The Impact of Bias Detection on the Web

Artificial intelligence is a buzzword that has long been used to elicit emotions from people. Whether those are feelings of fear and nervousness, or excitement and possibility, there is no doubt that AI is one of technology's greatest tools and, one of its greatest unknowns. While human-like robots and self-driving cars are not yet commonplace in society, AI has worked its way into most aspects of tech-driven everyday life. Notably, machine learning helps drive a lot of the personalized recommendations that internet users see daily on social media and the news. Machine learning also impacts industries like finance, production, and healthcare. Consequently, the negative aspects of using AI have also crept into new technologies. It is important that people with knowledge of machine learning combat its negative outcomes before they have a widespread impact. My group's senior design project will aim to increase the fairness of machine learning by identifying biased language corpora on the web. Not only will we utilize artificial intelligence to provide valuable insight into the public content on the internet, but we will also contribute to Aylin Caliskan's research in preventing biased machine learning tools from being released and used in technology.

Caliskan, in her paper "*Semantics derived automatically from language corpora contain human-like biases"*, concludes that many online texts that have been written or influenced by artificial intelligence contain evidence of long held human biases against certain groups or individuals. As machine learning spreads to more aspects of society, it is very important that engineers work to ensure that it does not contain human biases. Most people and industries that utilize machine learning are under the false assumption that since computers produce the information, there is no way that it can contain human prejudice. However, machine learning algorithms are trained on large sets of human-produced data that contains bias. Thus, the algorithms inevitably inherit humans' biases. For example, imagine that a college uses machine learning to screen applicants. Historically, colleges have been known to favor rich, white, male students. Therefore, most data points to the college admitting privileged groups with higher frequency. A machine learning algorithm can only be trained on human-created data, so it will inherit the idea that these privileged students should be favored. Something as small as a biased algorithm creates extremely unfair situations for first generation, low-income, prospective students. Our project will identify biased machine learning data and contribute to allowing algorithms to be trained on unbiased data.

In order to create a web tool to identify biased content, we will need to develop knowledge of how to write, train, and test a machine learning algorithm in Python. This presents some technical challenges, because our group is not very familiar with machine learning or Python. In addition, the largest technical challenge will be finding a proper, large dataset to train the algorithm. Fortunately, our advisor, Professor Caliskan, has done several years of research in biased machine learning and can help us find a usable set of data for the algorithm. In addition, we are currently taking her class on machine learning, which will prepare us to write an intelligent algorithm. As machine learning creeps further into everyday society, engineers must ensure that AI does not contain human biases that have driven prejudice, hate, and inequality since the beginning of time. Our group has the ability to combat the cycle of bias and inequality at its source, before biased machine learning algorithms are released for the public to use. Machine learning is going to be used by employers to screen resumes, healthcare providers who are identifying patients who receive organ donations, and financial institutions which determine who receives loans to buy a new home. All of these industries have had a historic record of discriminating against oppressed groups to maintain a cycle of inequality. My group's senior design project can contribute to ending this sort of discrimination in artificial intelligence before it spreads wider in society.