CS 6603 AI, Ethics, and Society
Final Exam

Kamran Munawar
kmunawar3@gatech.edu

1 FIND A RECENT PUBLIC ARTIFACT ADDRESSING AI MISUSE

1.1 Title

Visual Stereotypes of Autism Spectrum in DALL-E, Stable Diffusion, SDXL, and Midjourney

1.2 Release date

23rd Jul 2024

1.3 Link to artifact

https://arxiv.org/abs/2407.16292

1.4 Application/scenario/domain of misuse

AI based text-to-image models,generative AI, visual stereotypes, biases, autism

1.5 Regulated domain/protected class impacted

Race,Sex,Age,Disability

1.6 Link to evidence

https://arxiv.org/abs/2407.16292

2 SUMMARY OF BIAS IN PUBLIC ARTIFACT

This study identifies significant **biases** in the AI-generated images related to autism. The primary **biases** include the over representation of **white**, **male**, and **child** depictions, which reinforce the misconception that autism predominantly affects these groups. This **demographic bias** results from the **training datasets**' lack of diversity, which fails to adequately represent **non-white**, **female**, and **adult** individuals. Additionally, the study highlights the prevalence of **stereotypical** themes such as the puzzle piece symbol, the color blue, and depictions of isolation and emotional blandness. These stereotypes perpetuate harmful and inaccurate perceptions of autism, portraying autistic individuals in a limited and

often negative light.

To address these biases, the study suggests **diversifying the training datasets** to include a wider range of **demographics and cultural contexts**. It also recommends implementing **bias detection and correction techniques**, such as **reweighting algorithms**, to ensure **balanced and fair representations** in the AI-generated images. By improving the inclusivity and accuracy of these representations, the study aims to promote a more nuanced and respectful understanding of autism.

3 DISCUSSION OF ARTIFACT/EVIDENCE METRICS DEMONSTRATING BIAS

3.1 Privileged/unprivileged groups

In the context of this study, privileged groups are those over represented or depicted more favorably in the AI-generated images. The majority of characters depicted were white, indicating a racial bias in the training data. There was also a noticeable over representation of males, which aligns with the stereotype that autism is more common in males. Additionally, a disproportionate number of images depicted children, reinforcing the stereotype that autism is primarily a childhood condition. Conversely, unprivileged groups are those underrepresented or depicted in a stereotypical or unfavorable light. Non-white individuals, females, and non-binary individuals were significantly underrepresented in the images. Moreover, the generated images predominantly depicted children, with far fewer representations of adults, indicating a lack of representation for autistic adults.

3.2 Source(s) of data bias

The study identifies several types of data bias stemming from the limitations and characteristics of the data sources and methods used. A key issue is the Western centric nature of the training data, which is predominantly sourced from Western countries, thereby affecting the cultural and geographical diversity of the representations and leading to a lack of representation of non-Western cultures and perspectives. Additionally, media sources, such as images from media and online platforms, often reflect societal biases and stereotypes. For instance, media portrayals of autism may emphasize stereotypes like social isolation or lack of emotional expression, which are then mirrored in the AI-generated images.

Convenience sampling, which involves using easily accessible data sources like stock photo libraries or online image repositories, can introduce further bias, as these sources may not provide a representative sample of the diverse autistic population. Historical and societal biases are also often embedded in the training data, with frequent use of symbols like the puzzle piece or the color blue in autism-related imagery reflecting longstanding stereotypes.

Moreover, the design and training process of the AI models themselves can introduce bias. Certain models may be more prone to replicating and amplifying biases present in the training data due to their specific algorithms and learning mechanisms. This bias can result in a lack of diversity and perpetuation of stereotypes in the generated images.

3.3 Source(s) of sampling bias

The study acknowledges several types of sample bias arising from the inherent limitations and characteristics of the data sources and methods used. Firstly, the limited scope of the 53 prompts, despite being carefully designed, might not fully encompass the entire spectrum of autism-related experiences and concepts, potentially resulting in a biased representation of autism in the generated images. Secondly, while the study includes multiple images generated for each prompt, variability in the outputs can still occur, introducing bias if certain types of images are more likely to be generated than others. Lastly, the qualitative nature of evaluating stereotypes means that some biases may be more subtle and harder to quantify, with the personal biases of the evaluators potentially influencing outcomes, even with a standardized coding framework.

3.4 Sampling methods used to collect data

The sampling methods used in this study involve a structured approach to prompt engineering, expert selection of prompts, comprehensive inclusion of all generated images, and randomized pilot coding sessions. These methods ensure a thorough and reliable analysis of the biases and stereotypes present in the AI-generated images related to autism.

3.5 Correlations found in the data

The study discusses correlations found in the data to a considerable extent. It revealed that the choice of AI model significantly influences the perpetuation of stereotypes in the generated images, with Stable Diffusion exhibiting the highest

degree of stereotyping and SDXL the lowest. This correlation underscores the impact of the AI model on stereotype propagation. Additionally, the study identified recurring stereotypes such as the puzzle piece symbol, the color blue, and themes of isolation, which frequently appeared across all models. This suggests a strong correlation between these specific themes and the training data used. Furthermore, the images predominantly depicted white, male, and young individuals, indicating that the training data over represents these demographics. This reflects societal biases in the portrayal of autism, as seen in the generated images.

3.6 Bias and Fairness (or other) metrics used to identify differences in outcomes

In this study, various bias and fairness metrics were employed to identify differences in outcomes across the AI models. The degree of stereotyping was measured on a scale of 0 to 10 for each image, quantifying the presence and intensity of stereotypical features. Higher scores indicated a greater degree of stereotyping. By comparing the mean degree of stereotyping across different models, the study aimed to identify which model exhibited the most bias.

Figure 1 in the study provides a visual comparison of the degree of stereotyping across different AI models, highlighting the variability and extent of bias present in the generated images.The higher mean degree of stereotyping for Stable Diffusion suggests that this model may require more targeted adjustments to reduce bias in its outputs.The lower mean degree for SDXL and DALLE indicates that it might have better inherent mechanisms or training data that reduce bias.

Cohen's kappa coefficient was used to measure inter-rater agreement on the presence of stereotypes, assessing the consistency and reliability of the raters' evaluations.

Category	Observed Agreement	Expected Agreement	Cohen's Kappa
Child	0.8178	0.5171	0.6227
White	0.8097	0.6036	0.5200
Male	0.7611	0.5187	0.5037

Table 1—INTER-RATER RELIABILITY ASSESSMENT

High observed agreement indicates that raters often agreed on their evaluations.

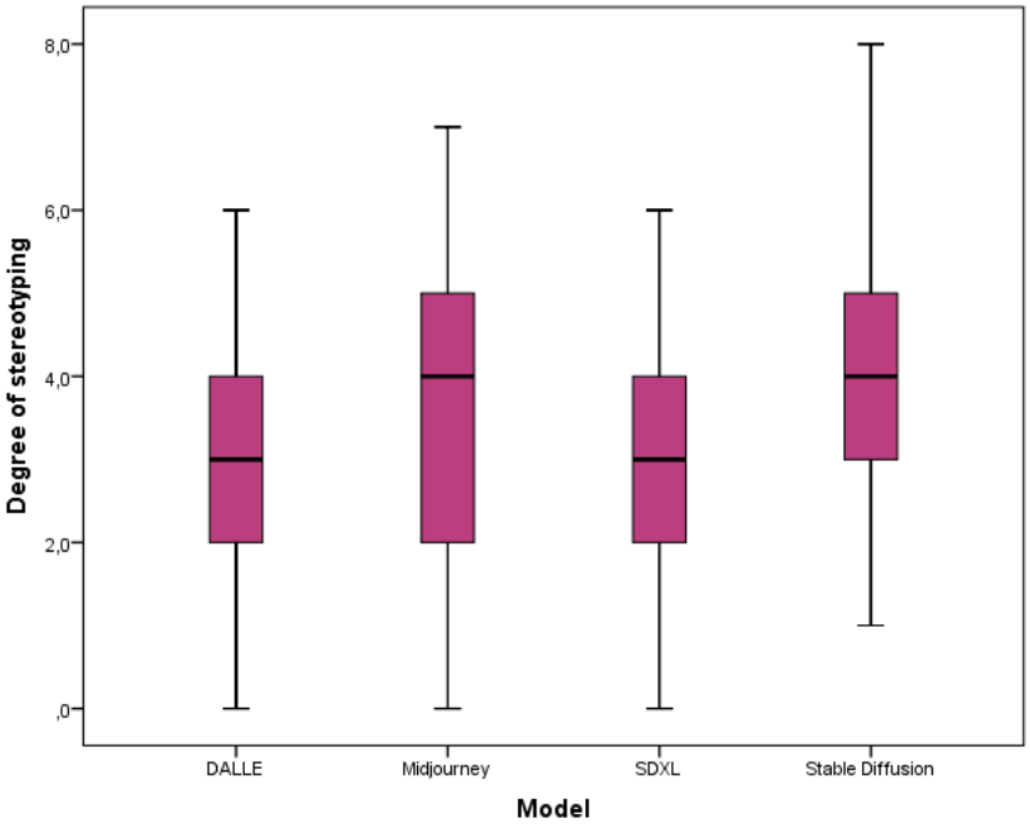

Figure 1—Degree of stereotyping

For instance, "White" had a high observed agreement of 0.8097. Expected agreement helps to understand what portion of the agreement is due to chance, with "White" having an expected agreement of 0.6036. Cohen's Kappa adjusts for chance agreement to provide a more accurate measure of reliability. Higher Kappa values indicate more reliable and consistent evaluations. For example, "White" had a Kappa of 0.5200, indicating good agreement beyond chance.

4 PROPOSE A METHOD FOR MITIGATING BIAS

If we can adjust the source of data bias, specifically the training data, we can mitigate bias in AI models. Given that the study did not provide details of the current proportions of each demographic group (e.g., race, gender, age) in the training dataset, we can use a general approach on how to apply the reweighting algorithm.

We need to calculate the reweighting factors required to modify the training data to address bias. Reweighting algorithms adjust the importance (weight) of different samples in the training dataset to create a more balanced representation of different groups. This process helps in mitigating bias and ensuring fairness in the model's outputs. The steps to achieve this are as follows:

1. Identify the privileged and unprivileged groups in your data.
2. Determine the current proportion of each group in the training dataset. For example:
   - Proportion of males,
   - Proportion of females,
   - Proportion of white individuals,
   - Proportion of non-white individuals, etc.
3. Define the desired proportion for each group to achieve fairness.
4. Compute the reweighting factors for each group by dividing the desired proportion by the current proportion. For example,

$$\text{Reweighting factor for males} = \frac{\text{Desired proportion of males}}{\text{Current proportion of males}}$$

5. Adjust the weights of the samples in the training dataset based on the reweighting factors. For instance, if a sample belongs to the male group, its weight is multiplied by the reweighting factor for males.

Figure 2 is showing Python code to implement above technique. Alternatively we can use Python libraries like "aif360.algorithms.preprocessing.Reweighing" to reduce biases during preprocessing step.

By applying these reweighting factors, you can modify the training data to achieve a more balanced and fair representation, thereby mitigating bias and unfairness in the model's outcomes.

```python
# Current proportions
P_m = data[data['gender'] == 'male'].shape[0] / data.shape[0]
P_f = data[data['gender'] == 'female'].shape[0] / data.shape[0]
P_w = data[data['race'] == 'white'].shape[0] / data.shape[0]
P_nw = data[data['race'] == 'non-white'].shape[0] / data.shape[0]

# Desired proportions
D_m = 0.50
D_f = 0.50
D_w = 0.50
D_nw = 0.50

# Reweighting factors
R_m = D_m / P_m
R_f = D_f / P_f
R_w = D_w / P_w
R_nw = D_nw / P_nw

# Apply reweighting factors
data.loc[data['gender'] == 'male', 'weight'] *= R_m
data.loc[data['gender'] == 'female', 'weight'] *= R_f
data.loc[data['race'] == 'white', 'weight'] *= R_w
data.loc[data['race'] == 'non-white', 'weight'] *= R_nw
```

Figure 2—reweighting using Python