CS 6603: AI, Ethics, and Society
Final Exam

Cleo Zhang

yzhang3761@gatech.edu

1 TASK 1

- Public Artifact:
    - Title – "The viral AI avatar app Lensa undressed me—without my consent."
    - Released – December 12, 2022
    - Link - https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent
- Application/Scenario/Domain of Misuse: Avatar Generation (**Predictive Algorithm**)
- Regulated Domain/Protected Class: Sex/Race
- Evidence:
    - Dataset like the one used to build Stable Diffusion: https://arxiv.org/abs/2110.01963
    - https://paperswithcode.com/dataset/laion-5b

2 TASK 2

In the article, the Author highlights the gender bias against women using the app Lensa, where female avatars are sexualized, objectified, and stereotyped. At the same time, Lensa generates realistic yet flattering avatars for male users. Furthermore, the Author's avatar became more real when her pictures went through male content filters. In addition to the bias against Females, the Author also discussed the discrimination against Asians. The Author, who has Asian heritage, received significantly more sexualized and pornographic avatars than their

white female colleagues. The AI model seemed to have picked up on the Author's Asian heritage and generated images of generic Asian women modelled on anime or video-game characters, often sexualized. The Author also notes that searching the training dataset used by the Stable Diffusion for keywords "Asian" brought back almost exclusively porn rather than neutral pictures.

The article also cites research by Aylin Caliskan, who studies biases and representation in AI systems, that shows AI training data is filled with racist stereotypes and explicit images of rape, which leads to AI models that sexualize women, especially women with identities that have been historically disadvantaged, such as minority races like Asian. In this application, **the existing bias and discrimination in the training dataset are reinforced** and applied to humans through AI model training. This is a typical example of biased sampling.

3 TASK 3

Table 1 — Privileged /unprivileged groups.

Protected Class	Privileged	Unprivileged
Sex	*Male*	Female
Race	*White*	Asian

- Misleading graphs: the abundance of images on the internet that depict women in a state of undress or with minimal clothing, along with pictures that **promote sexist or racist stereotypes**.
- Sources of Data Bias: the utilization of Stable Diffusion, which is an open-source AI model used by Lensa for generating avatars based on textual prompts, which is trained by LAION-5B, an image data set compiled by scraping from the internet.
- Sources of Sampling Bias: the training dataset does not fully reflect the actual population and is filled with historical stereotypes against certain groups of people. This is a **biased sampling**.

- Sampling Methods Used to Collect Data: what the Author used is closest to **the stratified sampling** – as described in the article, data are sampled when the user is white, male, female or has Asian heritage.
- Correlations found in the data: In the article, the authors show that whether Lensa generates pornified avatars or not has correlations with the skin colour and gender of the person on the photo used by the user. **But Correlation does not mean Causation.**
- Outcome measures: Table 2 shows the **Frequency Distributions**

Table 2 – Frequency table of Author's avatar generation.

Sex	Images	Members	Count
Female	topless	2	16
	extremely skimpy clothes and overtly sexualized poses	1	14
	normal	0	70
Male	topless	2	0
	extremely skimpy clothes and overtly sexualized poses	1	0
	normal	0	100

Given the sampled size of 100, we can calculate **the Margin of Error** as below:

$$Margin\ of\ Error = \pm 1/\sqrt{n} = \pm 1/\sqrt{100} = \pm 10\%$$

Given that 30 (topless, extremely skimpy clothes and overtly sexualized poses) out of 100 avatars are biased, we can say that **at a 95% level of**

confidence, the true rate of biased images of an Asian female can be within 20% to 40%.

To compute the **Average** biased score of female and male groups, we can first label topless, extremely skimpy clothes and overtly sexualized poses as normal as 2, 1, 0 in Table 2.

$$Avg_{female} = (2 \times 16 + 1 \times 14 + 0 \times 70)/100 = 0.46$$

$$Avg_{male} = (2 \times 0 + 1 \times 0 + 0 \times 100)/100 = 0$$

- Bias & Fairness (or other) metrics used to identify differences in outcomes: **Disparate Impact** and **Statistical Parity Difference**, computed as below:

$$\text{Disparate Impact } = Avg_{female} \div Avg_{male} = N/A$$

$$\text{Statistical Parity Difference} = Avg_{female} - Avg_{male} = 0.46$$

Disparate Impact has no practical meaning in this case, so we primarily rely on the Statistical Parity Difference value to identify outcome differences.

4 TASK 4

By exploring the dataset mentioned in the Evidence above, I learned that the demographic in the training set is different from the actual population. For example, the number of men in the dataset is more than the number of women, and the mainstream race is much larger than the minority groups. More importantly, the sample collection is based on the Western cultural context, which can likely introduce unintentional bias and misunderstanding to the AI model in a subtle way. Therefore, we can first start with the dataset to mitigate such bias. A possible solution would be collecting samples in different cultural contexts, mixing them, and training AI models using them with the same weighting.

As a part of pre-processing, we may apply the reweighting algorithm to modify the training data features and labels. Given the calculation of the Statistical Parity Difference in the previous section, we may introduce an

additional factor of 0.46 to manipulate the record if the observation's gender is male so that the Avg_{male} increases and Statistical Parity Difference decreases.

After applying the above mitigation methods, I expect to have a more balanced and less biased database as an input to train the AI model. Therefore, I expect the outcomes from the model to be less biased, and most outcome measures or matrices fall into a reasonable range of acceptance.