SEVEN
Anna, Aundre, Kai, Kaitlyn, Ye Wei

M6.3 | Deliverable H: Research Report

Table of Contents

Research Question

*How might we elevate or **improve college students' understanding** of **algorithmic bias' causes** and **effects** within popular gen-AI platforms such as ChatGPT?*

Executive Summary

Activities Conducted

For approaching our research, we established our interview protocol (think-aloud study) and found potential candidates for our interviews. We each conducted 1 think-aloud study each, following our protocol, including pre-interview questions, a **think-aloud study** using a standardized prototype, and post interview questions. Following, we conducted **interpretation sessions** for each, sharing insights. We then utilized **affinity clusters** to better form and mull over our insights to further understand our data. We created 2 models to better understand our users: user **personas**, and a **customer journey map**.

High Level Insights

1. A users' ability to identify bias is inherently linked to their understanding of discrimination, bias, and prejudice.
2. There is a common belief that AI tasked with menial or non-creative tasks won't produce biased consequences.
3. User could benefit from a better understanding of the development and operations behind Gen-AI models when identifying algorithmic biases.
4. Distrust arises in generative AI, especially on human-centric topics, due to lack of transparency, single-perspective responses, and inherent skepticism.
5. Bias detection in generative AI is influenced by external factors and improves when bias is a focal point of discussion.

Next Steps

Our next steps are to ideate new solutions to better improve knowledge on generative AI bias. We will utilize speed dating for this process along with a group review session to refine and iterate upon our ideas.

Research Methods

Summary: Our group conducted a think-aloud study simulating image generation tasks to better understand the breadth and depth of college students' knowledge of bias in generative AI platforms.

Goal: Given that our research question is seeking to elevate or improve the current understanding, we needed to determine to what extent college students are able to identify and evaluate bias in gen-AI outputs. Therefore, we needed to get a sense of how students observed and interacted with the gen-AI platform, allowing us to see what they thought was worth commenting on.

Participants: We found a five college students to participate in our think-aloud study. We recruited students from our personal networks but avoided students who were already familiar with this project (such as our UCRE classmates) or particularly knowledgeable about our research topic (experts).

Process & Methods: We followed the script included in our Revised Research Guide and created pre-scripted generated image results for each task prompt. This ensured that each participant experienced the same process regardless of the interviewer and variances in gen-AI output and helped us researchers better identify commonalities and recurring themes among their responses.

Insights with Supporting Evidence

Insight #1

A users' ability to identify bias is inherently linked to their understanding of discrimination, bias, and prejudice.

Much of the bias identified, stemmed from the participants' prior knowledge/experience with bias, discrimination, and prejudice, whether it was recognizing a common stereotype or relating the outputs to societal norms. Furthermore, most of the participants felt that improving media literacy could increase the awareness of algorithmic bias across users.

Insight #2

There is a common belief that AI tasked with menial or non-creative tasks won’t produce biased consequences.

With menial tasks there is an expected output from users, for instance when prompted to code the AI should produce code that follows the prompter’s specifications. In this case, it is easier for users to verify the quality of the generated content because it’s about accuracy. However, with more creative work, like generating an image of a generic situation, it can be more challenging for some users to identify biases because generic situations can be interpreted in a vast amount of ways. The accuracy in this case is more aligned with a user’s beliefs than it is with one standardized answer.

Insight #3

User could benefit from a better understanding of the development and operations behind Gen-AI models when identifying algorithmic biases.

Many of the participants attributed the bias produced by the Gen-AI to a lack of diverse training data and developers without necessarily understanding how and where the outputs for a prompt come from. Users with prior knowledge of gen-AI bias tended to feel more confident when tasked with its identification. The realization of the black box effect between users and the models led us to believe that more awareness of how these systems work would lead to more confidence in distinguishing biased from non-biased AI generated content.

Insight #4

Distrust arises in generative AI, especially on human-centric topics, due to lack of transparency, single-perspective responses, and inherent skepticism.

To build trust, AI systems should provide clear source citations, offer multiple perspectives on subjective topics, and enhance transparency in response generation. Given that skepticism is heightened for intangible concepts like emotions and external information, AI should incorporate explainability features that clarify reasoning behind outputs. Users tend to distrust AI more after encountering biased content, so systems should allow for easy comparison of alternative responses to expose and mitigate bias. Additionally, as AI-generated images often make untrustworthiness more apparent, visual outputs should include context, disclaimers, or interactive elements to help users assess credibility.

Insight #5

Bias detection in generative AI is influenced by external factors and improves when bias is a focal point of discussion.

To enhance bias detection, AI systems should reduce confusing or overly abstract results that may hinder users' ability to identify bias. When generating content like images, it is crucial to offer more diverse visual and stylistic variations, as a lack of diversity can make biases more apparent. Clearer framing in prompts and pre-engagement discussions on bias can help users critically assess outputs, especially when they are made aware of bias risks upfront.

Appendix

Table of Contents

Interpretation Notes

	Note		User Profile
SH1-1	The user cares more about how close is the result to the prompt instead how biased the result is.		SH1 - Stakeholder One
SH1-2	The reason why gen-AI is biased is because it was trained with old data.		CMU Student
SH1-3	It is easier for the user to find biases in the results than the prompts.		Major in ETC
SH1-4	Confusing results will effect people's judge on bias.		ChatGPT, Everyday Use, Fix bugs
SH1-5	Use AI to improve efficiency.		
SH1-6	Gen-AI cannot help solve some detailed problems.		
SH1-7	Users care more about how quickly they can regenerate the result to receive the result they need.		
SH1-8	Understanding machine learning can help people find bias problems more easily.		
SH1-9	People only want to regenerate the result if the results they received are not good for them.		
SH1-10	The participant had different standard on general prompts and detailed prompts.		
SH1-11	The old dataset has too many biased problems		
SH1-12	The participant can easily find biased problems in race and gender.		
SH1-13	Don't care too much about the diversity in the results.		
SH1-14	The participant didn't learn more about bias problems in gen-AI after the experiment.		
SH1-15	After seeing white male docotor images, the pariticipant only thought the result was missing the women doctors.		
SH1-16	People will realize the possible biased results in the prompts at second thoughts		
SH1-17	It harder for users only using text generaing AI to find biases.		
SH1-18	People's standard on some words like "healthy " varies a lot.		
SH1-19	The distraction will make users less alert to bias.		
SH1-20	Bias will not influence this participant's motivation in using AI.		

	Note		User Profile	
SH2-1	Expectations of Gen-AI output depends on use case		SH2 - Stakeholder ___	
SH2-2	For writing editing tasks, expectations are higher for quality output		CMU Student	
SH2-3	For technical questions, expectations are much lower		Major in ECE	
SH2-4	Trustworthiness/quality "only matters if I don't know the answer already"		Sometimes uses gen AI tools	
SH2-5	Considers how believable the output is as a gut reaction when evaluating AI		most questions about course topics	
SH2-6	For technical questions, can do a quick google search to fact check info			
SH2-7	Never fully trusts gen-ai because there is no reasoning or explainability			
SH2-8	Believes all users should be skeptical of the accuracy of results			
SH2-9	Bias comes from training data and usually reflects societal biases			
SH2-10	Why are all the images of "doctors" all men?			
SH2-11	Thought that the results showed a clear gender bias and lacked representation			
SH2-12	Why are all the images of "healthy people" all men?			
SH2-13	When all the images also look really similar in style and composition, the lack of diverstiy becomes even more apparent.			
SH2-14	Concerned about lack of diversity in images generated			
SH2-15	"Wouldn't want to use a platform with obvious bias in the results"			
SH2-16	Lack of diverse results reduces users' trust in the overall result			
SH2-17	Easier to identify bias when the model is generating multiple results for one prompt			
SH2-18	Popular gen-ai platforms (e.g. ChatGPT, Claude) typically only generate one response making it harder to see bias without comparing responses.			
SH2-19	Most students use gen-ai for general overviews, so most prompts likely wouldn't generate similarly harmful images.			
SH2-20	These results fine for simple use cases (e.g. remembering a concept)			
SH2-21	These results fine for simple use cases (e.g. creating an educational presentation)			
SH2-22	Identified bias more easily because it was discussed before the generatiom task.			
SH2-23	Bias is not necessarily harmful as shown with the retail example.			
SH2-24	Biased results reflect a lack of diversity in training data.			
SH2-25	Biased results can also reflect a lack of diversity in model developers.			
SH2-26	"Would expect Gen-AI to produce similar results to what I imagine in my head"			
SH2-27	Would be helpful to see how satisfied other people are with similar queries			
SH2-28	Would be helpful to see a disclaimer about possible bias along with search results			
SH2-29	Would be helpful to see a confidence meter in how accurate the model is being			
SH2-30	Curious to see what kind of data the model is pulling from to generate these images			

Interpretation Notes (cont.)

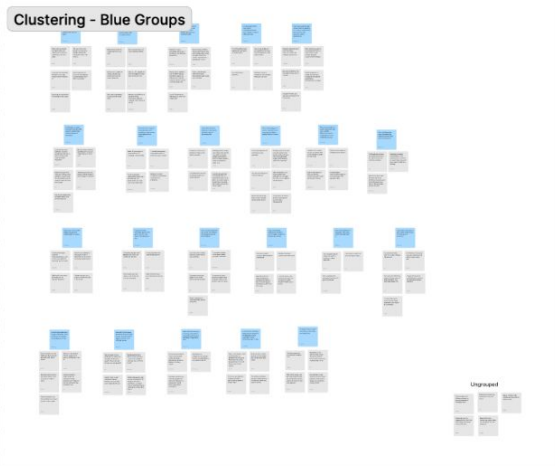
	Note	User Profile
SH3-1	Believes AI is more susceptible to bias due to the answer only being one perspective	SH3 - Stakeholder Three
SH3-2	Believes AI is more susceptible to bias due to no citations for information	CMU Student (Junior)
SH3-3	Doesn't trust AI for external information, only for improving own personal ideas because easy to fact check latter	ECE Major
SH3-4	Felt that all prompts were unbiased as they were broad and general	Uses AI tools frequently for coding + coursework (mainly GPT4o)
SH3-5	AI-generated doctors match general public perception of doctors	
SH3-6	Confused when AI didn't even generate people in the "person working in retail shop" prompt	
SH3-7	Noticed that images of POC look less like "people" than white males generated previously	
SH3-8	Felt that images reinforced stereotypes of who can actually work prestigious jobs	
SH3-9	Inferred that bias in last prompt partly due to actual bias and partly due to technical/user error since all images looked exactly the same	
SH3-10	"Healthy" prompt didn't actually generate a wide variety of healthy people (ie. non-white)	
SH3-11	After completing the think-aloud users trust of gen-AI decreased	
SH3-13	AI-generated images make the untrustworthiness of gen-AI tools more apparent than other gen-AI content	
SH3-13	Confused about the definition of a biased prompt since felt that an ideal AI should avoid leading prompts regardless	
SH3-14	Interprets potential bias in technical writing (resumes, emails) and coding as less harmful	
SH3-15	Interprets potential bias in technical writing (resumes, emails) and coding as more difficult to identify	
SH3-16	Couldn't come up with potential interfaces/features that could help users better identify bias	
SH3-17	Emphasized that even with AI content warnings -- many users will glaze over them and still blindly trust AI	
SH3-18	Felt that core solution to better evaluating bias would be stronger education/emphasis on liberal arts (history, english)	
SH3-19	Saw interactive collaborative workshops focusing on media literacy as a potential solution	
SH3-20	Expressed uncertainty in motivating people who don't enjoy liberal arts to participate in media literacy educational experiences	

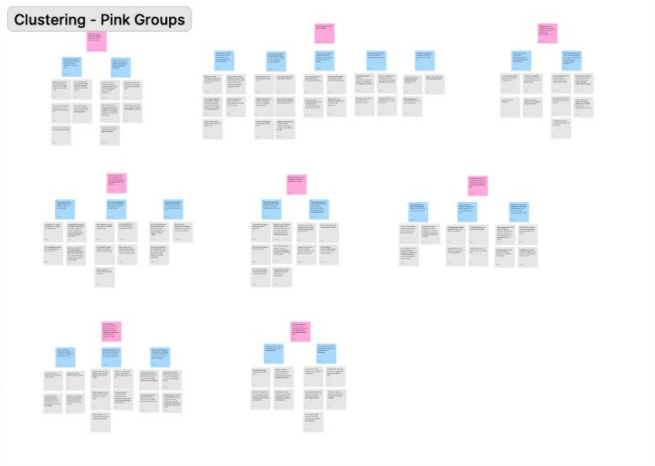
Interpretation Notes (cont.)

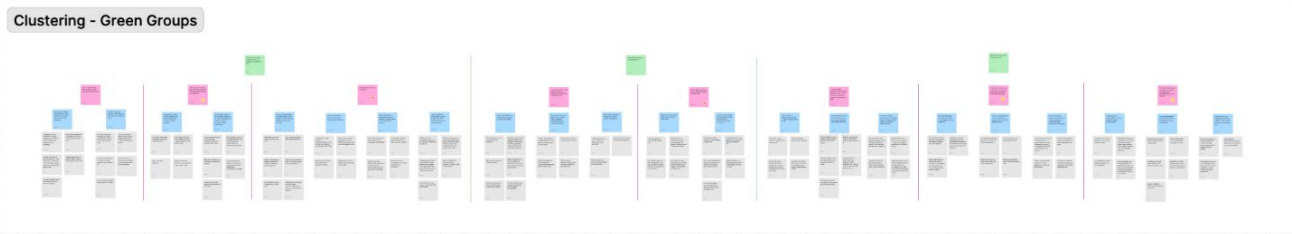
	Note		User Profile
SH4-1	User is more skeptical of the outputs of gen-AI after initial encounter with biased generated content		SH4 - Stakeholder Four
SH4-2	Noted lack of diverse interpretations of the prompt given to AI (focus on convenience store over other types of retails like clothing)		CMU Student
SH4-3	Similarly, much of the bias detected related to the absence of diversity from the various generated images (predominantly white males depicted)		Finance Major
SH4-4	Skepticism of AI also stems from previous readings about gen-AI		occasional gen-AI use (mostly for school)
SH4-5	Seems less skeptical of AI's capabilities with less creative applications, like graph creation		
SH4-6	The feeling of familiarity with gen-AI brought more confidence to identifying algorithmic bias		
SH4-7	Uncertainty with the relationship between prompt and bias led to a want for more focus on the algorithms rather than their usage		
SH4-8	Detecting possible bias from prompt leads to a heightened awareness of bias in the output (what is the definition of healthy, does healthy account for mental health)		
SH4-9	Stereotypes from society can help to easier identify bias within generated content, seen with the doctor example (doctors being stereotypically white men)		
SH4-10	More abstract results confuses users on if bias is present (some doctor images were neckdown confusing user)		
SH4-11	Assistance with prompt creation could benefit users in auditing bias from AI		
SH4-12	Some casual users don't take much notice of the role UI features have on their experience utilizing AI		
SH4-13	Harm can be perceived from any bias detected (racial, gender, health)		
SH4-14	"More prompts seem to worsen the resullt" in reference to image generation		
SH4-15	Results should match the specificity of the prompt, "I guess person is white man"		
SH4-16	Bias from AI, while harmful, can also enhance the perception that AI is dangerous		
SH4-17	Bias was easy to detect because participants were given comparisons across examples, rather than singular responses like popular gen-AI give		
SH4-18	While being aware of bias, the bias displayed across images stood out to participants, so smaller sizes of responses may be challenging to detect bias within		
SH4-19	Expectations of what general prompts should produce is based on the person's societal outlook		
SH4-20	Users' unawareness to the actual functions of generative AI can lead to frustration with unmet expectations of what will be generated from a prompt		

	Note		User Profile
SH5-1	Believes AI more for answers on more tangible things such as code		SH5 - Stakeholder Five
SH5-2	Believes AI is capable of making mistakes, but overall trust depends model being used		Boston University Student
SH5-3	Distrusts AI's capability for more untangible concepts such as emotions, etc.		CS Major
SH5-4	Felt that the healthy person prompt was biased		Uses AI tools frequently for coursework (mainly GPT and Claude)
SH5-5	Seems skeptical on image responses and differences between them		
SH5-6	Confused when AI didn't even generate people in the "person working in retail shop" prompt		
SH5-7	Uncertain whether bias from prompt or bias from AI caused bias within image for healthy person prompt		
SH5-8	Believes that bias in AI may behave like an "echo chamber" of sorts where AI may also give a person answers it may think they want to hear		
SH5-9	Noticed visual and stylistic image details generated in doctor prompt, rather than racial or gender bias at first		
SH5-10	Trust of gen-AI after think aloud decreased slightly		
SH5-11	Believes that bias in gen- AI is ineveitable		
SH5-12	Having a prompt at bottom of gen-AI somewhere easily readable would help user know to think about bias in general more		
SH5-13	Confused about the definition of a biased prompt since felt that an ideal AI should avoid leading prompts regardless		
SH5-14	More easily identified bias because of parameters of think aloud since they were made aware of it		
SH5-15	Confusion on bias in prompt in spefically within retail prompt.		
SH5-16	Emphasis that users may disregard for bias in AI overall		
SH5-17	Expected bias further after first couple of biased image series		
SH5-18	Gen-AI bias was not important to participant, but participation made them more aware		
SH5-19	Was confused on more additional information that may help, confusion on specfics that could help (answer was very general)		
SH5-20	Evaluated bias through typical themes within society and social media		

Affinity Clusters: Figjam Link







Model #1: Persona

Mike

Student at Boston University

From Tulsa | 19 years old

Skeptical — Open-minded

Barely Utilizes AI — Everyday User

Neutral on Social Issues — Cares about Social Issues

BIO

Mike is a sophomore undergraduate student who occasionally uses generative AI for convenience, such as summarizing articles or answering quick questions. He doesn't think much about AI beyond its usefulness and doesn't actively consider its ethical implications. He hasn't engaged with discussions about AI bias and doesn't see it as a major concern, as long as AI tools work efficiently for him.

MOST USED GEN AI

- ChatGPT
- ClaudeAI
- StableDiffusion

AI KNOWLEDGE

- How It Works
- Knowledge of AI Bias
- Reportability of Bias

GOALS AND MOTIVATIONS

- Uses AI mainly for efficiency and convenience
- Wants AI to streamline schoolwork, not add complexity
- More interested in practical applications than ethical discussions
- Open to learning about AI bias only if it directly impacts AI performance

FRUSTRATIONS

- Gets annoyed when AI hallucinates wrong answers
- Dislikes when AI tools are slower or harder to use than just Googling something

PAINPOINTS

- Does not think AI bias is relevant to his daily use of AI
- Unaware of AI ethics discussions and doesn't seek them out
- Only cares about AI when it stops being useful
- Sees AI as a tool, not a topic of interest

Model #1: Persona (cont.)

Maya

Student at Carnegie Mellon University

From LA | 20 years old

Skeptical — Open-minded

Barely Utilizes AI — Everyday User

Neutral on Social Issues — Cares about Social Issues

BIO

Maya is a junior at Carnegie Mellon University who utilizes generative AI at times to learn about new things, and help her with coursework at times. She does not know much about bias within generative AI, but is up to date on current issues.

MOST USED GEN AI

- ChatGPT
- ClaudeAI
- StableDiffusion

AI KNOWLEDGE

- How It Works
- Knowledge of AI Bias
- Reportability of Bias

GOALS AND MOTIVATIONS

- utilize AI to learn more about unknown topics
- break down information
- receptive to learning about issues
- open to learning but may initially struggle to see why AI bias is relevant to her studies.

FRUSTRATIONS

- when AI does not provide correct details
- when AI does not provide helpful responses

PAINPOINTS

- does not know what AI bias is or why it matters
- may not be immediately engaged without relatable examples
- trusts AI day to day without second thought

Model #1: Persona (cont.)

Alex

Student at Carnegie Mellon University

From Seattle | 22 years old

Skeptical — Open-minded

Barely Utilizes AI — Everyday User

Neutral on Social Issues — Cares about Social Issues

BIO

Alex is a senior undergraduate student majoring in Computer Science with a minor in Ethics. Through coursework and independent research, they have developed a deep understanding of AI bias and its implications in areas like hiring, law enforcement, and healthcare. Alex actively participates in discussions about responsible AI and fairness in machine learning models, often critiquing AI systems for perpetuating discrimination. They believe that AI should be transparent, accountable, and equitable, and they are passionate about advocating for ethical AI practices.

MOST USED GEN AI

- ChatGPT
- ClaudeAI
- StableDiffusion

AI KNOWLEDGE

- How It Works
- Knowledge of AI Bias
- Reportability of Bias

GOALS AND MOTIVATIONS

- Wants to help create fair and unbiased AI systems
- Advocates for transparency and accountability in AI decision-making
- Interested in policy and regulation around AI ethics
- Seeks hands-on experience with AI fairness tools and frameworks

FRUSTRATIONS

- Finds it difficult to convince non-technical people that AI bias is real and harmful
- Annoyed by AI models that reproduce stereotypes despite fairness efforts
- Struggles with bias mitigation trade-offs in AI development

PAINPOINTS

- Worries that AI bias can have serious real-world consequences
- Limited access to unbiased training data makes solving the problem harder

Model #2: Customer Journey Map

Jimmy

Scenario: Jimmy is an undergraduate college student who frequently uses AI every day to help him complete various tasks.

Expectations

- Hope for less bias in AI results.
- Hope for results that align more closely with the prompt and needs.
- Able to quickly identify bias in the results.

	Input prompt	Generate results	Observe results	Reflect on results
Doing	• Check the purpose and use the certain type of AI tools. • Input prompt.	• Wait for AI to generate results.	• Check the result of AI. • Compare the result with the prompt. • Pick the result. • Adjust the prompt. • Regenerate the results.	• Think how to improve the prompt to get the result that users want. • Want to reduce biased results
Thinking	• An ideal gen-AI system should provide me with results it expects while employing potential safeguards. • I was confused on whether a prompt might or might not produce gen-AI bias. • I felt that additions to the gen-AI interface could provide more awareness surrounding gen-AI bias.	• I prioritize speed and accuracy when evaluating the usefulness of a gen-AI result. • I have some inherent degree of skepticism regarding gen-AI outputs.	• I'm more skeptical of gen-AI results related to more "human" concepts such as dealing with personal emotions. • I lack trust in gen-AI especially because AI provides no clear sources for the information it outputs. • I lack trust and see through bias in gen-AI because there isn't multiple perspectives/results provided. • When processing gen-AI results, I was concerned by the lack of diversity.	• I trust gen-AI for more menial, straightforward tasks such as graph creation or editing grammar. • I feel that gen-AI bias is less harmful or apparent for menial, straightforward tasks such as code correction, editing grammar, etc. • When processing gen-AI results, I could identify bias by comparing marginalized and non-marginalized groups
Feeling