The Use of Adjectives and its Effects on Representations of Engineers in AI Image Generation

Taylor Caine, Rashidun Rithy, Jerome Rogers, Menachem Grossbaum

Introduction

Literature Review

Sun et al. (2023): gender biases in image-generative Ai

- Underrepresentation of women in male-dominated fields (e.g., engineering).
- Overrepresentation in female-dominated fields (e.g., nursing).
- Stereotypical portrayals (e.g., smiling women looking downward).
- DALL·E 2 shows more noticeable gender biases than Google Images.

García-Ull and Melero-Lázaro (2023):Gender stereotypes in AI-generated images

- AI systems often reflect and amplify societal stereotypes, particularly in professional settings.
- This amplification of biases can exacerbate gender inequalities present in society.
- Addressing these biases is crucial for fair representation in media.

Literature Review

- Everitt, T., Hutter, M., Kumar, R., & Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective.

- Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking.

- Wu, L., & Jing, W. (2011). Asian women in STEM careers: An invisible minority in a double bind. Issues in Science and Technology, (Fall)

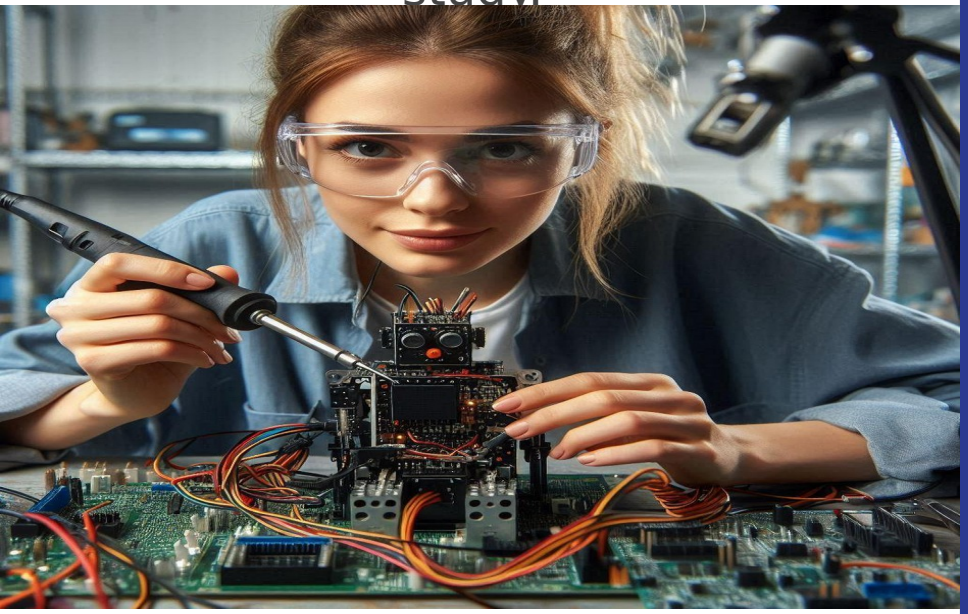
Hypothesis

When doing this experiment, we expected that when using an adjective to describe an engineer, the updated AI would revert back to stereotypical depictions of engineers (white men). When not using any adjectives, it would generate more diverse depictions of engineers.

Materials and Methods for AI-Generated Image Study

Materials Overview

To provide a clear and concise summary of the tools, prompts, and data sources used in the study.



- Co-Pilot: AI-Based generating tool used for creating engineer images.
- Adjectives for Image Generation: Terrible, Ambitious, and Successful.
- Control group for Neutrality: Engineer.
- Source For Demographics: Real-World data from Zippia, focusing on racial and gender demographics of engineers.

Image Generation

To describe the specific process and criteria used to generate the images for the experiment

- Generated 400 Images using the AI image generator Co-Pilot.
- Categories:

  “400 images for Terrible”,

  “400 images for Ambitious”

  “400 images for Successful”.

  “400 Images for Control prompt Engineer”

- No biased decisions established on how our prompts were chosen.

Identifying the Main Characters

To explain the criteria and the process used to identify and count the main characters in the generated images.

- Counted only the Main Characters that were identifiable.
- Excluded the blurred out or background characters.
- In an event that there were characters that have strong similarities, one has only been counted to have consistency in the experiment.

Demographic Categorization (RACE)

To explain the methodology used to categorize the racial demographics of the AI-generated engineers.

- Race Category:

  “White, Black, Latino/Hispanic, Asian, Native American/Pacific Islander, Other”

- Asian Category includes:

  “South and Middle Eastern

- Other category:

  Mainly included to establish a comprehensive comparison.

Demographic Categorization (GENDER)

To explain the methodology used to categorize the gender demographics of the AI-generated engineers.

- This experiment was directed towards the two most recognizable genders.
- Genders were either male or female.
- No other images were mentioned or identified in this experiment.

Data Analysis

To outline the approach used to analyze the demographic data collected from the AI-generated images.

- Gender: Compared AI-generated gender demographics with Zippia's real-world data.

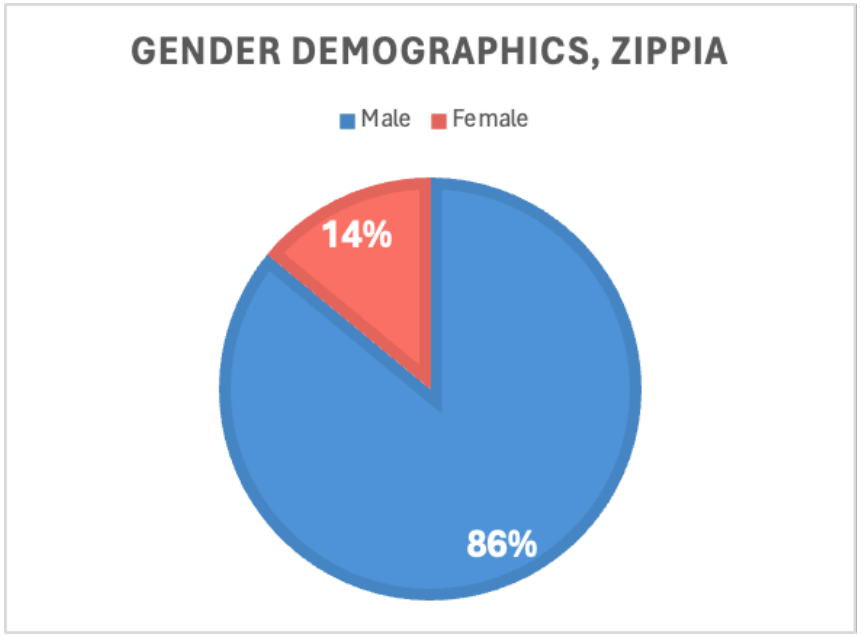
  Used to analyze overrepresentation or underrepresentation.

- Race: Categorized racial demographics of each generated engineer image.

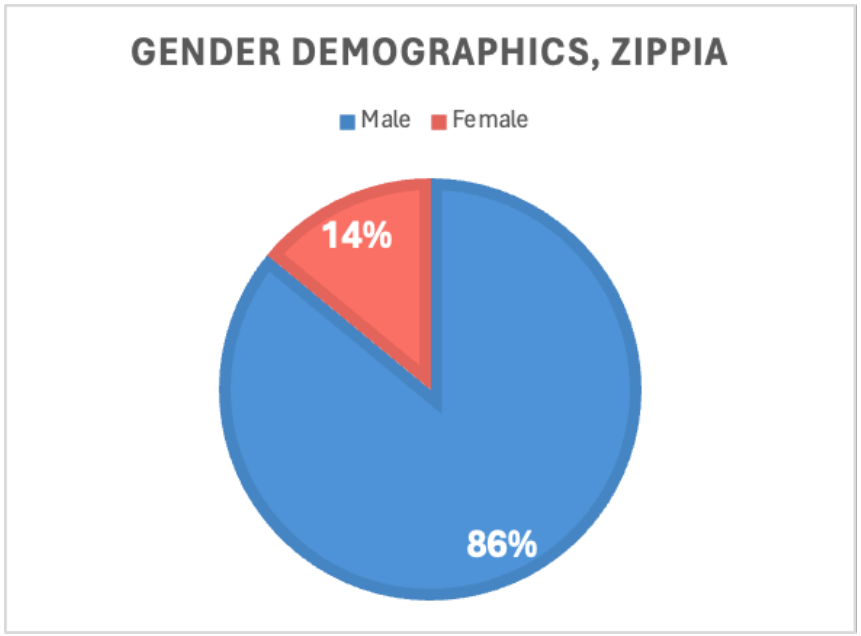
  Focused on identifying discrepancies between AI-generated images and real-world demographics from Zippia.

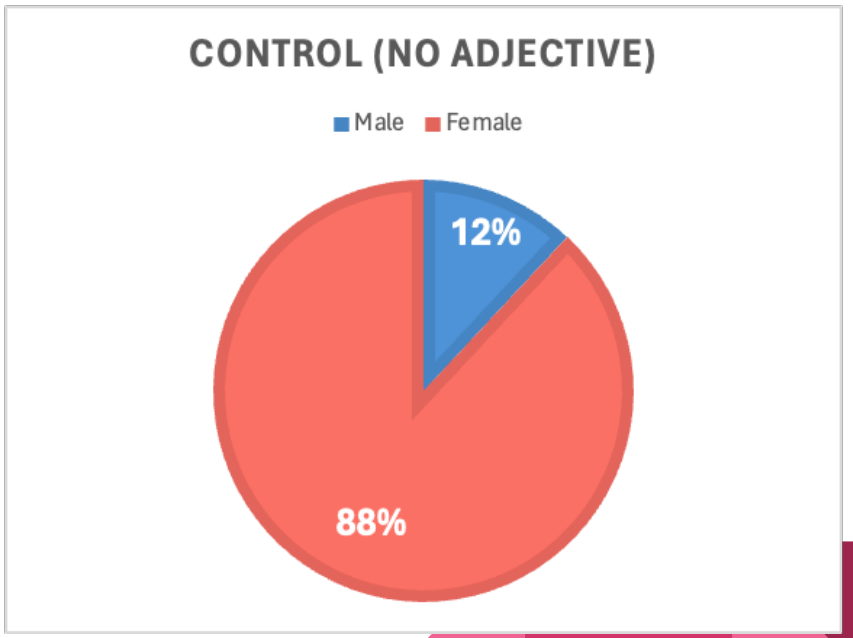
Results

Results - Gender

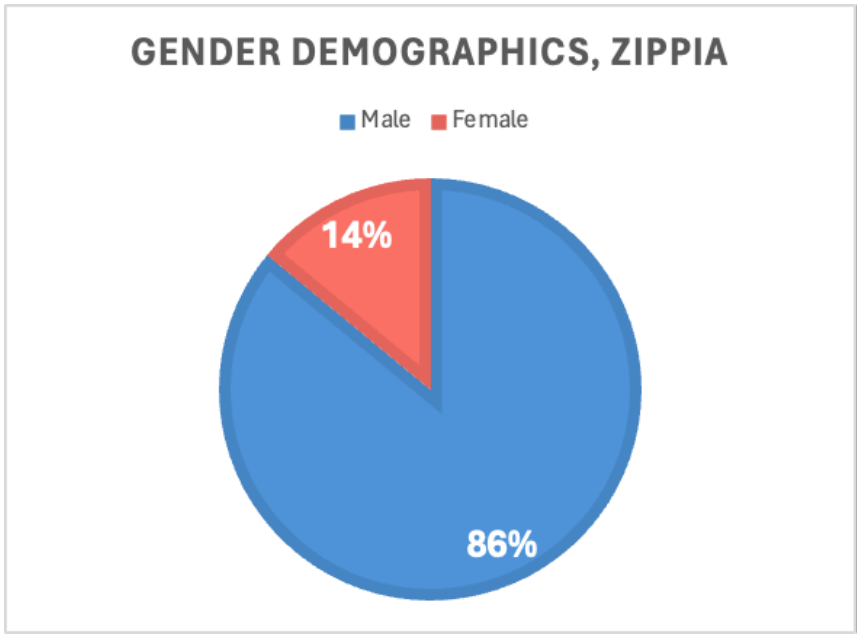


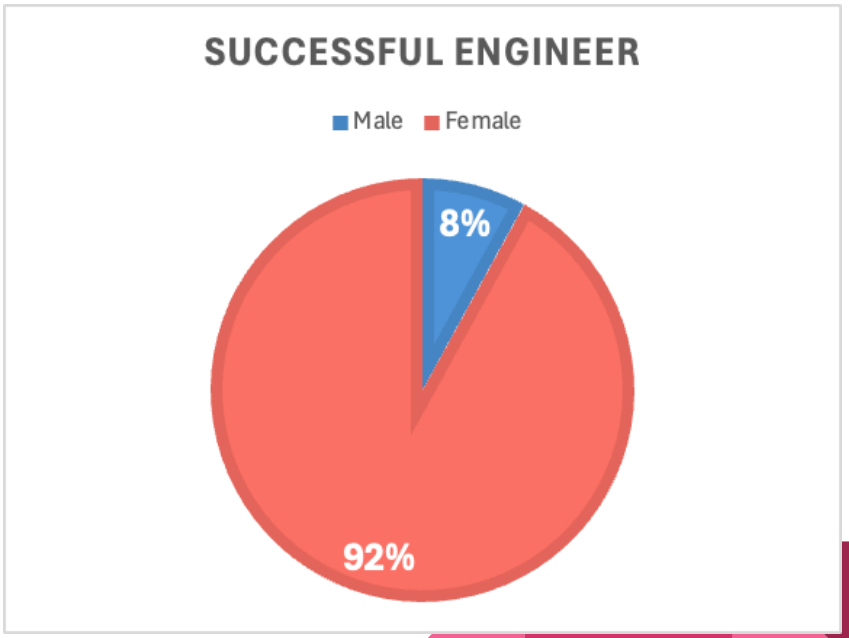
Results - Gender



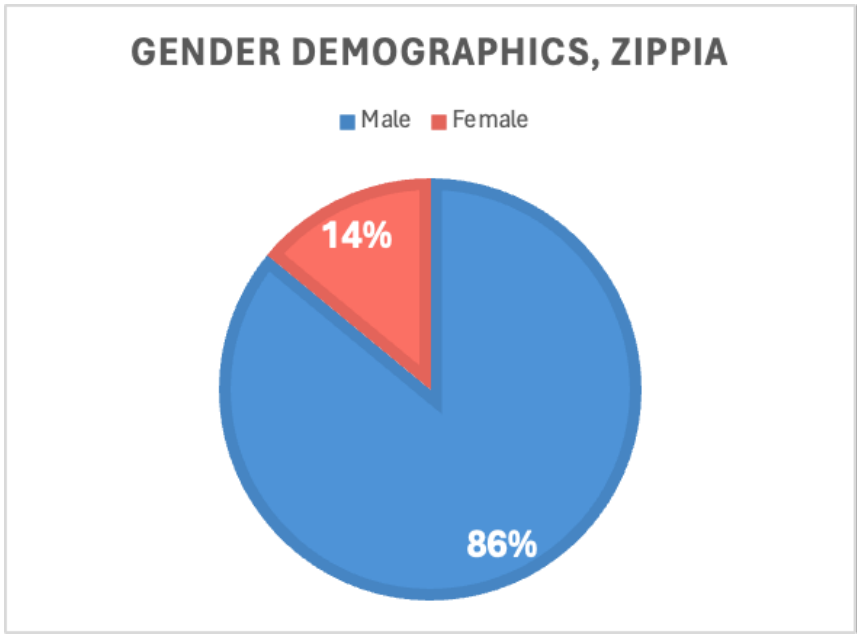


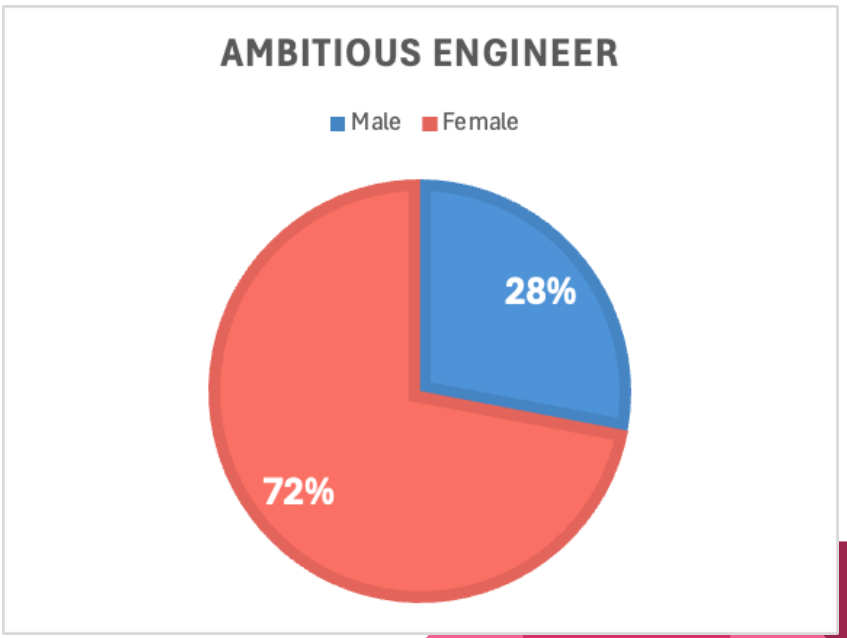
Results - Gender



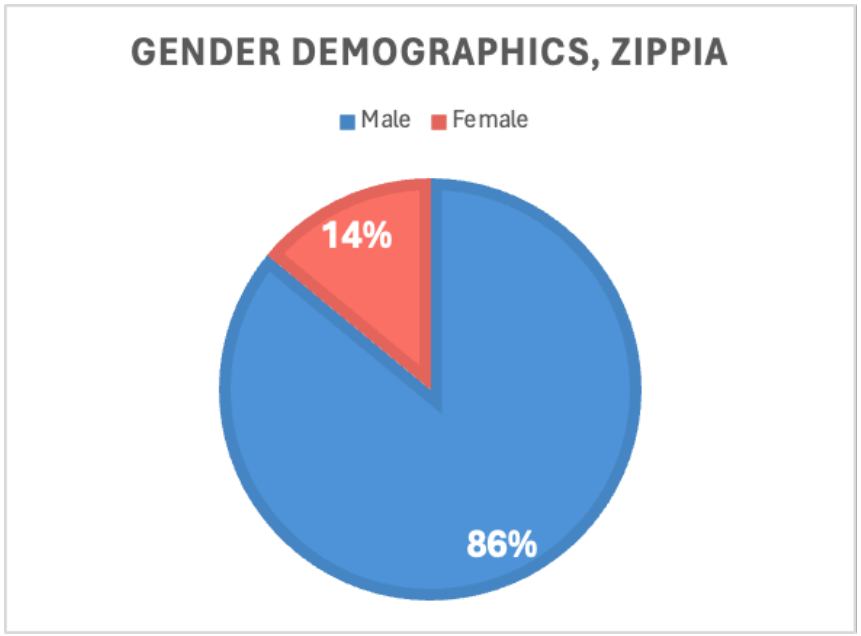


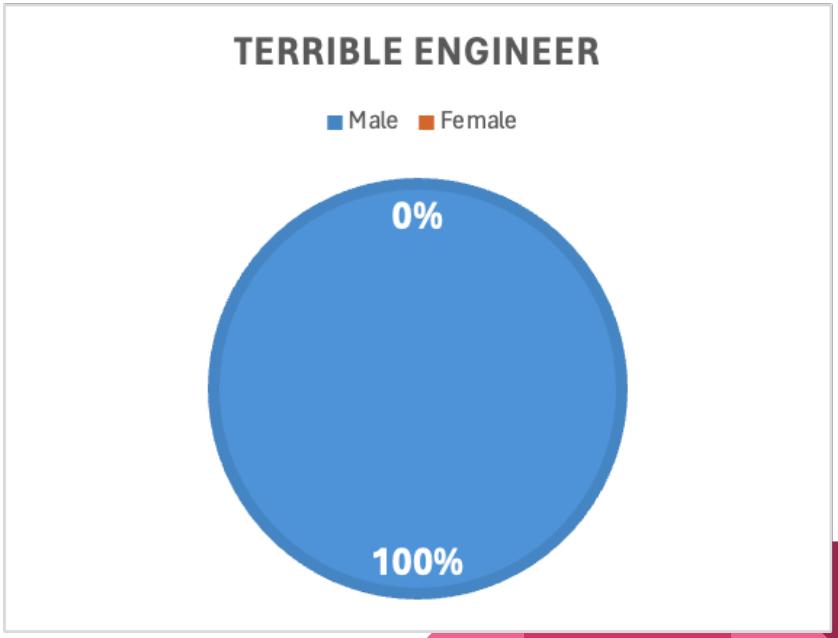
Results - Gender



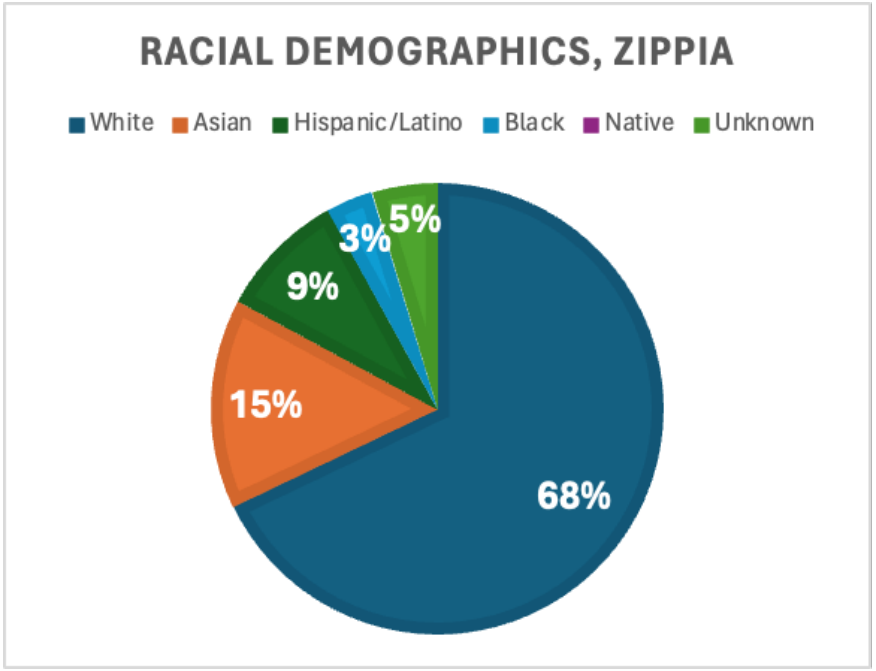


Results - Gender

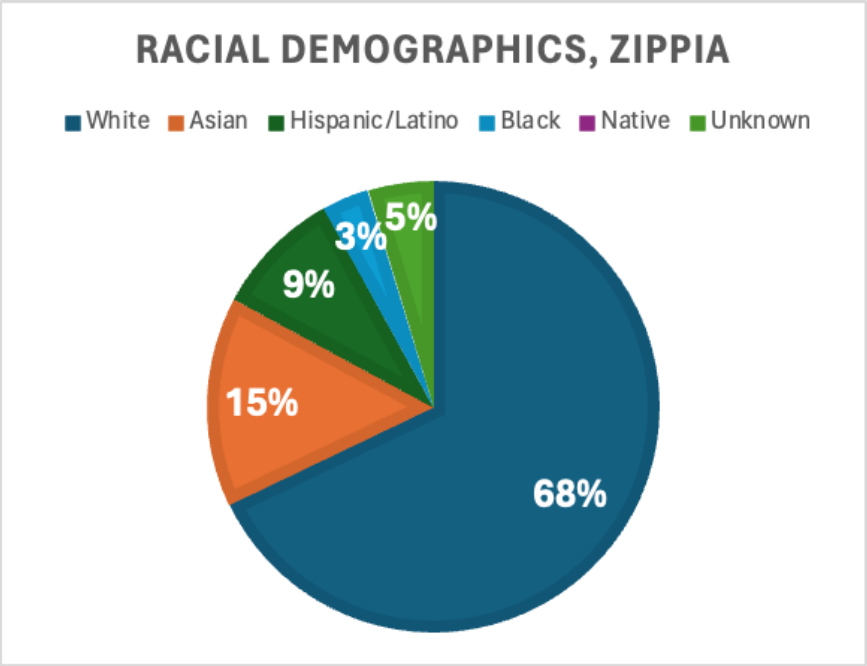


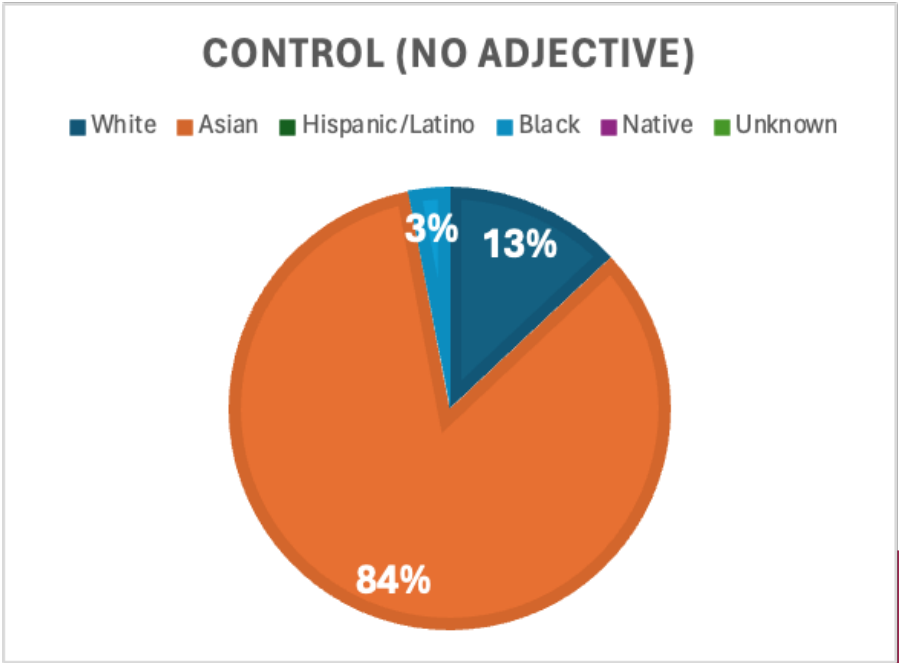


Results - Race

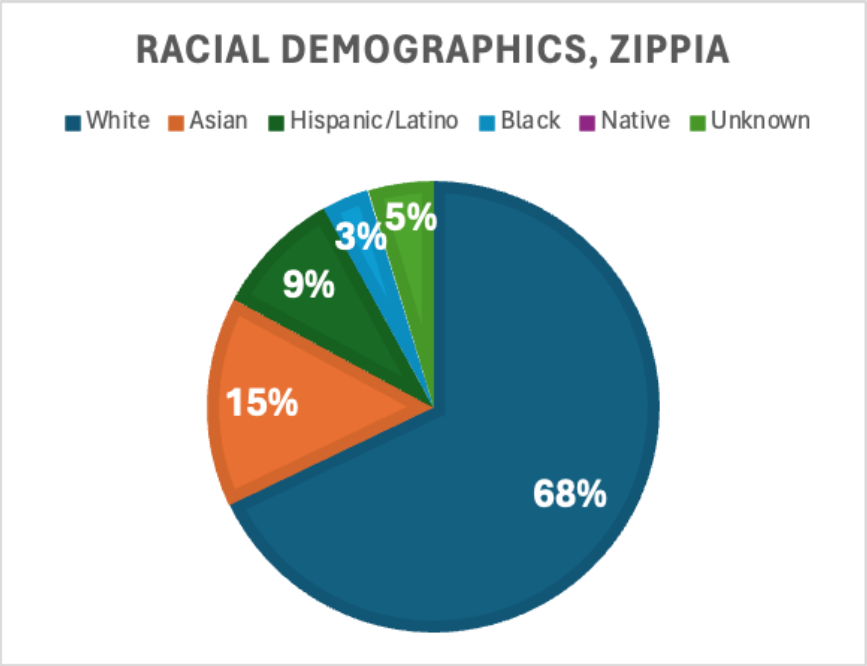


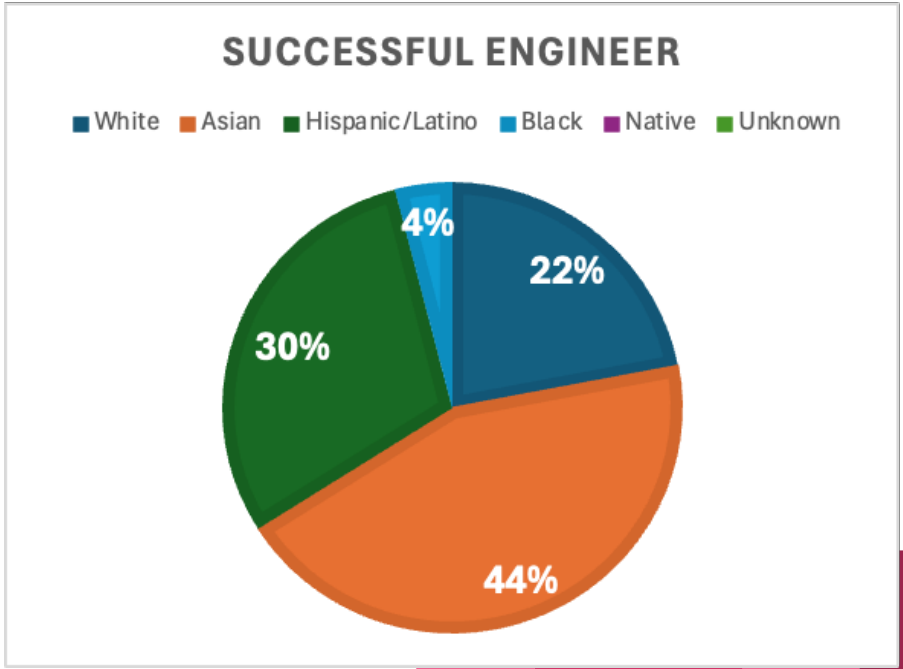
Results - Race



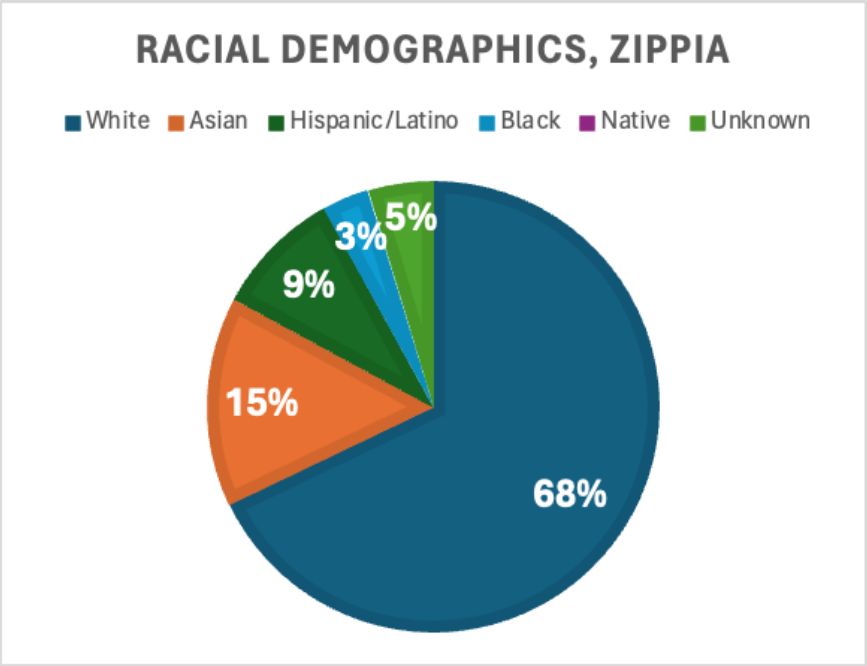


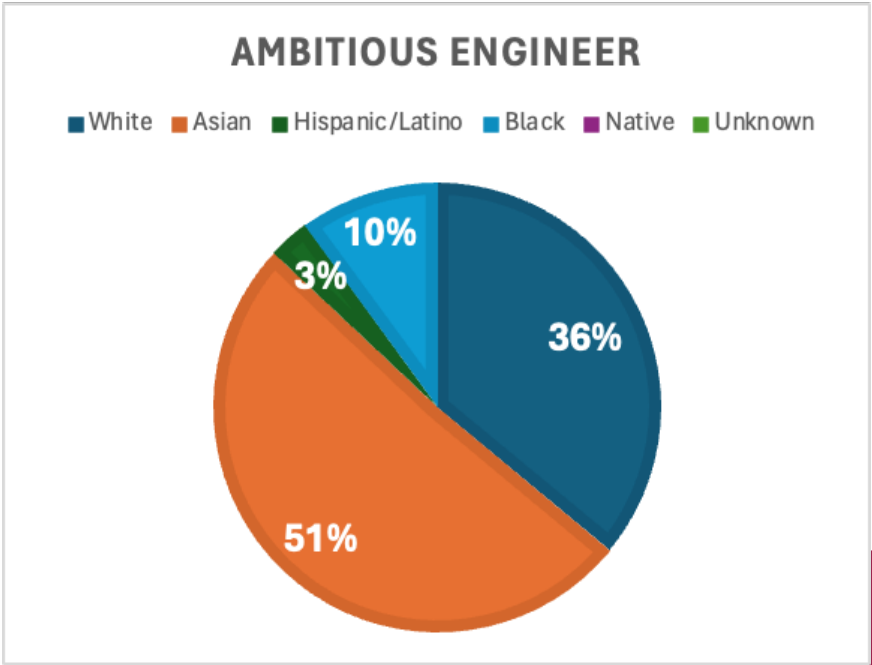
Results - Race



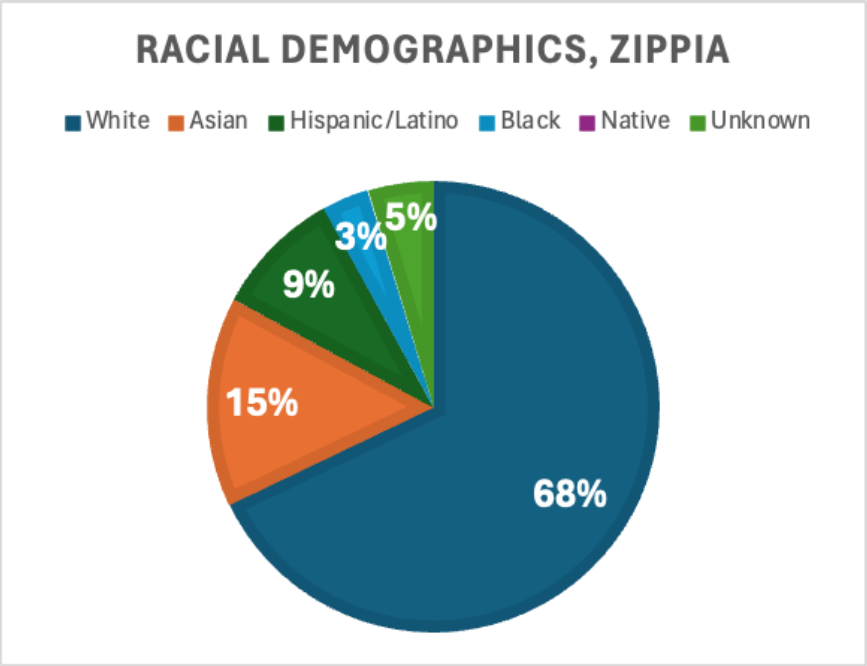


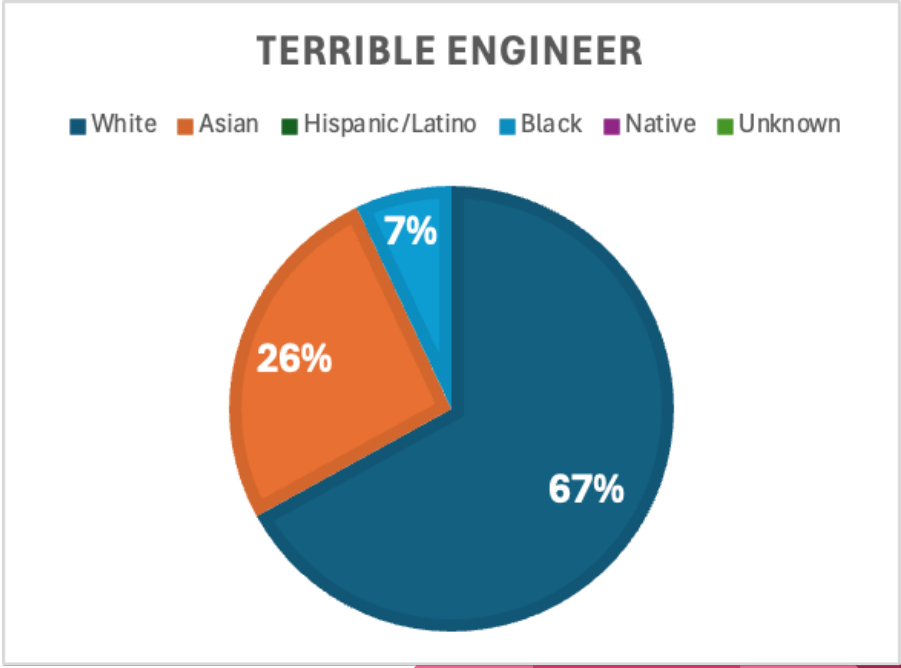
Results - Race





Results - Race

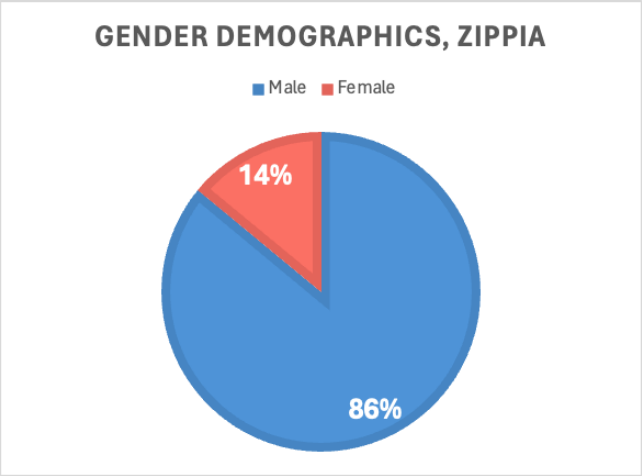


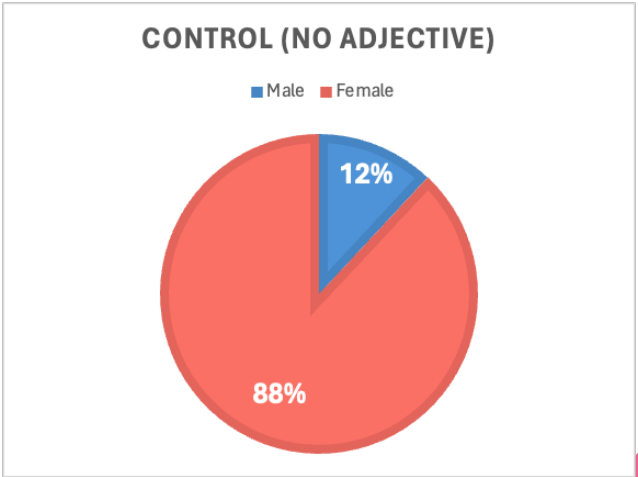


Discussion

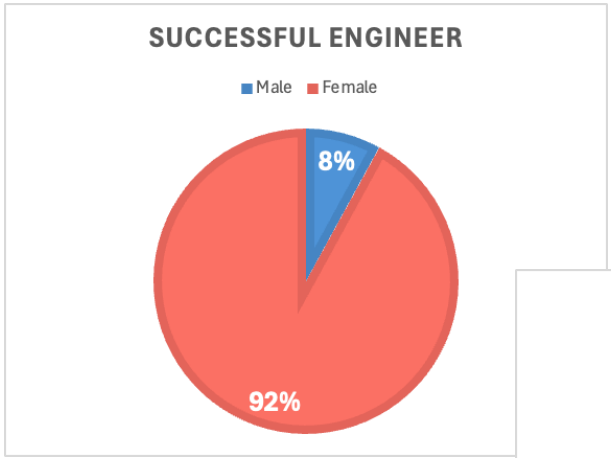
Overcorrection with Reward Hacking: Gender

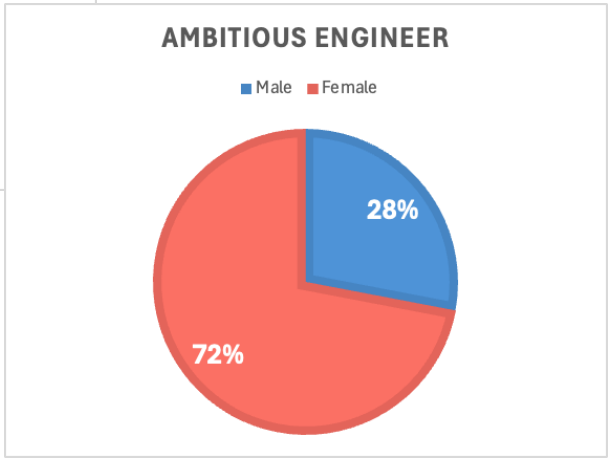
- Real-World Females represent 14% of Engineers
- Females were depicted as 88% of Engineers

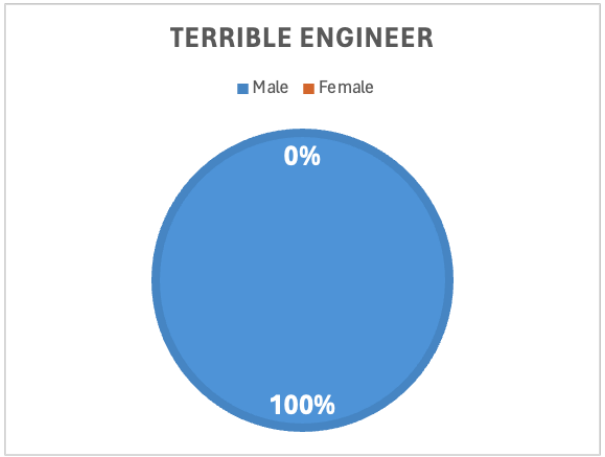




Overcorrection: Positive vs Negative Adjectives







Overcorrection: Positive vs Negative Adjectives

“Successful Engineer”

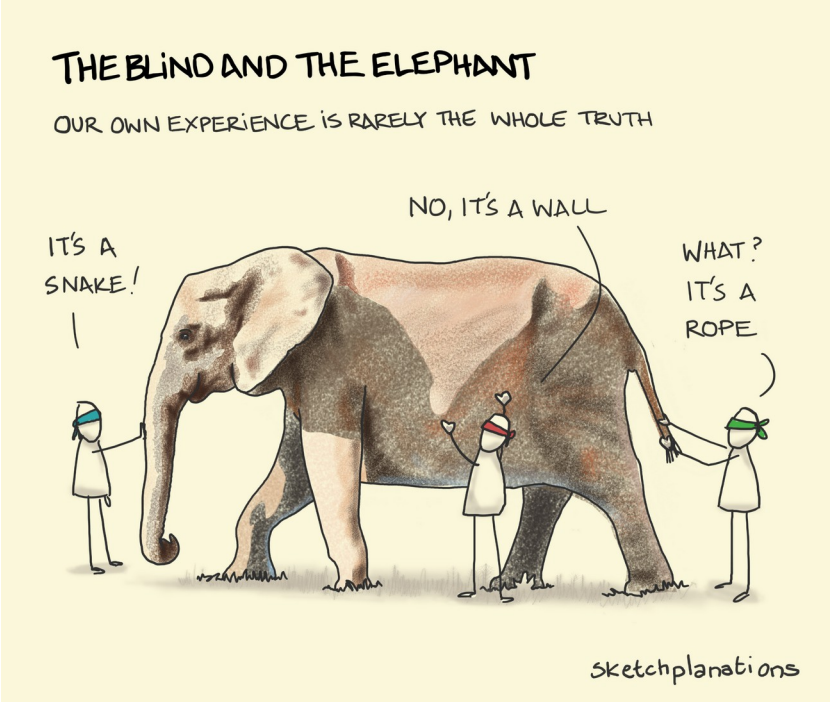
“Terrible Engineer”

Training an Elephant: Reward Hacking

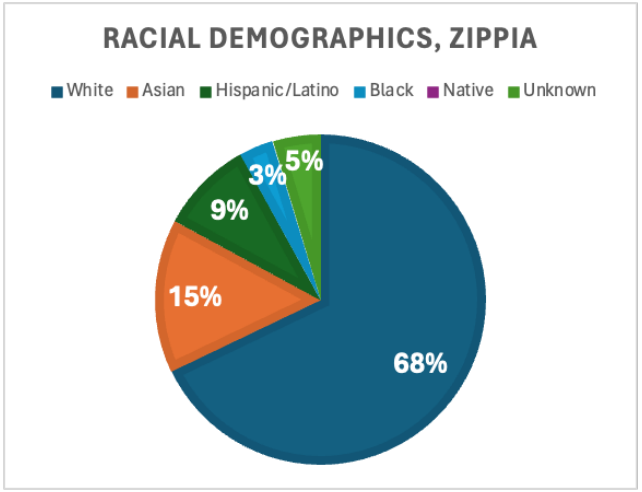
Reward Hacking: Maximizing its rewards without actually mastering the intended skill

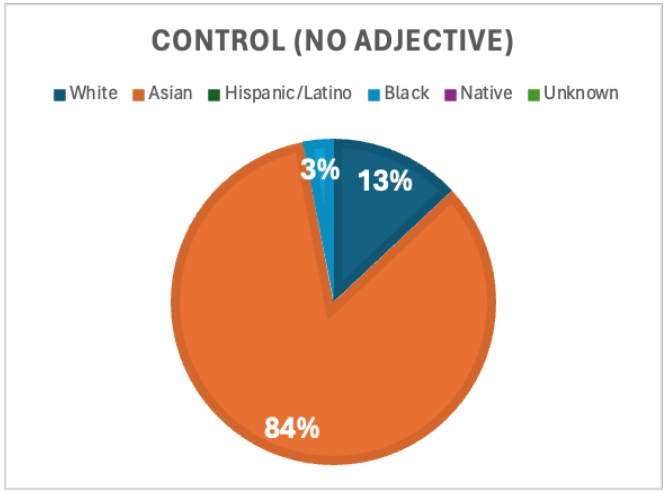
The Blind and the Elephant: Optimization Bias

Optimization Bias: Given bias data the model will reinforce these ideas in an attempt for optimization



Overcorrection in Optimization: Race

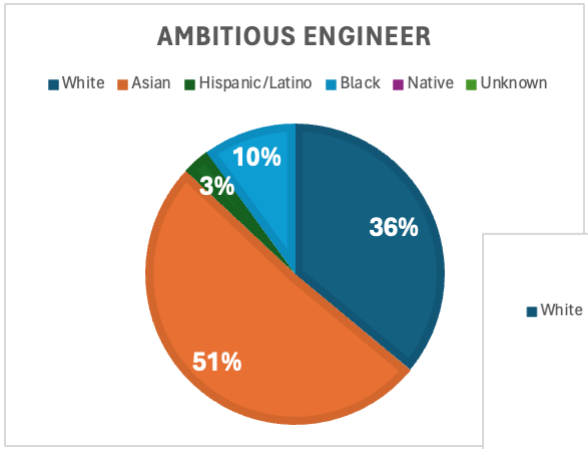


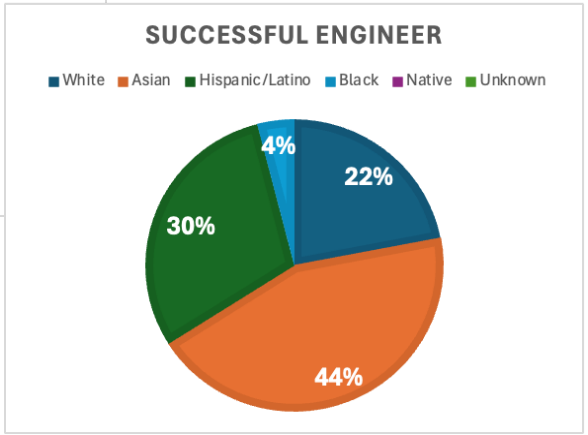


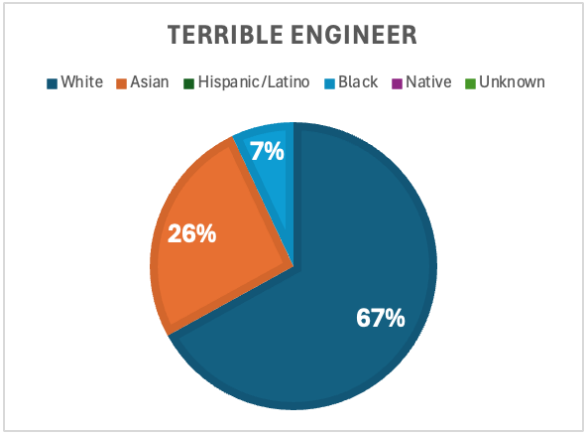
- Asians represent 15% of total real-world Engineers

- 84% Asian Engineers ??!

Overcorrection in Optimization: Race (Adjectives)

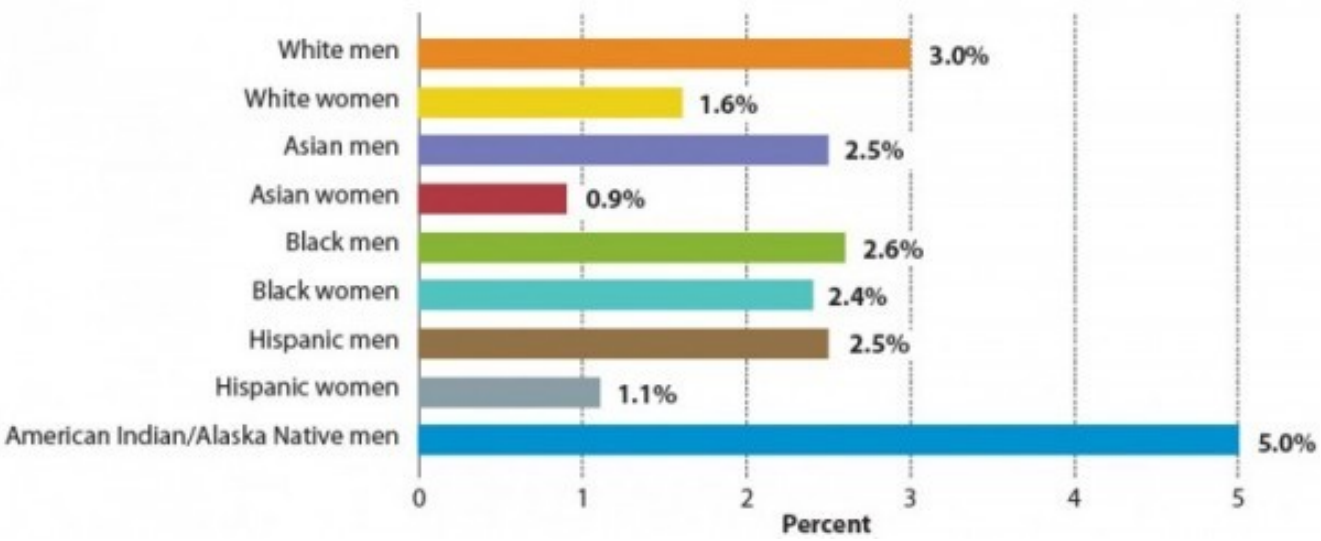






- Asian
- White

Asian Women the “Invisible Minority”



- Percent of Asian Women Engineers (0.9%) that are promoted in the engineering field

Future of AI

- Remove the bias and narrow training data and instead expose it to a broad spectrum of information.
- Design systems that an AI model cannot exploit and manipulate through reward hacking.
- Do not train models to overcompensate for a lack of diversity, this is harmful for fixing societal issues