Supplementary Materials

1. Distribution of responses for Bias Case Sets

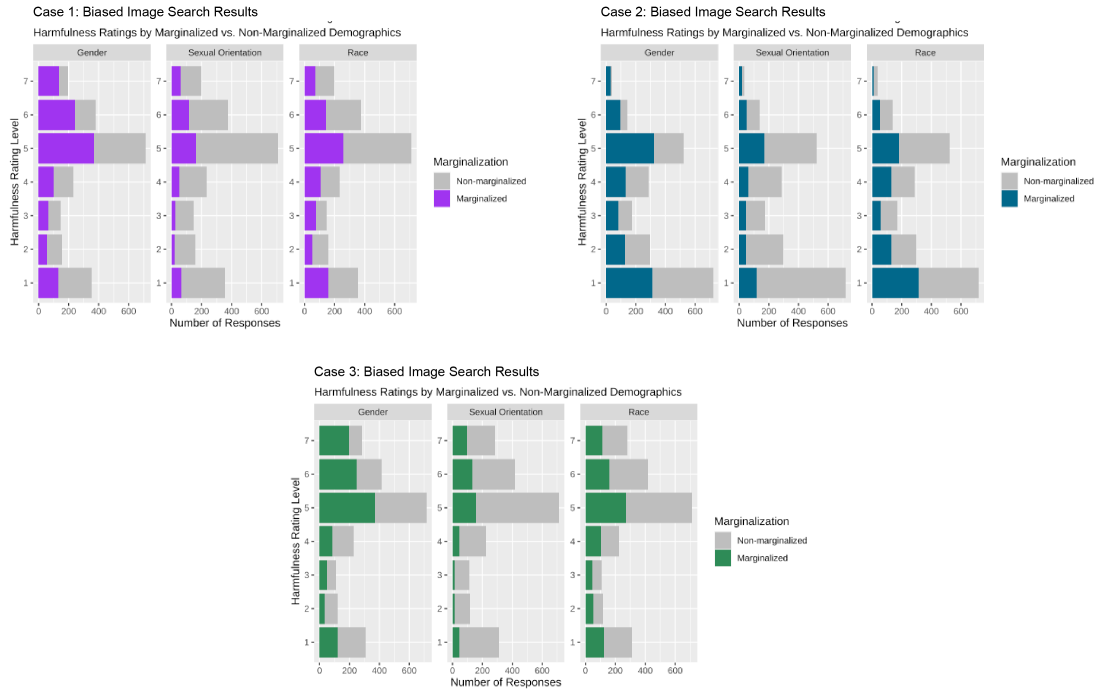

Fig. 1: Distribution of responses for all three Bias Case Sets: For Case 1, Case 2, and Case 3, above, there are three separate graphs shown side-by-side for each case; the first graph for each case shows the scores given by respondents who belonged to marginalized vs. non-marginalized gender demographics; in the second graph, marginalized vs. non-marginalized sexual orientation demographics; and in the third graph, marginalized vs. non-marginalized race demographics. On the X-axis, the three graphs for each case show the number of survey respondents who gave the image search results for the given bias case a Harmfulness Rating score of 1 - 7, where 1 = 'Totally Unharmful' and 7 = 'Totally Harmful.' The Y-axis is bias the 'harmfulness' rating level the survey respondents gave the image search results for the given case, and ranges from 1 = 'Totally unharmful to 7 = 'Totally harmful. In the Case 1 graphs, in purple, each graph shows the scores given by survey respondents belonging to marginalized demographics; in gray, each graph shows the scores given by survey respondents who do not belong to marginalized demographics. In the Case 2 graphs, in teal, each graph shows the scores given by survey respondents belonging to marginalized demographics; in gray, each graph shows the scores given by survey respondents who do not belong to marginalized demographics. In the Case 3 graphs, in green, each graph shows the scores given by survey respondents belonging to marginalized demographics; in gray, each graph shows the scores given by survey respondents who do not belong to marginalized demographics. Each case shows three charts, comparing ratings of non-marginalized vs marginalized groups with respect to gender, sexual orientation, and race.

2. Survey Questions

The following questions are for one of the experiment cases. We iterated all the four case sets in a randomized order and randomly selected one case out of each case category to each participant.

A. Questions about Detection of Harmful Bias and Discrimination in the case

These are the search results for "[INSERT SEARCH QUERY]" from an image search system. It was

captured from a [INSERT INTERFACE (COMPUTER / PHONE)] screen. Please look at the results carefully and answer all the questions below.
[INSERT IMAGE]
Q1: Please select the statement that indicates how much or how little harmful or unharmful bias and discrimination you think the algorithmic system is generating with this image. There are no right or wrong answers. (Choices: "Totally harmful", "Very harmful", "Somewhat harmful", "Neither harmful nor unharmful", "Somewhat unharmful", "Very unharmful", "Totally unharmful")
Q2.1: Why do you believe this algorithmic system is generating [INSERT ANSWER FROM Q1] bias and discrimination? There are no right or wrong answers. Please explain your choice in a sentence starting with "I ... because . . . ". [Text entry]
(If the answer to Q1 is "Totally harmful", "Very harmful", "Somewhat harmful", or "Neither harmful nor unharmful")
Q2.2: What kinds of harmful bias and discrimination do you perceive and Who do you think might be harmed in this case? [Text entry]

B. Questions about Everyday Discrimination Experience

Q9.1: In your day-to-day life how often have any of the following things happened to you? ("You are treated with less courtesy or respect than other people": "Almost everyday", "At least once a week", "A few times a month", "A few times a year", "Less than once a year", "Never"; "You receive poorer service than other people at restaurants or stores": "Almost everyday", "At least once a week", "A few times a month", "A few times a year", "Less than once a year", "Never"; "People act as if they think you are not smart": "Almost everyday", "At least once a week", "A few times a month", "A few times a year", "Less than once a year", "Never"; "People act as if they are afraid of you": "Almost everyday", "At least once a week", "A few times a month", "A few times a year", "Less than once a year", "Never"; "You are threatened or harassed": "Almost everyday", "At least once a week", "A few times a month", "A few times a year", "Less than once a year", "Never")
(If the answer to Q9.1 is "A few times a year" or more frequently to at least one proxy.)
Q9.2: What do you think is the main reason for these experiences? (Select all that apply) (Choices: "Your Ancestry or National Origins", "Your Gender", "Your Race", "Your Age", "Your Disability", "Your Religion", "Your Height", "Your Weight", "Some other Aspect of Your Physical Appearance", "Your Sexual Orientation", "Your Education", "Income Level", Other: [text entry]")
Q10: Please select from the following that which most accurately represents the answer to the following question: Mary has 5 apples. If Josh gives her 3 oranges and 4 apples, how many apples will she have? (Choices: "7", "9", "8", "6")

C. Questions about Socio-Technical Knowledge

Q11: Do you have close friends / family members who are members of racial minority (including all non-white ethnicity) groups? (Choices: "Yes", "No", "I don't know")
Q12: Do you have close friends / family members who are members of gender identity minority (including all non-male) groups? (Choices: "Yes", "No", "I don't know")
Q13: Do you have close friends / family members who are members of sexual orientation minority (including all non-heterosexual) groups? (Choices: "Yes", "No", "I don't know")
Q14: How frequently would you say you encounter news/media about bias in society? (Choices: "Daily", "Weekly", "Monthly", "A few times a year", "Never", "I don't know")
Q15: How frequently would you say you encounter news/media about bias in algorithms or automated systems used by people or organizations? (Choices: "Daily", "Weekly", "Monthly", "A few times a year", "Never", "I don't know")
Q16: How familiar are you with the use of algorithmic systems (e.g., email spam filter, social media, amazon recommendations, etc)? (Choices: "Extremely familiar", "Moderately familiar", "Neither familiar nor not familiar", "Moderately not familiar", "Not familiar at all")
Q17: How aware do you feel you are with issues related to societal biases (e.g., racial biases, gender biases)? (Choices: "Very aware ", "Somewhat aware", "Neither aware nor not aware", "Not very aware", "Not at all aware")

D. Questions about Demographics

Q18: What is your age? (Choices: "18-24", "25-34", "35-44", "45-54", "55-64", "65 or above")

Q19: What is your Race/Ethnicity? Please select all that apply. (Choices: "White", "Black or African American", "American Indian or Alaska Native", "Asian", "Native Hawaiian or Other Pacific Islander", "Hispanic", "Two or More Races", "Other Race [Text Entry]")

Q20: What is your Gender Identity? (Choices: "Male", "Female", "Trans Male/Trans Man", "Trans Female/Trans Woman", "Genderqueer/Gender Non Conforming", "Different Identity", "Rather not say")

Q21: Does your current gender differ from the one you were assigned at birth? (Choices: "Yes", "No", "Rather not say")

Q22: What is your Sexual Orientation? (Choices: "Heterosexual", "Homosexual", "Bisexual", "Asexual", "Other [Text Entry]")

Q23: What is the highest degree or level of school you have completed (if you're currently enrolled in school, please indicate the highest degree you have received). (Choices: "Less than a high school diploma", "High School degree or equivalent", "Some college, no degree", "Associate degree", "Bachelor's degree", "Master's degree", "Professional degree", "Doctorate")

Q24: What is the zip code of your current residence? [Text Entry]

E. Debriefing Session

We previously informed you the purpose of the study is asking for your opinion about algorithmic systems. The goal of our research is to find out factors influencing users' bias detection in algorithm systems. The following case (we showed you previously in this survey) has been identified as having harmful algorithmic bias and discrimination against [ADDRESS THE IMPACTED DEMOGRAPHICS].

Q25: Before taking this survey, have you seen or heard about this image search result and its harmful impacts before? [ITERATIVELY INSERT THE IMAGE PREVIOUSLY SHOWN IN THE SURVEY] (Choices: "Yes", "No", "I'm not sure")

3. Participant Demographics

Table 1: Demographics statistics of our survey sample ($N = 2,179$).

Demographic Characteristics	*N*	%
Gender		
Female	969	44.47%
Male	1132	51.95%
Genderqueer	41	1.88%
Prefer not to disclose	11	0.50%
Transgender	21	0.96%
Different Identity	5	0.23%
Sexual Orientation		
Heterosexual	1677	76.96%
Gay, Lesbian, Bisexual, or Asexual	454	20.84%
Other	48	2.20%
Race		
American Indian or Alaska Native	6	0.28%
Asian	226	10.37%
Black or African American	283	12.99%
Hispanic	142	6.52%
Native Hawaiian or Pacific Islander	3	0.14%
White	1314	60.30%
Other Race	10	0.46%
Two or More Races	195	8.94%
Age		
18–24	309	14.18%
25–34	702	32.22%
35–44	524	24.05%
45–54	313	14.36%
55–64	206	9.45%
65+	125	5.74%
Education		
Less than a high school diploma	23	1.06%
High school degree or equivalent	275	12.62%
Some college, no degree	495	22.72%
Associate Degree	222	10.19%
Bachelor's degree	821	37.68%
Master's degree	258	11.84%
Professional degree	43	1.97%
Doctorate	42	1.93%

[1] For gender, our source data from U.S. Census only have female and male percentage.
[2] For Race, our source data from U.S. Census doesn't have Middle Eastern as a separate race.

4. Diagram of the Overview of the Study Procedure

5. Filtering Low-Quality Responses using LLM

When we further inspected our data, we found that a number of participants' responses asking for their rationale behind their ratings (Q2) were contradictory to their ratings (Q1). For example, one participant responded *"I believe this is harmful"* while rating the case as *"Very unharmful"*. To filter out likely noise in our data, we used a Large Language Model (LLM)[1] to label participants' responses in the open-ended questions, following best practices in prior work [5]. All responses underwent a de-identification process in compliance with our IRB. To validate the use of LLM, we sampled 100 participants from the original 2,201, yielding 400 responses to Q2. Two human coders labeled these as *"Yes"* (the participant felt the case contains harmful bias) or *"No"*. After achieving an intercoder reliability of 0.920, exceeding the recommended 0.7 threshold [6], we used the 400 hand-coded responses as ground truth, and divided this

[1] https://openai.com/product/gpt-4

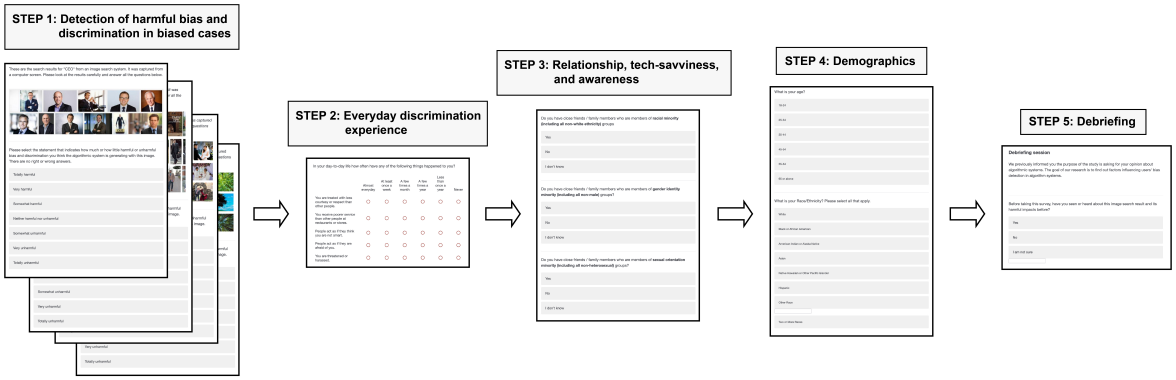

Fig. 2: **Overview of the study procedure.** In each survey, participants were asked to complete 5 steps: (1) Rate four different cases of image search results in random order, in terms of how harmful the bias and discrimination is and offer rationale in open-ended questions; (2) Answer questions about their everyday discrimination experiences; (3) Answer questions regarding their social and technical knowledge; (4) Answer demographic questions; (5) Complete the debrief session.

set into a validation set (200 instances) and a test set (200 instances). After iterative prompt engineering, we settled on a final prompt that had the highest accuracy level of 98.7% and a kappa value of 0.974 on the validation set, which is: *"For the given statement about an algorithmic system, identify whether it indicates the user feels the system contains harmful biases. Answer the question with "Yes", "No", "Uncertain"."* We then used the LLM to analyze the remaining 8,804 responses from all 2,201 participants. A human coder reviewed and hand coded all responses labeled as *"Uncertain"*. We identified 22 participants that had inconsistencies between their answers to Q1 and Q2. Our final data set contains 2,179 participants after excluding 22 participants.

6. Cases We Use in the Survey Experiment

Fig. 3: **Bias Case Set 1 - 1:** The image search results of "professor style" on Google Images were identified as having gender bias because of only displaying images of male [4]. We used the screenshot of the image search results presented in the literature, which was captured from a computer screen [4], and cropped out all the text in the screenshot to use in the survey experiment.

Fig. 4: **Bias Case Set 1 - 2:** The image search results of "doctor" on Google Images were identified as having gender bias because of only showing white male doctors [4]. We used the screenshot of the image search results presented in the literature, which was captured from a computer screen [4], and cropped out all the text in the screenshot to use in the survey experiment.

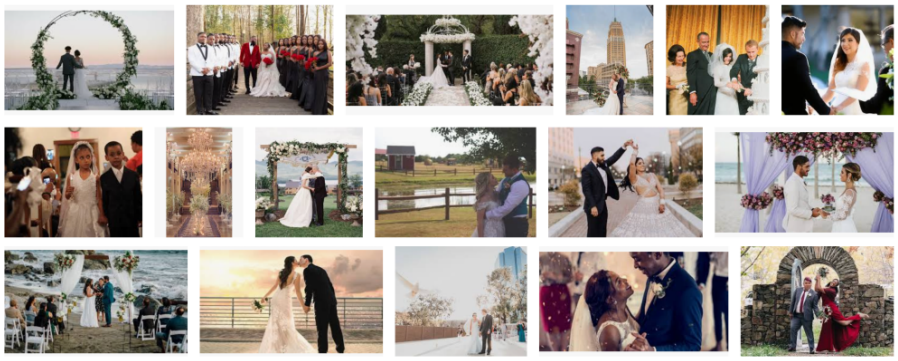

Fig. 5: **Bias Case Set 2 - 1:** The image search results of "weddings" on Google Images were identified as having sexual orientation bias because of only portraying heterosexual couples [2]. One researcher recaptured a screenshot of the first page of Google image search results for "weddings" from a computer interface and cropped out all text, https://www.google.com/search?sca_esv=562916950&sxsrf=AB5stBiKKLCxOFQn5F4dw4Kz0KzS07bWCg:1693964114308&q=weddings&tbm=isch&source=lnms&sa=X&ved=2ahUKEwjfja6K7JSBAxXsGFkFHTO7BCkQ0pQJegQIDRAB&biw=1728&bih=994&dpr=2, Accessed: 2022-10-22.

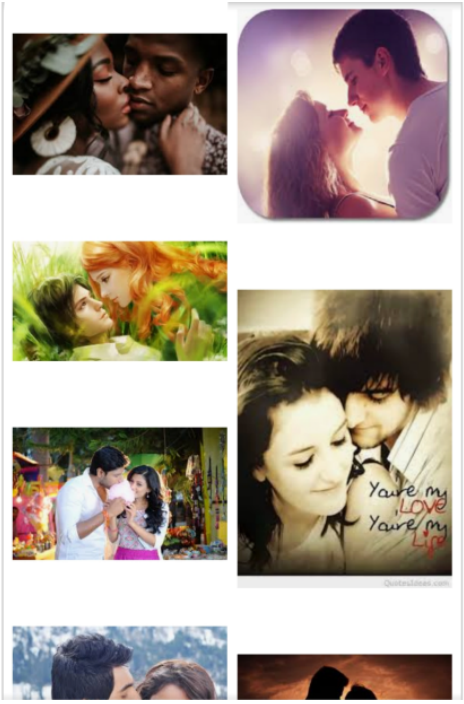

Fig. 6: **Bias Case Set 2 - 2:** The image search results of "romantic couples" on Google Images were identified as having sexual orientation bias because of only portraying heterosexual couples [2]. We used the screenshot of the image search results presented in the literature, which was captured from a phone screen [2], and cropped out all the text in the screenshot to use in the survey experiment.

Fig. 7: **Bias Case Set 3 - 1:** The image search results of "babies" by Microsoft search engine Bing were identified as having racial bias because of displaying white babies [3]. We used the screenshot of the image search results presented in the literature, which was captured from a phone screen [3], and cropped out all the text in the screenshot to use in the survey experiment.

Fig. 8: **Bias Case Set 3 - 2:** The image search results of "CEO" on Google Images were identified as having racial bias because of only showing images of white male [1]. We used the screenshot of the image search results presented in the literature, which was captured from a phone screen [1], and cropped out all the text in the screenshot to use in the survey experiment.

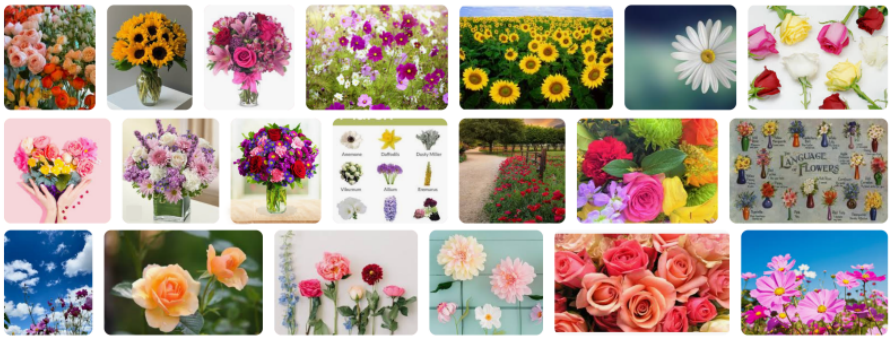

Fig. 9: **Neutral Case Set - 1** The image search results of "flower" on Google Images were considered as a neutral case because of not involving human subject. One researcher captured a screenshot of the first page of Google image search results for"flowe" from a computer interface and cropped out all text, https://www.google.com/search?q=flower&sca_esv=562916950&tbm=isch&source=hp&biw=1571&bih=874, Accessed: 2022-10-19.

Fig. 10: **Neutral Case Set - 2** The image search results of "tree" on Google Images were considered as a neutral case because of not involving human subject. One researcher captured a screenshot of the first page of Google image search results for"flowe" from a computer interface and cropped out all text, https://www.google.com/search?q=tree&sca_esv=562916950&tbm=isch&source=hp&biw=1571&bih=874, Accessed: 2022-10-19.

References

1. Brekke, K. Google image search has a gender bias problem. 2015. Accessed: 2023-09-05.

2. Devos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems* (2022).
3. Kleiman, Z. Artificial intelligence: How to avoid racist algorithms. 2017.
4. NOBLE, S. U. Algorithms of oppression: How search engines reinforce racism. NYU Press, 2018.
5. Xiao, Z., Yuan, X., Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 75–78, 2023.
6. McDonald, N., Schoenebeck, S., & Forte, A. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019, ACM New York, NY, USA.