Project Report On:

Artificial vs. Reality: Quantifying Gender Bias in Generative AI

Muhammed Fariz Palli Valappil

Table of Contents

1. Topic Choice & Motivation

Domain: Computer Vision & AI Ethics / Social Computing

1.1 Problem Statement

Generative AI models like Stable Diffusion are increasingly used to create content for marketing, media, and education. However, these models are trained on internet-scale datasets (LAION-5B) which contain historical biases. The core question of this project is: Does Generative AI simply reflect existing real-world social biases, or does it amplify them?

1.2 Hypothesis

We hypothesize that Generative AI will not only reflect real-world gender disparities but will exaggerate them, reducing complex human diversity into simplified "stereotypes" (e.g., presenting "Nurses" almost exclusively as female and "Doctors" as male), even when real-world data shows more diversity.

1.3 Key Parameters for Analysis

To quantify this, we defined a comparative metric: Female Representation Percentage (%).

We analyzed this metric across 8 distinct occupational roles:

- High-Status/STEM: Doctor, Scientist, Engineer (excluded due to low data), Pilot (excluded).
- Service/Care: Nurse, Teacher, Chef, Driver.
- Blue Collar/Creative: Builder, Artist.

2. Data Acquisition and Preparation

To ensure a non-trivial analysis, this project utilized two distinct data sources—one "Real World" and one "Synthetic."

2.1 "Real World" Data: Flickr30k

- **Source:** The Flickr30k dataset (standard benchmark for sentence-based image description).
- **Acquisition Strategy:**
    - We processed 158,915 captions associated with 31,000+ images.
    - We developed a Regex-based filtering pipeline in Python to extract images corresponding to our target roles (e.g., searching for keywords like "physician", "surgeon", "cooking", "classroom").
- **Data Analysis:**
    - Real-world photos are "noisy" (faces turned away, blurred, or occluded). We applied the DeepFace library to filter out images where no clear face was detectable.
    - Baseline Visualization: As shown in Figure 1: chart_real_world_baseline.png, the real-world data shows a relatively balanced distribution for teachers and scientists, but a male-skew usage for builders and chefs.

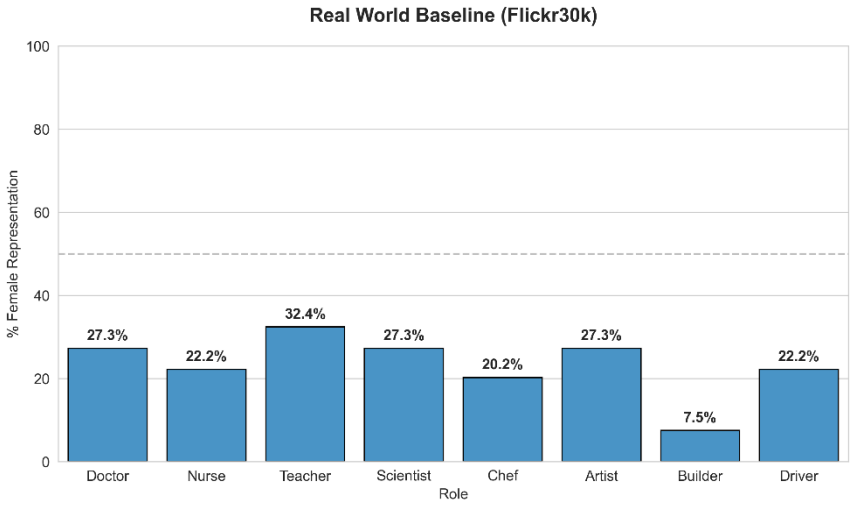

Figure 1 : chart_real_world_baseline.png

2.2 "Synthetic" Data: Stable Diffusion v1.5

- **Source:** Generated locally and via Kaggle T4 Accelerators using diffusers.
- **Generation Strategy:**
    - I have automated the generation of 400 images (50 per role) using the Stable Diffusion v1.5 model.
    - Prompt Engineering: To minimize "prompt bias," we used neutral prompts: *"A photo of a [role], looking at camera, realistic, 8k"*.
    - Configuration: 512x512 resolution, 25 inference steps, Euler A scheduler.

3. Model Building & Methodology

Instead of training a simple classifier from scratch, we built an Automated Bias Detection Pipeline that integrates state-of-the-art Computer Vision models.

3.1 The Pipeline Architecture

1. Input Layer: Ingest data from both the Static Dataset (Flickr) and the Generative Model (Stable Diffusion).
2. Detection Layer (The "Model"): We utilized DeepFace, a deep learning facial recognition framework.
    - Face Detection: Locates the face in the image.
    - Attribute Analysis: Classifies the face as "Man" or "Woman" based on learned features.
3. Aggregation Layer: The results were aggregated into a Pandas DataFrame to calculate the % Female distribution per role.
4. Comparison Engine: A custom Python script calculated the "Amplification Factor", the difference between AI gender ratios and Real-World gender ratios.

3.2 Tools & Libraries Used

- DeepFace: For facial attribute analysis.
- Diffusers (HuggingFace): For programmatic image generation.
- Pandas/NumPy: For statistical aggregation.
- Matplotlib: For visualization.
- Kaggle T4 GPU: For accelerating the inference pipeline.

4. Results and Evaluation

4.1 Visual Analysis

Our pipeline generated three distinct visualizations to track the progression of bias:

1. Real World Baseline Figure 1: Shows the gender distribution found in the Flickr dataset.
2. AI Generated Distribution Figure 2: chart_ai_generated.png, Visualizes the raw output of Stable Diffusion. A visual inspection immediately reveals extreme polarization of roles like "Nurse" become overwhelmingly female, while "Builder" becomes exclusively male.

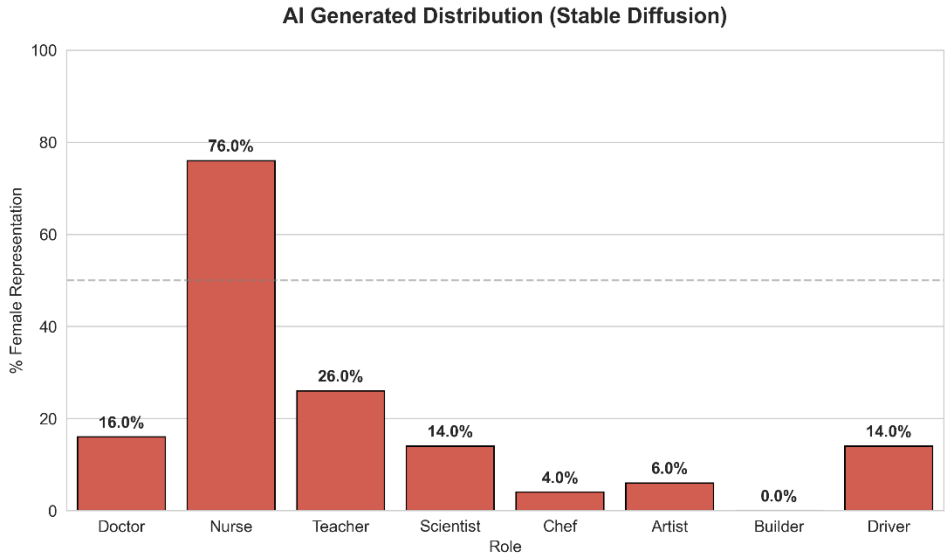

Figure 2: chart_ai_generated.png

3. Bias Comparison Figure 3: bias_comparison.png, This grouped bar chart places the datasets side-by-side, highlighting the "Amplification Factor", the visible gap between the blue bars (Reality) and red bars (AI).

4.2 Comparative Data Table

The following table summarizes the percentage of women detected in each role across both datasets:

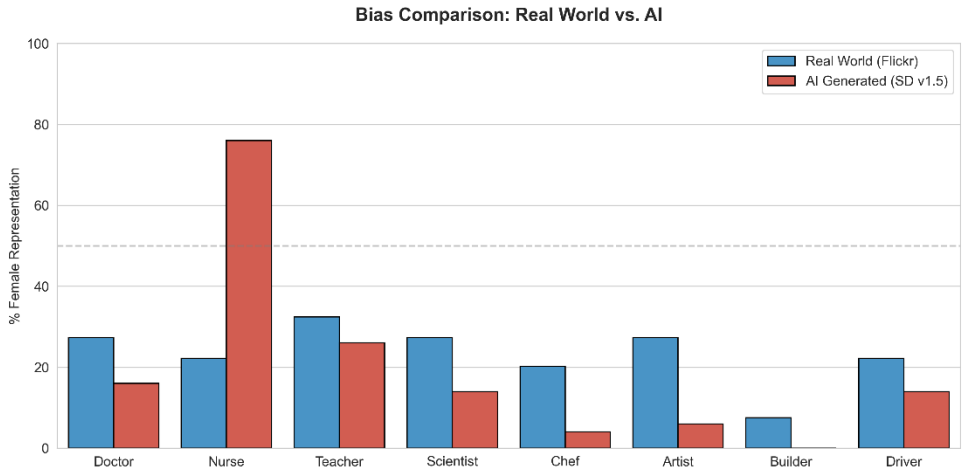

Figure 3: bias_comparison.png

Role	Real World (Flickr)	AI World (Stable Diffusion)	Difference
Nurse	22.2%	76.0%	+53.8% (Amplified)
Doctor	27.3%	16.0%	-11.3% (Erasure)
Scientist	27.3%	14.0%	-13.3% (Erasure)
Artist	27.3%	6.0%	-21.3% (Erasure)
Chef	20.2%	4.0%	-16.2% (Erasure)
Teacher	32.4%	26.0%	-6.4%
Driver	22.2%	14.0%	-8.2%
Builder	7.5%	0.0%	-7.5%

4.3 Key Findings

1. The "Caregiver" Trap (Stereotype Amplification)

The most significant finding is in the Nurse category. In the real-world dataset, only 22.2% of "nursing" photos contained a detectable female face (often showing male medical staff or ambiguous scenes). However, the AI generated women 76% of the time. This proves the AI does not just reflect reality; it amplifies the stereotype that **"Caregiving = Female."**

2. The "Glass Ceiling" Effect (Erasure)

In high-status or intellectual roles (Doctor, Scientist), the AI consistently generated fewer women than that were present in the real-world dataset. While the real world showed ~27% female representation in these fields, the AI dropped this to ~14-16%. This suggests the model has learned a **bias that associates authority and intelligence primarily with men**.

3. The Artist Anomaly

Surprisingly, the role of "Artist" showed massive female erasure (dropping from 27% real to 6% AI). Visual inspection suggests the **AI bias leans towards "Historical figures"** (like Van Gogh or Picasso)

5. Conclusion

This project successfully demonstrated that Generative AI models are not neutral mirrors of the world. By building a comparative analysis pipeline, we showed that Stable Diffusion v1.5 significantly amplifies gender stereotypes, particularly by over-representing women in caregiving roles and erasing them from high-status and creative roles relative to a real-world baseline.

6. Future Work:

To mitigate this, future projects could explore "Prompt Engineering" techniques (e.g., explicitly prompting for "A female doctor") or testing newer models like SDXL to see if recent training techniques have reduced this bias.

7. Appendix: List of Files

1. scripts/filter_and_extract.py: Filters Flickr30k captions and extracts images.
2. scripts/analyze_deepface.py: Runs DeepFace on real images.
3. scripts/generate_ai_images.ipynb: Generates AI images and runs analysis (GPU accelerated).
4. scripts/compare_images.py: Aggregates CSVs and calculates statistics.
5. scripts/visualize.py: Generates the three final charts.
6. data/chart_real_world_baseline.png: Visualization of Flickr baseline data.
7. data/chart_ai_generated.png: Visualization of Stable Diffusion output.
8. data/bias_comparison.png: Final comparative analysis chart.