Law and Artificial Intelligence

Predicting Algorithmic Misgendering

Nov 10, 2025

Context: Understanding Bias in AI Gender Classification

Artificial Intelligence (AI) systems trained on biased datasets can reinforce and even amplify societal prejudices. One common example is gender classification algorithms, which are often trained using datasets with inherent biases. These biases can lead to misgendering, where AI incorrectly labels individuals, particularly those who do not conform to traditional gender norms. In this activity, you will analyze the statistical composition of the *CelebA dataset*, a widely used dataset for facial attribute recognition, to determine whether the dataset is likely to cause misgendering in an AI model trained on it.

Objective

- Try to anticipate the outcome of a specific AI system by analysing the data used to train it.
- Understand how AI training datasets can contribute to biased outcomes.
- Analyse statistical distributions in a dataset to predict potential algorithmic biases.
- Develop critical thinking around ethical AI development.

Instructions

- Examine the distribution of the male variable across the dataset.
- Analyse how other facial attributes (e.g., makeup, facial hair, hairstyle) correlate with the male.
- Observe if certain facial attributes are disproportionately assigned to one gender.
- Identify any imbalances that could affect an AI's gender classification accuracy.

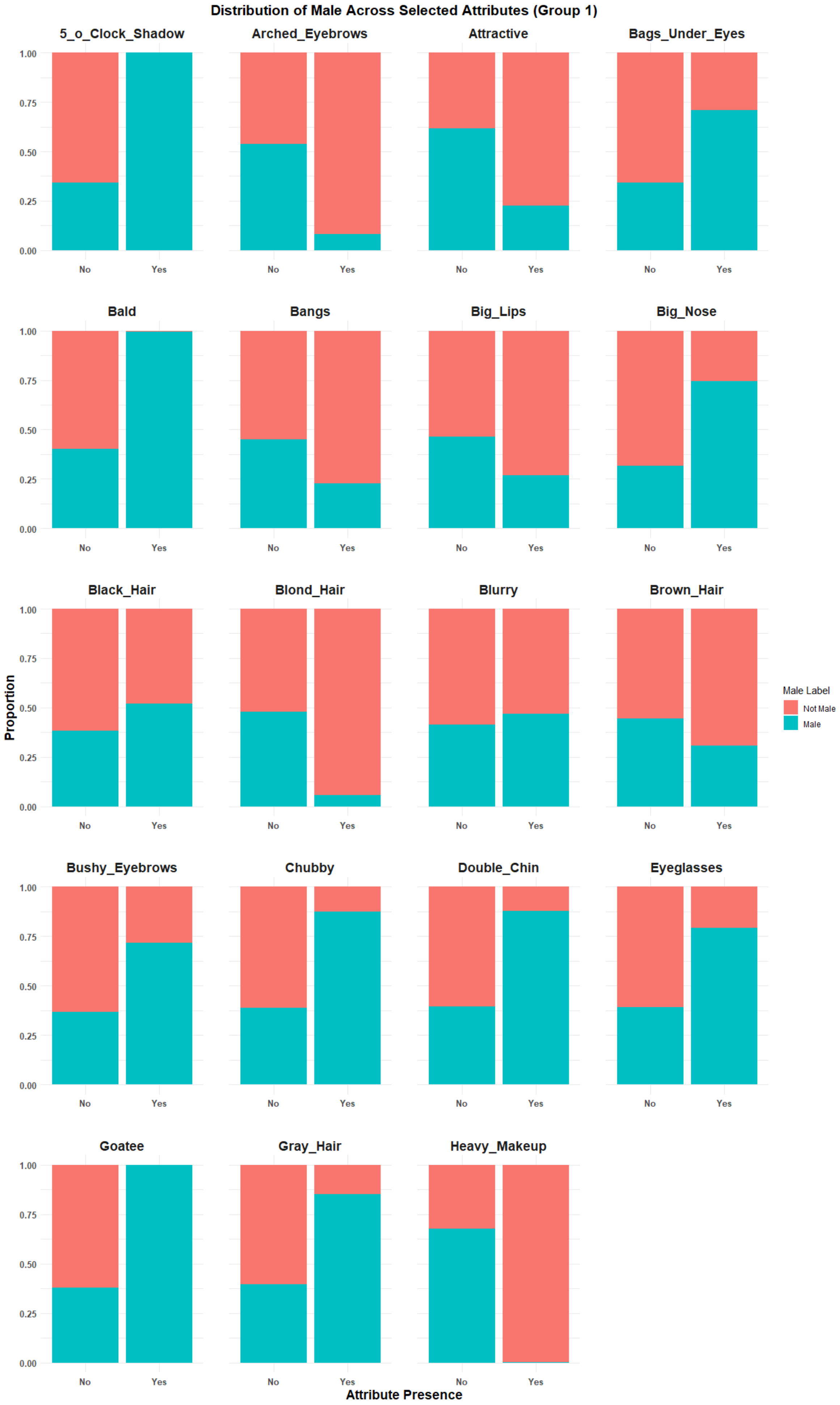
Discussion Questions[1]

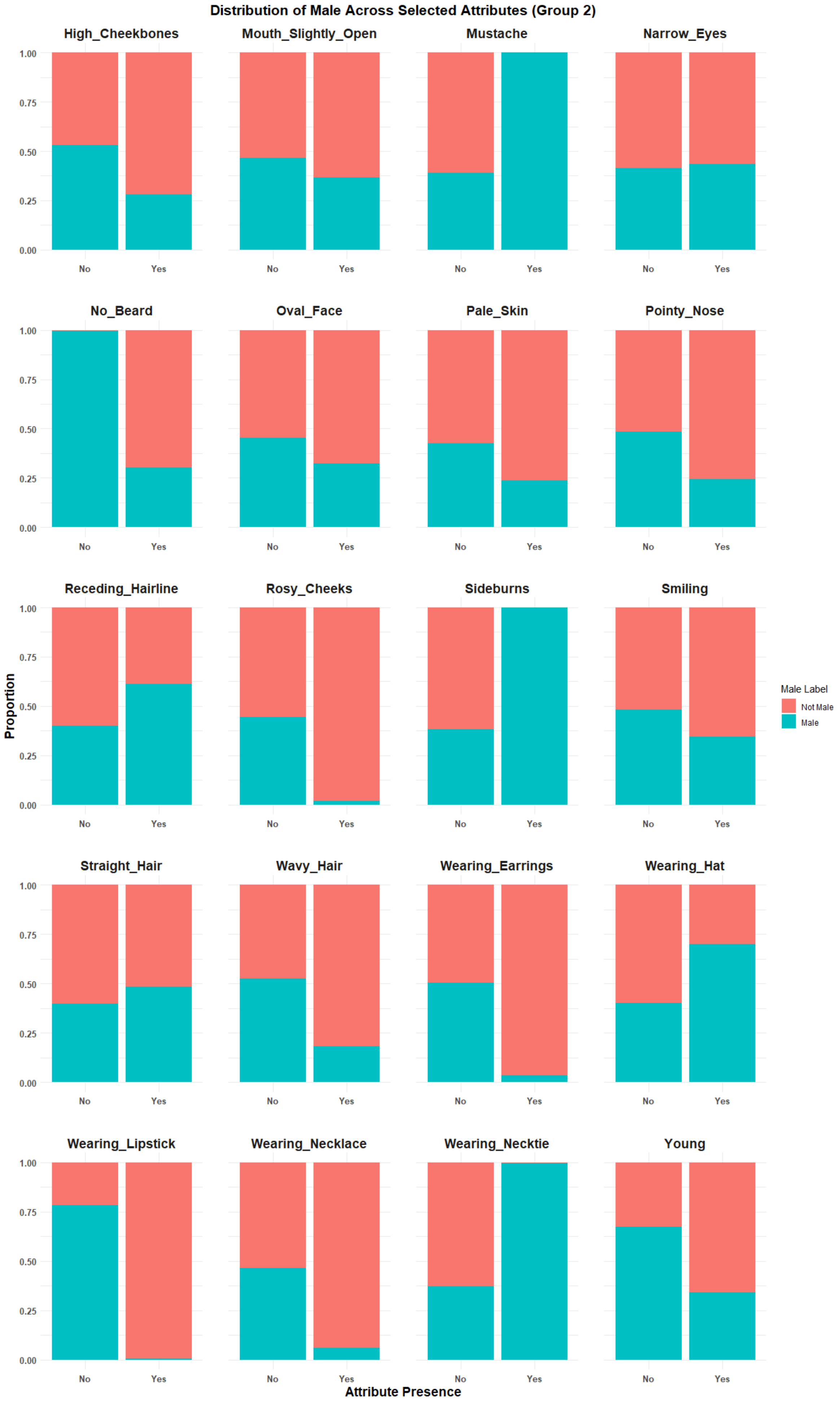
1. Does the dataset appear balanced in terms of gender representation?
2. Which facial attributes are most strongly correlated with the Male label?
3. How might these correlations lead to AI misgendering individuals, particularly non-binary people?
4. If an AI model is trained on this dataset, what biases might it exhibit?
5. What steps could be taken to mitigate bias in AI gender classification models?

Plot Interpretation

The following plots represent the proportion of images labeled as **Male** and **Not Male** across different facial attributes. Each facet in the plots corresponds to a specific attribute used in training the dataset. The bar heights indicate the proportion of images assigned to each gender label for a given attribute. For example, if we consider the attribute **Wearing Lipstick**, we might observe that a significantly higher proportion of images labeled as **Not Male** have this attribute, while very few images labeled as **Male** do.

[1] These questions may help you focus the scope of the debate.


Distribution of Male Across Selected Attributes (Group 1)
5_o_Clock_Shadow
Arched_Eyebrows
Attractive
Bags_Under_Eyes
Bald
Bangs
Big_Lips
Big_Nose
Black_Hair
Blond_Hair
Blurry
Brown_Hair
Bushy_Eyebrows
Chubby
Double_Chin
Eyeglasses
Goatee
Gray_Hair
Heavy_Makeup
Proportion
1.00
0.75
0.50
0.25
0.00
No
Yes
Attribute Presence
Male Label
Not Male
Male


Distribution of Male Across Selected Attributes (Group 2)
High_Cheekbones
Mouth_Slightly_Open
Mustache
Narrow_Eyes
No_Beard
Oval_Face
Pale_Skin
Pointy_Nose
Receding_Hairline
Rosy_Cheeks
Sideburns
Smiling
Straight_Hair
Wavy_Hair
Wearing_Earrings
Wearing_Hat
Wearing_Lipstick
Wearing_Necklace
Wearing_Necktie
Young
Proportion
1.00
0.75
0.50
0.25
0.00
No
Yes
Attribute Presence
Male Label
Not Male
Male