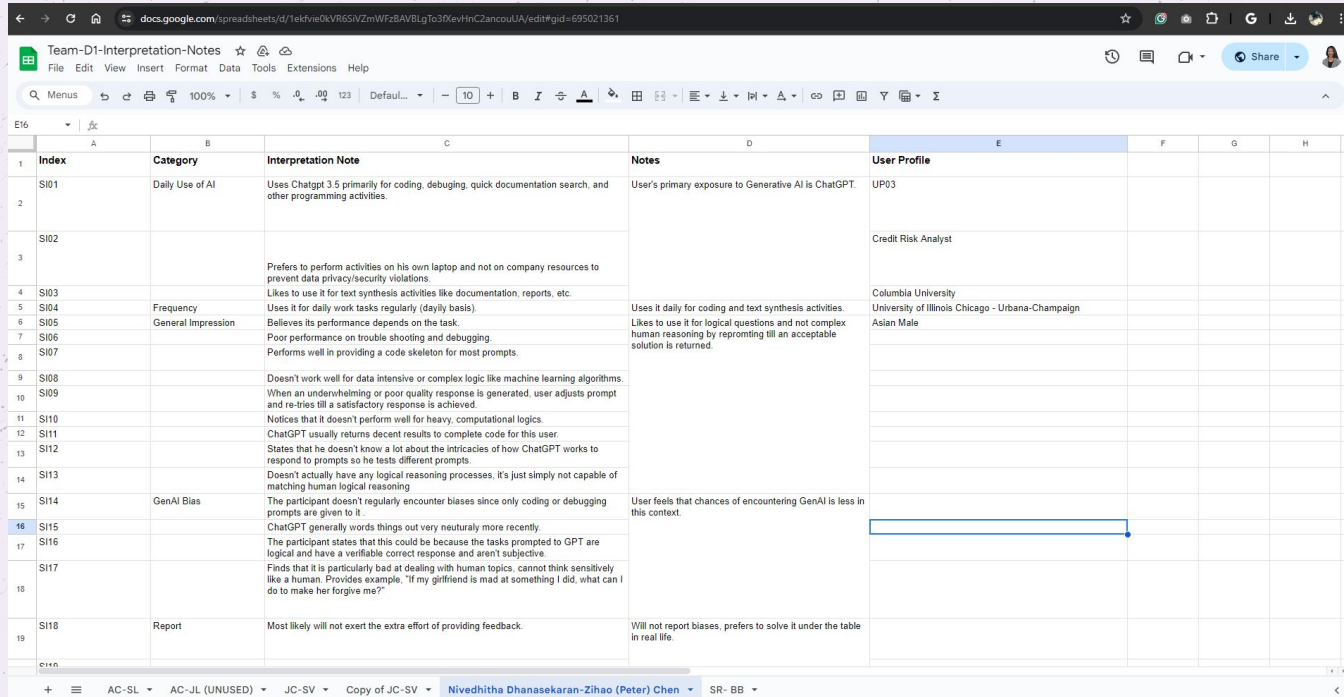
The effectiveness of UI/UX elements as prompts for feedback is closely tied to the ease with which users can engage with them. If the effort to provide feedback is perceived as too high or the reminder strategy is not compelling, users are less likely to participate. This ties back to the psychological principle of effort justification, where the perceived benefits of an action must outweigh the effort required. If providing feedback is simple and seamlessly integrated into the user experience, it's more likely to be utilized. Moreover, the common user response to re-prompting during unsatisfactory AI interactions exemplifies a behavioral tendency towards the path of least resistance. Users favor immediate and straightforward actions over more complex ones, such as submitting detailed feedback.

Appendix: Interpretation Notes



Appendix: Interpretation Notes

<https://drive.google.com/drive/folders/1EKQhguqMtPpeDy5lw1ol-CxEL8CQu2tz?usp=sharing>



The screenshot shows a Google Sheets spreadsheet with the following data:

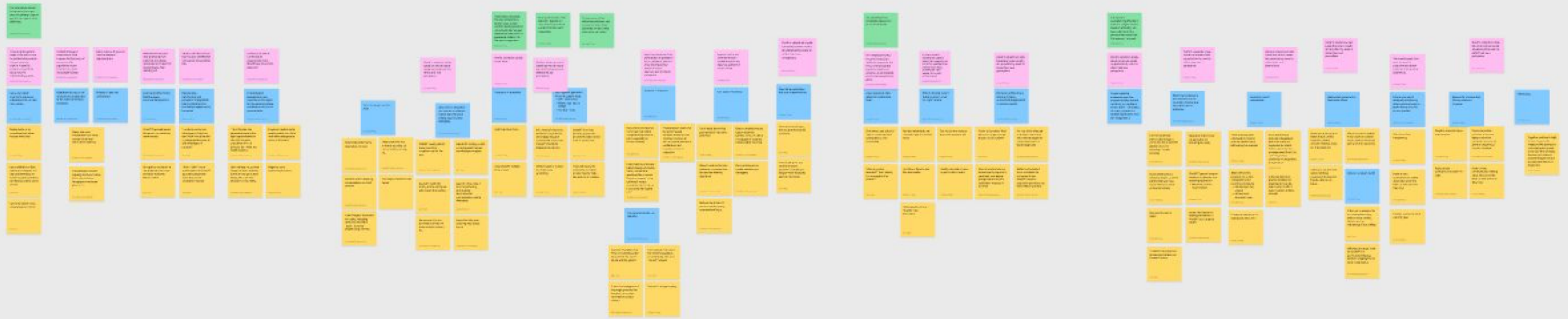
Index	Category	Interpretation Note	Notes	User Profile
SI01	Daily Use of AI	Uses Chatgpt 3.5 primarily for coding, debugging, quick documentation search, and other programming activities.	User's primary exposure to Generative AI is ChatGPT.	UP03
SI02				Credit Risk Analyst
SI03		Prefers to perform activities on his own laptop and not on company resources to prevent data privacy/security violations.		Columbia University
SI04	Frequency	Likes to use it for text synthesis activities like documentation, reports, etc.	Uses it daily for coding and text synthesis activities.	University of Illinois Chicago - Urbana-Champaign
SI05	General Impression	Believes its performance depends on the task.	Likes to use it for logical questions and not complex human reasoning by reprompting till an acceptable solution is returned.	Asian Male
SI06		Poor performance on trouble shooting and debugging		
SI07		Performs well in providing a code skeleton for most prompts.		
SI08		Doesn't work well for data intensive or complex logic like machine learning algorithms.		
SI09		When an underwhelming or poor quality response is generated, user adjusts prompt and re-tries till a satisfactory response is achieved.		
SI10		Notifies that it doesn't perform well for heavy, computational logics.		
SI11		ChatGPT usually returns decent results to complete code for this user.		
SI12		States that he doesn't know a lot about the intricacies of how ChatGPT works to respond to prompts so he tests different prompts.		
SI13		Doesn't actually have any logical reasoning processes, it's just simply not capable of matching human logical reasoning		
SI14	GenAI Bias	The participant doesn't regularly encounter biases since only coding or debugging prompts are given to it.	User feels that chances of encountering GenAI is less in this context.	
SI15		ChatGPT generally words things out very neutrally more recently.		
SI16		The participant states that this could be because the tasks prompted to GPT are logical and have a verifiable correct response and aren't subjective.		
SI17		Finds that it is particularly bad at dealing with human topics, cannot think sensitively like a human. Provides example, "If my girlfriend is mad at something I did, what can I do to make her forgive me?"		
SI18	Report	Most likely will not exert the extra effort of providing feedback.	Will not report biases, prefers to solve it under the table in real life.	

Appendix: Affinity Clustering



Appendix: Affinity Clustering

<https://www.figma.com/file/BGVlKWsz6MlvXN4bZBwRq/Team-D1-Team-Contract-Spring24?type=whiteboard&node-id=0-1&t=t3k5pH1Hh8a5x8oU-0>



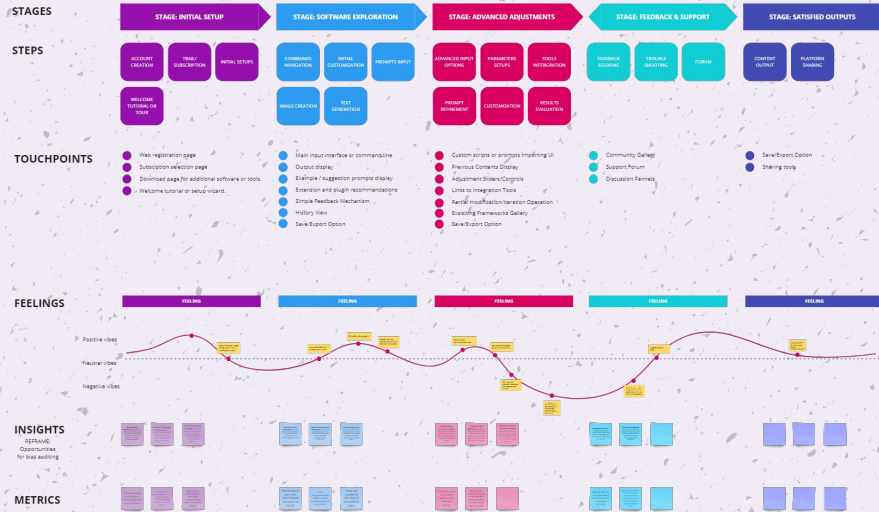
Appendix: Modeling



Appendix: Modeling – Journey Map

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

User Journey Map



Reflection

Contributions

- Amanda Cheng
 - All slides Research Report
- Nivedhitha Dhanasekaran
 - All slides Research Report
- Alec Chen
 - Executive Summary (revisions)
 - Research Method Review (revisions)
- Sofia Reyes Franco
 - Research Report (insight #2 & 4 and other revisions)

Attendance

- Friday
 - No meeting (midterm week)
- Sunday Morning
 - Amanda
 - Nive
 - Alec
- Sunday Evening (Final Submission for F, H)
 - Amanda
 - Nive
- Tuesday Evening (Final Submission for I)
 - Amanda
 - Nive

Thank You