11-16-2023

AI and Concerns on Bias and Correctness in Engineering

Stacy Bjorgaard
University of North Dakota, stacy.bjorgaard@und.edu

AI and Bias

For each scenario, discuss:

- What are the pros and cons of using AI in this scenario?
  - What benefits and/or harm might result from this use?
  - Who is most likely to benefit or be harmed?
- What do you see as possible concerns around bias, privacy, transparency, and ethical use?
- What questions can we ask of this use of AI to expose what is problematic and identify transparent measures to inform stakeholders?

Scenario 1

State University uses Hephaestats, an AI software, to analyze student data gathered through student ID card usage with a phone app serving as proxy. Using this data, the software identified key indicators that, in combination, predict whether a student would drop out with 94% accuracy. The reasons ranged from predictable administrative issues (for example, an overly ambitious or unsuitable combination of courses) to external factors (for example, balancing a job with school, domestic responsibilities). The administration used the information provided by Hephaestats to adjust certain aspects of the campus environment in order to nudge students towards better behavior. For example, advisors encouraged students to change courses to make their schedules more manageable. Hephaestats also supplied faculty with profiles of at-risk students to help them better understand why an individual student might be struggling and suggested targeted approaches for helping them.

Scenario 2

A financial aid office uses AI-powered chatbots to answer student questions on financial aid and other financial and administrative inquiries. The chatbots are designed to recognize common questions and provide answers based on policy and personal data released when a student enters their ID. The office has seen significant savings in time with phone interaction now replaced by the software for the vast majority of inquiries. They recognize chatbots are limited in their ability to interpret language nuances and cannot provide personalized feedback to individual students. They also believe that students prefer to interact with the chatbots rather than staff because they feel no judgment or negative emotion from the chatbot.

Scenario 3

Strategeion, a large organization, had become so overwhelmed with the number of resumes received they had to cease hiring to deal with the backlog. A group of Strategeion's developers interpreted the messages from HR as a call for help and implemented an AI-based hiring tool that scans resumes and selects candidates for interviews. Diagnosing the problem as a simple issue of information overload, this group of developers expected it could be solved by implementing a program to automatically pre-sort resumes according to a candidate's desirability, optimizing especially for projected "fit" within the company. The software uses tone of voice, language, gestures, and facial expressions to create personality profiles for use in hiring processes.

Scenario 4

In April 2017, United Airlines flight 3411 was preparing to take off from Chicago when flight attendants discovered the plane was overbooked. When not enough people were willing to give up their seats in exchange for vouchers and hotel accommodations, United called airport security officers, who boarded the plane and forcibly removed Dr. David Dao, a passenger on the flight. The officers ripped Dr. Dao out of his seat and carried him down the aisle of the plane, nose bleeding, while onlookers captured the scene with their phones. An algorithm likely determined that based on what Dr. Dao had paid for his ticket, what time he had checked in, how often he flew on United, and whether he was part of a rewards program, he was one of the least valuable customers on the flight at the time.

Scenario 5

In March 2019, the Detroit Police Department had run a grainy photo of a Black man with a red cap from Shinola luxury store's surveillance video after a theft through its facial recognition system, made by a company called DataWorks Plus. The system returned a match with an old driver's license photo of Robert Williams. Investigating officers then included Williams's license photo as part of a photo line-up, and a Shinola security contractor (who wasn't actually present at the time of the theft) identified Williams as the thief. The officers obtained a warrant, which requires multiple sign-offs from department leadership, and Williams was arrested. Williams's wrongful arrest was a direct result of the facial recognition system. Two more instances of false arrest have since been made public. Both are also Black men, and both have taken legal action.

Scenario 6

An engineering professor uses AI grading software to grade assignments in a large class to save time and provide consistent grades. Several students in the class received grades significantly lower than they expected and sought redress with the faculty member. In conversations with the company rep, it became apparent that the grading software was trained on a dataset that reflected certain types of solutions. The software penalized students who used different and valid approaches.

What is the AI Bill of Rights?

- From White House, nonbinding, as of October 2022:
- Safe and effective systems:
  - People protected from “unsafe or ineffective AI systems” through “pre-deployment testing and risk mitigation.”
- Non-discrimination:
  - People “should not face discrimination by algorithms and systems should be designed and used in an equitable way.”
- Built-in Data Protection:
  - People should be free from “abusive data practices via built-in protections and you should have agency over how data about you is used.”
- Knowledge and Transparency:
  - “You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.”
- Opting Out:
  - People should have the ability to “opt out” and have access to individuals “who can quickly consider and remedy problems” they experience.

UN Principles for Ethical Use of AI

- Do no harm/protect and promote human rights
- Have a defined purpose, necessity, and proportionality
- Prioritize safety and security, with risks identified
- Be built on fairness and non-discrimination
- Be respectful of individuals' right to privacy
- Be sustainable (socially and environmentally)
- Guarantee human overside & not impinge on human autonomy
- Be transparent and explainable
- Be responsible and accountable to appropriate authorities
- Be inclusive and participatory