Bypassing Machine Learning Models to Beat the System

Drew Kinsey

Many students, including me, are familiar with ChatGPT, a large language model, and the ability to use it to aid or do busywork. While it is considered cheating by schools, we continue to use this to make our lives easier. A new machine learning model was created with the goal of predicting our personality traits based on a single picture. Our future employers could use this model in addition to the other hiring practices for our future careers. What I propose is people will find a way to bypass this model with photo editing tools to present themselves as a better job candidate.

First, the model needs to be understood before people can try and bypass it. The model takes a single picture of a person's face and predicts their personality in the big 5 traits. These traits are openness, conscientiousness, extraversion, agreeableness, and neuroticism. By putting us in these categories, our potential employers could decide if we are a good fit for the job.

The model was trained using photos from LinkedIn from graduates with a full-time MBA degree. The number of people trained on the model was around 96,000. An ethical issue from this can be a person's privacy. It is unknown if the people were briefed on their photos being used to train this model.

Personality screenings are not new as a hiring process for a job. Employers have used questionnaires to try and find a person's personality in the past. The appeal of the new model is these questionnaires could become obsolete, streamlining the hiring process.

Many people know they have subconscious bias. A computer system is seen by many people as an unbiased tool. However, machine learning models are not perfect. They are trained and created by humans. While the model claims to control for many biases including race, age, and attractiveness, there still may be some biases the system has that was not controlled by the people making the model. This should be scary to us as college students since many of us do not have the work experience which could combat a personality screening in a hiring process.

Since this model presents itself as unbiased, with claims to control biases, a potential employer can use this model to justify their hiring decisions. An employer who uses this model as a factual unbiased tool should be held accountable for their decision to use this. Since I believe most employers do not know machine learning models can be biased, I believe they will use this model as fact.

Just like many of our fellow students use large language models to combat busy work, I believe the way to combat this new model would be photo editing software. It is in human nature to find a way around guidelines to seek success. While the proposition of editing our photos presents other ethical issues, an employer accepting a machine learning model as total truth has

larger implications. If the model is biased, a group of people will certainly be harmed. The group of people will have a more difficult time finding jobs which can lead to poverty. A person altering their photo to present themselves as a better candidate has less ethical implications. If this person is truly a bad fit for a company based on their personality, they can be fired, making room for a new person.

As the model is used more, the biases and favoritism for different personality traits will increasingly show. This data can be used to find what makes each photo fit in the big 5 personality traits. With this people can potentially alter their photos to make themselves more appealing to personality traits geared towards their careers.

Just like fellow students use ChatGPT to aid in assignments, people will alter their photos to make themselves favorable for their potential careers. While some see this as cheating, I see it as beating the system.