Data Analysis Report:

Survey Response of Algorithmic Bias by WeAudit

Data Visualization, Insights and Hypothesis

Team Nasa GenAI Vanguard

Understanding the Data

1. **What information is useful?**
    1. Detailed user demographic details like age, gender, location,
    2. Information on user familiarity with societal and algorithmic bias, prior exposure to bias in text-to-image products
    3. User opinions on level of harm for curated examples on gender, racial, neutral biases
2. **What information is not useful?**
    1. Too much metadata information related to the survey
    2. RecaptchaScore, RelevantIDFraudScore useful for data cleaning and filtering but aren’t directly useful
    3. Very specific examples for understanding how we can leverage user auditing for GenAI bias mitigation since only one use-case is explored

Understanding the Data (cont.)

1. **What research questions can be answered via this dataset?**
   a. How do demographic factors influence perceptions of bias and discrimination in algorithmic systems?
   b. Is there a correlation between geographic location and sensitivity to algorithmic bias?
   c. What themes emerge from textual responses about why individuals find certain algorithmic outputs harmful or unharmful?
2. **Are there any limitations of this data?**
   a. This data exhibits social desirability bias as it self-reported by users in the survey
   b. This data is a narrow exploration of understanding user sentiments towards text-to-image generative models which is only a small part of GenAI systems

Hypotheses and Questions

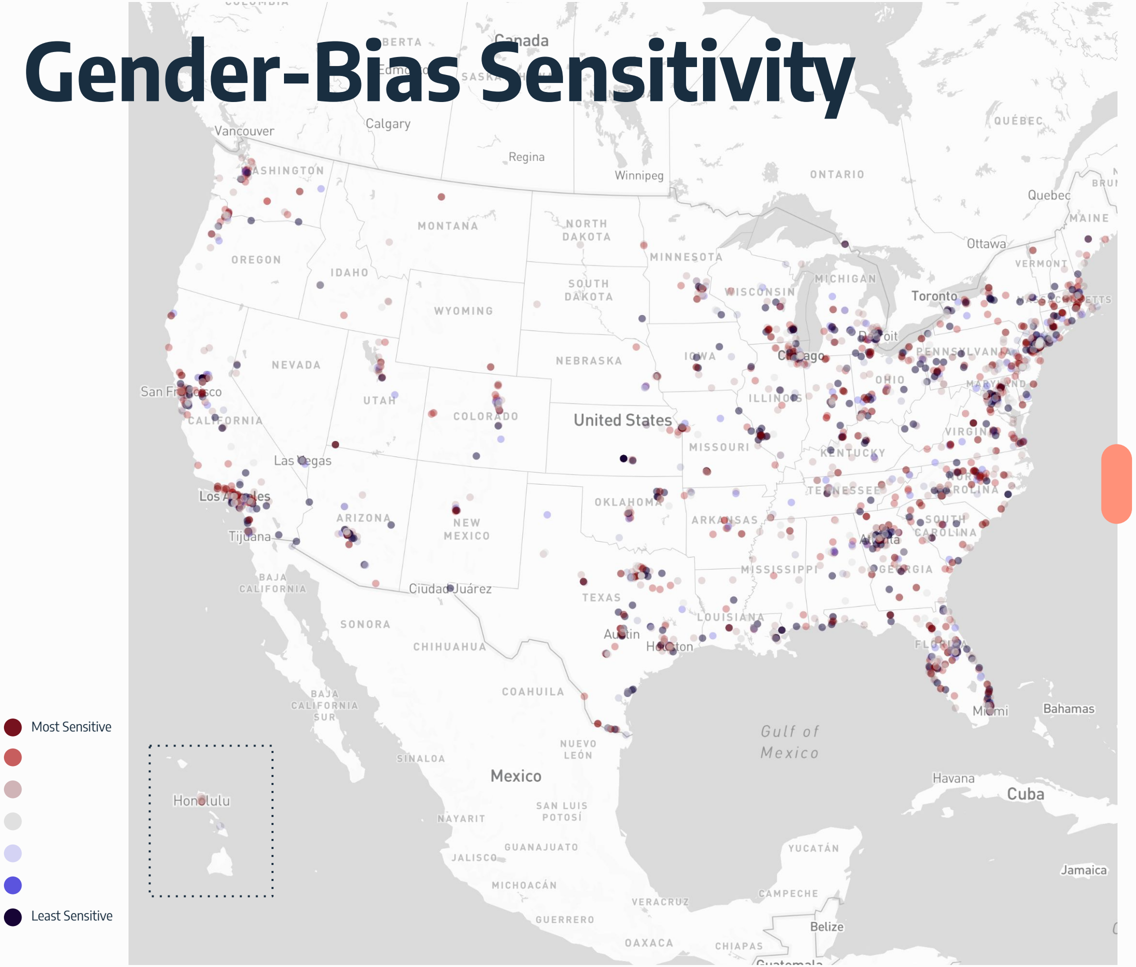
1. Is there a correlation between geographic location and sensitivity to algorithmic bias?
2. How do demographic factors of users influence perceptions of bias and discrimination in algorithmic systems?
3. What themes emerge from textual responses about why individuals find certain algorithmic outputs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

01

Visualization 1

Is there a correlation between geographic location and sensitivity to algorithmic bias?

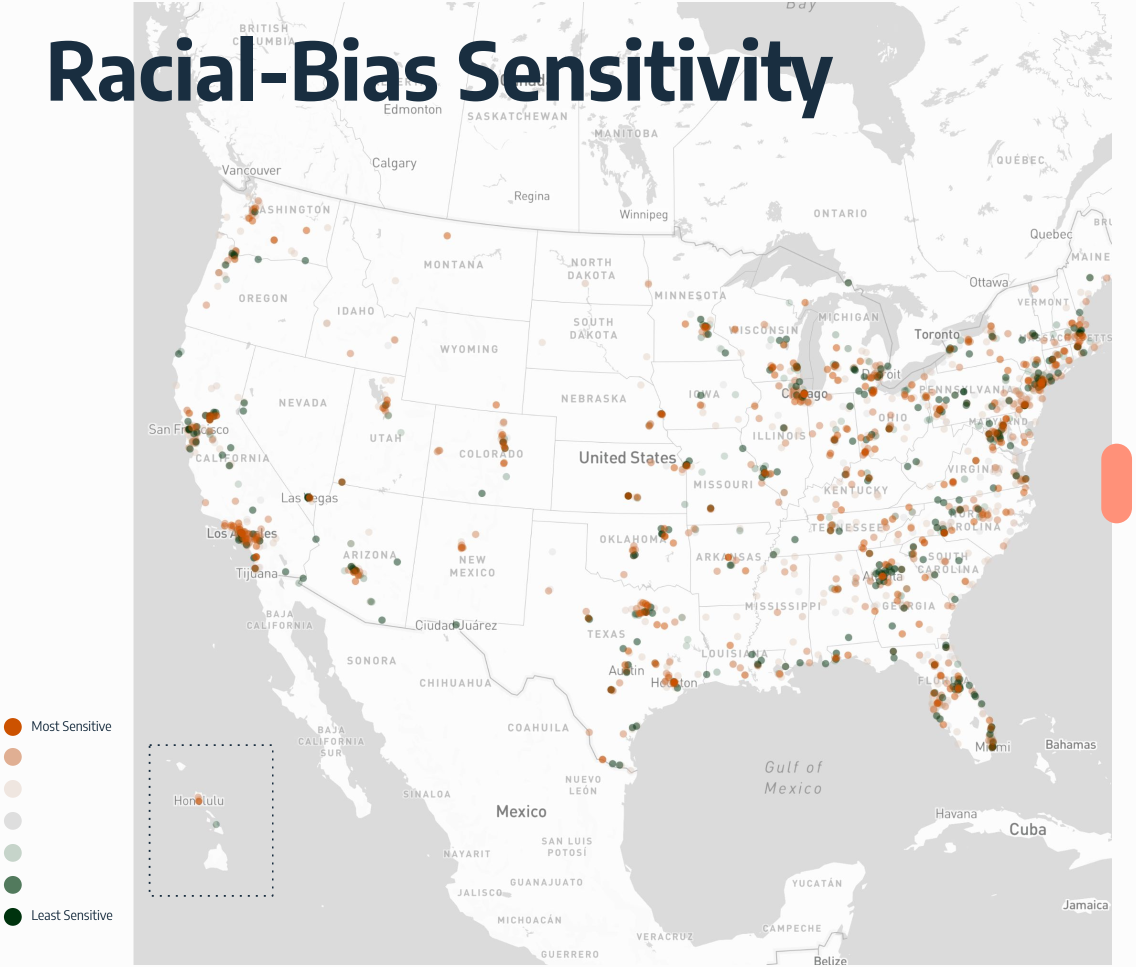
Gender-Bias Sensitivity



In this visualization, 1286 participants perceived the content as harmful to different extents, while 663 consider the content to be unharmful.
(N=2180)

The data indicates that the participants' sensitivity to identify gender bias is not correlated to specific geographic region.

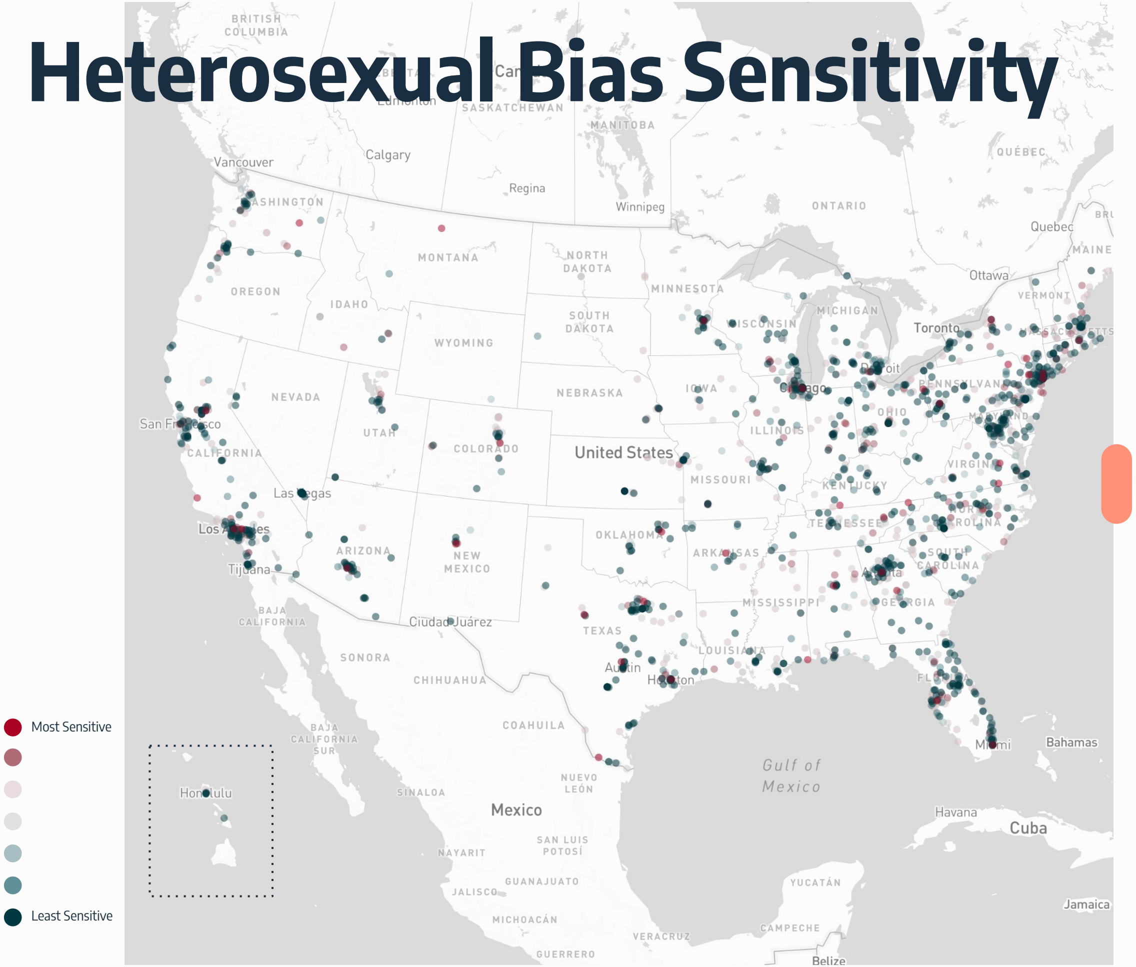
Racial-Bias Sensitivity



From a total of 2180 participants, 1415 participants perceived the content that is racial biased as harmful to different extents, while 540 consider the content to be harmful.

Although we found that certain states had more pronounced feedback for racial bias content, the overall sensitivity was also independent of geographic location.

Heterosexual Bias Sensitivity



From a total of 2180 participants, 702 participants perceived the content as harmful to different extents, (Only 78 of them considered such bias to be TOTALLY HARMFUL within), 1191 consider the content to be harmful, and 287 participants remain neutral.

The data indicates that the participants' sensitivity towards Heterosexual bias is not correlated to geographic region.

02

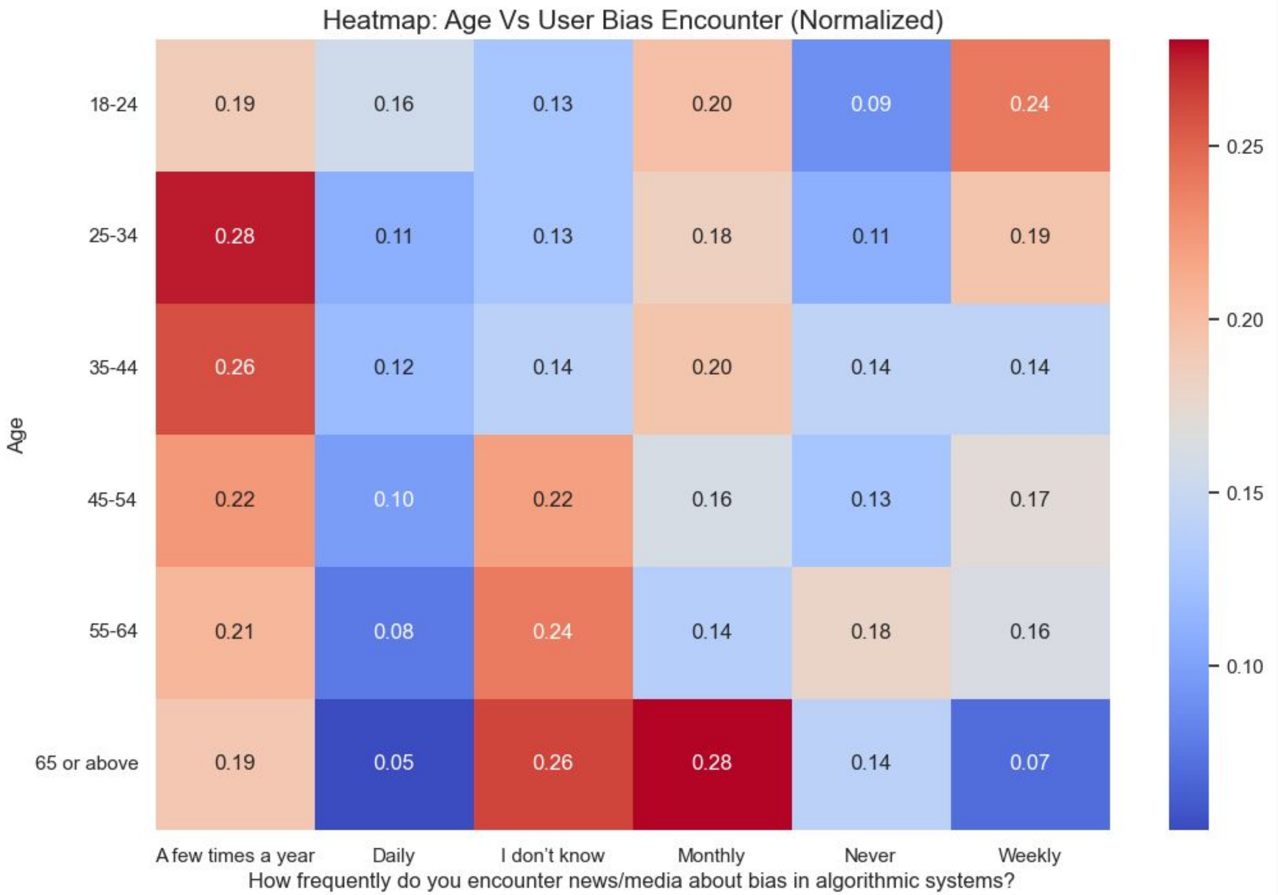
Visualization 2

How do demographic factors of users influence perceptions of bias and discrimination in algorithmic systems?

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.

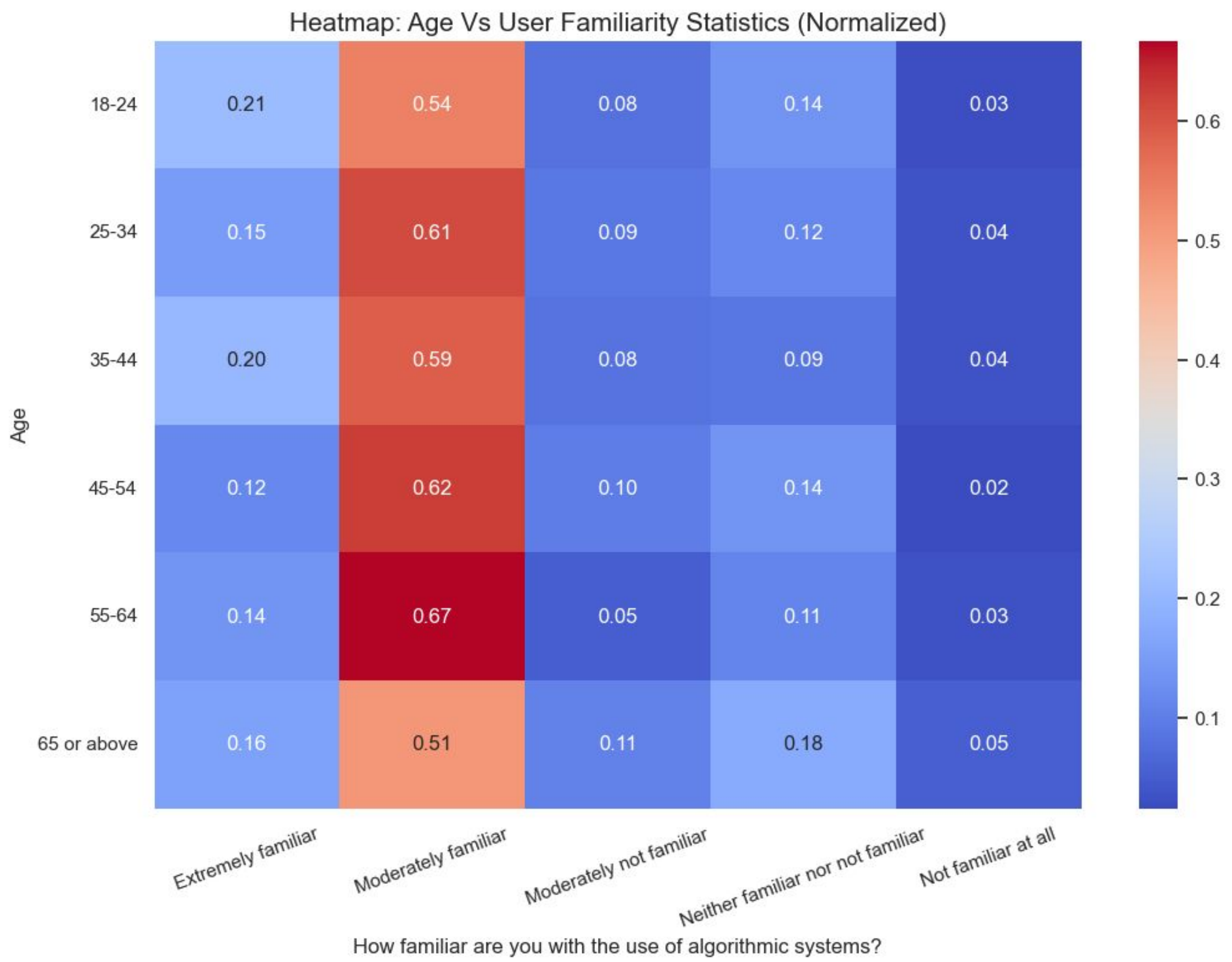
All age groups encounter such news a couple of times yearly. However, it is a rare occurrence daily.

PROJECT: Our user-auditing frequency should be adjusted to account for user-overload for testing when they encounter news articles and try to test systems.



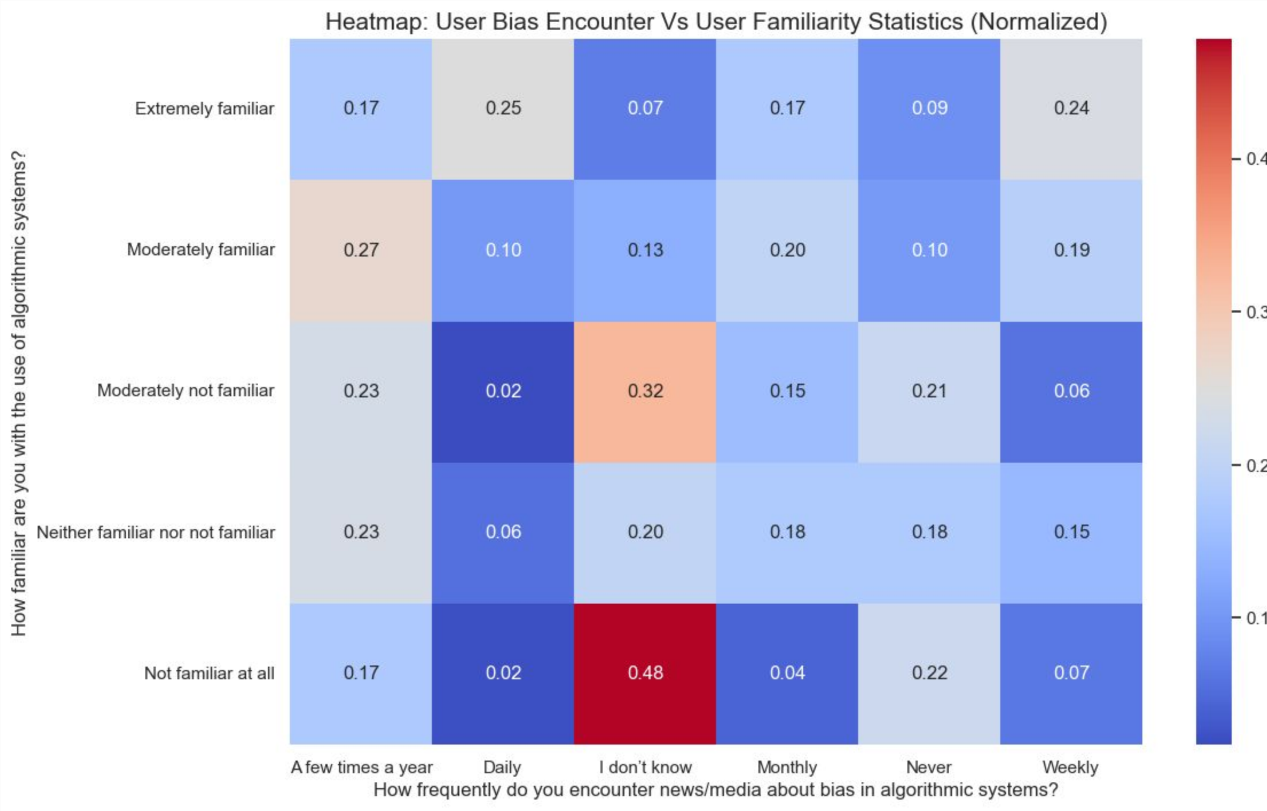
Age	A few times a year	Daily	I don't know	Monthly	Never	Weekly
18-24	0.19	0.16	0.13	0.20	0.09	0.24
25-34	0.28	0.11	0.13	0.18	0.11	0.19
35-44	0.26	0.12	0.14	0.20	0.14	0.14
45-54	0.22	0.10	0.22	0.16	0.13	0.17
55-64	0.21	0.08	0.24	0.14	0.18	0.16
65 or above	0.19	0.05	0.26	0.28	0.14	0.07

Almost all age groups are now familiar with using algorithmic systems, including, 55+. This inference may need further exploration and triangulation as this could be due to desirability bias due to self-reporting or sampling issues.

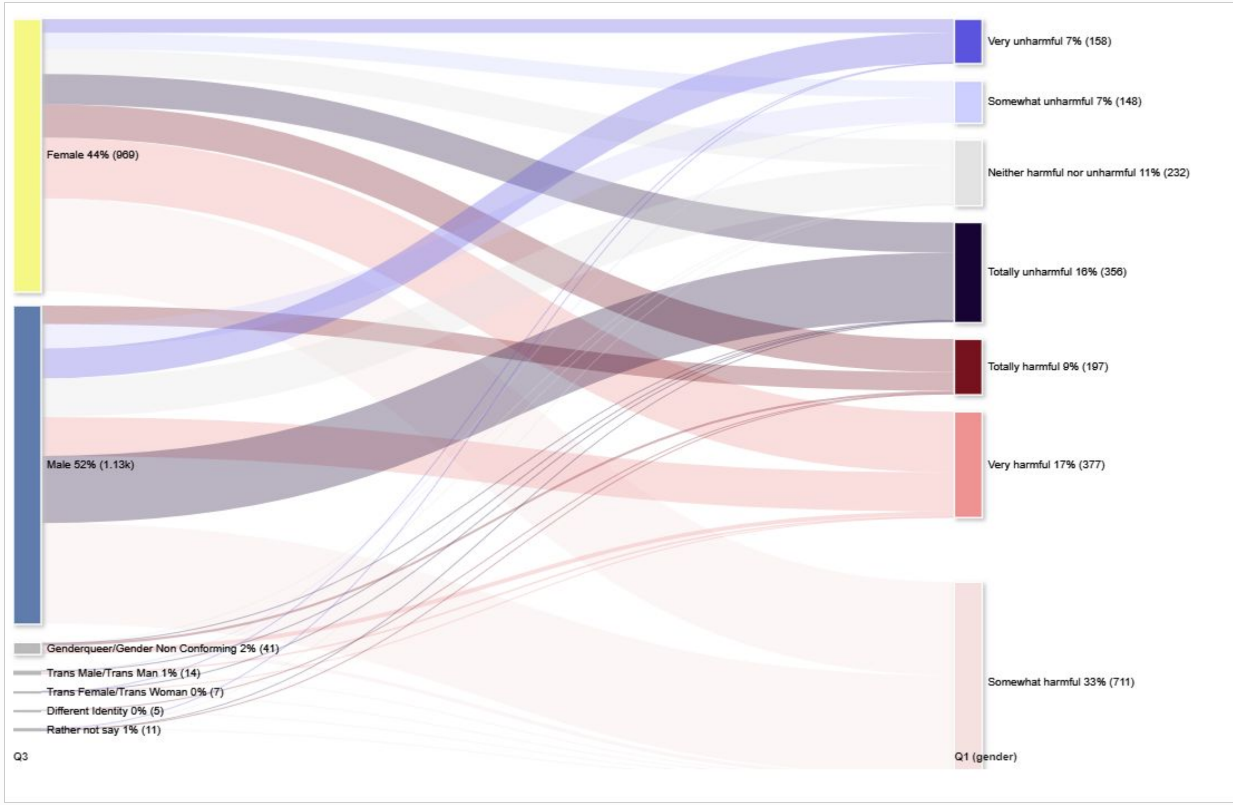


Usage of algorithmic systems and the encounter frequency of algorithmic biases seems weakly correlated despite all age groups being familiar with them.

This inference also needs to be checked.


Heatmap: User Bias Encounter Vs User Familiarity Statistics (Normalized)

We also discovered that gender may influence how participants respond to these questions. Generally, women are more attuned to gender bias.

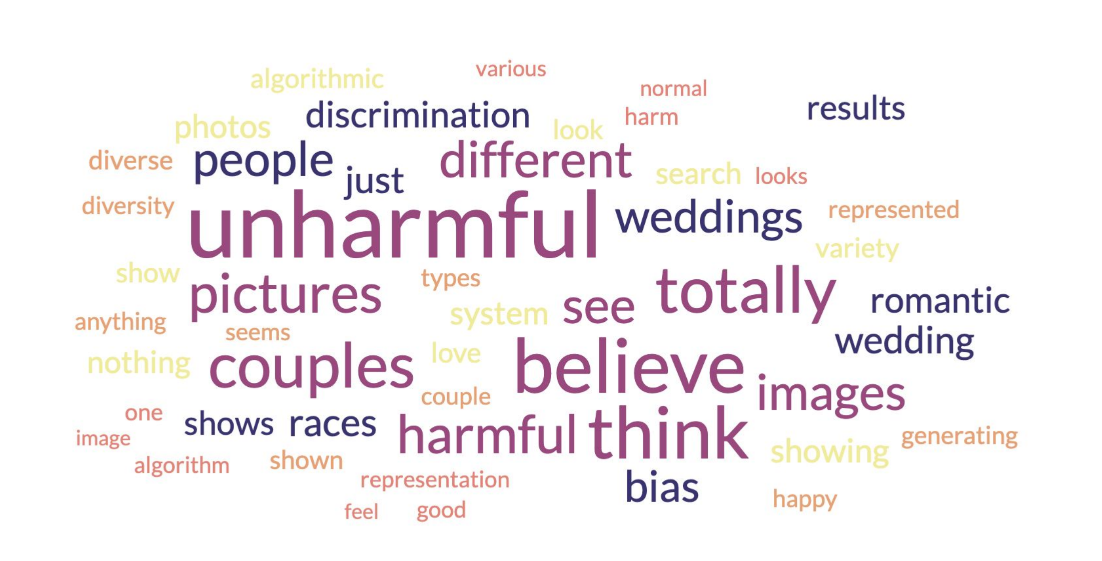


03

Visualization 3

What themes emerge from textual responses about why individuals find certain algorithmic outputs harmful or unharmful?

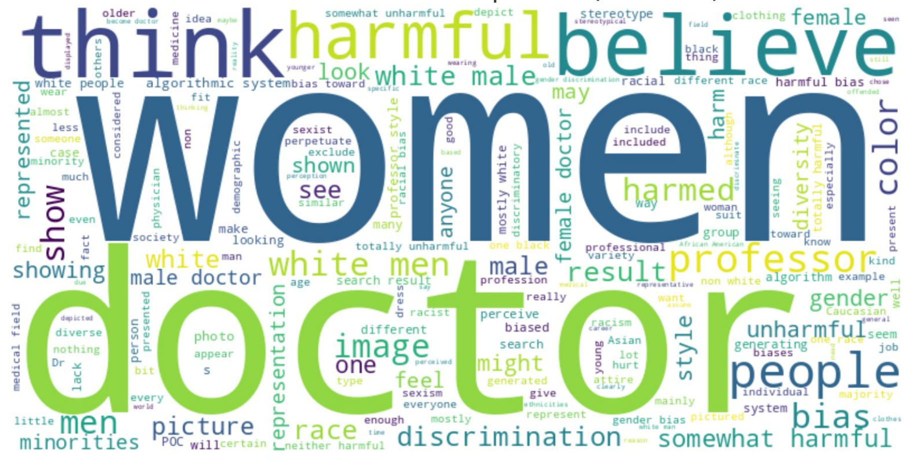
Example: Opinions on “Totally Unharmful”



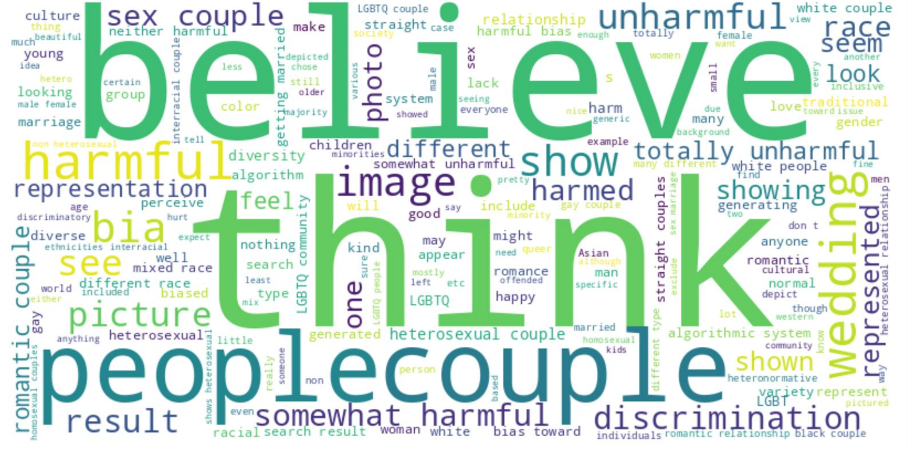
This word cloud was formed from the responses for the prompt “Why do you believe the bias and discrimination this algorithmic system is generating is totally unharmful?” (Column AR).

This word cloud was formed from the responses of the 721 people (out of 2197) that answered the previous prompt relating to heterosexual bias as “totally unharmful”.

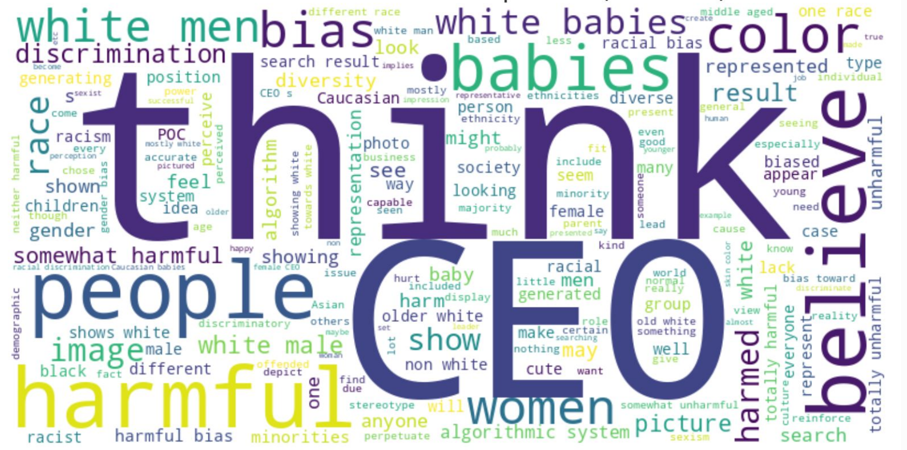
Word Cloud of Text Responses (answer0)



Word Cloud of Textual Responses (answer1)



Word Cloud of Textual Responses (answer2)



Individual word clouds were generated for the responses by the users by combining information from multiple columns.

- answer0: gender bias
- answer1: sexuality bias
- answer2: racial bias
- answer3: neutral (not included)

Topic Modelling

Gender Bias

```
Topic 1:  white women men

Topic 2:  doctors unharmful just

Topic 3:  unharmful style professor

Topic 4:  color people person

Topic 5:  white people black

Topic 6:  women men biased
```

Insights from the Data

1. There does not appear to be any visible correlation between location and sensitivity on bias
    a. The average sensitivity appears to be similar across all three maps—there are no areas which seem to have more or less sensitivity than others
    b. It is worth noting that parts of the West and Midwest are underrepresented
2. People appear to be much more sensitive to racial and gender biases than they are towards bias on sexual orientation
    a. There doesn’t seem to be an immediate connection between the biases, either: being strongly sensitive on one bias does not indicate the same on the others

Assignment Post-mortem

- Since there were about 2181 rows and 110 columns of data to work with, we had to spend significant time cleaning up the data using Python scripting (pandas):
  - Filtering out the columns we needed for data analysis brought down the our column count to 36
  - We also combined the text from multiple columns like the ‘`Why do you believe the bias and discrimination this algori…`’into one single column for textual data analysis
  - Rows with NaN values were dropped and the data was evaluated for duplicates
- We conducted our analysis for visualizations for structured and unstructured data primarily in python. With more time, we would have liked to try Tableau or other software but no one had prior experience with them.

Resources

Mapbox Studio
https://studio.mapbox.com/

Free Word Cloud Generator
https://www.freewordcloudgenerator.com/generatewordcloud

TOPIC MODELING

https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a