Candidate Interest #2

Neural Network Bias Mitigation

Neural networks are a subset of machine learning, and rely on training data to "learn" from the data and then produce predictions. The neural networks that will be focused on in this study relate to computer vision, as well as facial recognition software commonly used by police and big tech companies such as Microsoft and IBM.

There has been well documented evidence that facial recognition neural networks have shown biases against certain groups of people (marginalized individuals, such as people of color and transgender individuals). Bias in this context is a synonym of prejudice. Many have brought attention to the potential ways that neural networks could receive these biases, but they have been inconclusive about how to mitigate said biases.

How do neural networks receive bias against marginalized people? How do these biases manifest within the code? And how can we potentially mitigate those biases?