Avoiding Algorithmic Bias

adapted from Lori Carter

College admissions

1. List at least 5 things that might have been taken into consideration in granting you admission to your college or university.
2. It is likely that there was a point-based rubric used to predict your probable success at this institution based on some of the criteria you mentioned in question 1. For example, you might get a certain number of points for a high GPA, a certain number for each extra-curricular activity etc. Why do you think this method was used (if indeed it was)? What are the benefits of this method?
3. What are the drawbacks of a point-based method (such as described in question 2)?

Two Scenarios

The two scenarios below are excerpted from "How Algorithms Rule Our Working Lives" by Cathy O'Neil (author of *Weapons of Math Destruction*)

After reading, you will discuss:

- What is the goal of the algorithm(s) mentioned? Who determines this goal?
- Who benefits from this algorithm? Who is potentially harmed?
- Whose data / perspective is included? Whose is left out?
- What are the potential consequences of this algorithm (good and bad)? Think of stakeholders at various levels.

Scenario 1: Financial Market Crash of 2008

After the financial crash of 2008, it became clear that the housing crisis and the collapse of major financial institutions had been aided and abetted by mathematicians wielding magic formulas. If we had been clear-headed, we would have taken a step back at this point to figure out how we could prevent a similar catastrophe in the future. But instead, in the wake of the crisis, new mathematical techniques were hotter than ever, and expanding into still more domains. They churned 24/7 through petabytes of information, much of it scraped from social media or e-commerce websites. And increasingly they focused not on the movements of global financial markets but on human beings, on us. Mathematicians and statisticians were studying our desires, movements, and spending patterns. They were predicting our trustworthiness and calculating our potential as students, workers, lovers, criminals.

This was the big data economy, and it promised spectacular gains. A computer program could speed through thousands of résumés or loan applications in a second or two and sort them into neat lists, with the most promising candidates on top. This not only saved time but also was marketed as fair and objective. After all, it didn't involve prejudiced humans digging through reams of paper, just machines processing cold numbers.

Few of the algorithms and scoring systems have been vetted with scientific rigor, and there are good reasons to suspect they wouldn't pass such tests. For instance, automated teacher assessments can vary widely from year to year, putting their accuracy in question. Tim Clifford, a New York City middle school English teacher of 26 years, got a 6 out of 100 in one year and a 96 the next, without changing his teaching style. Of course, if the scores did not matter, that would be one thing, but sometimes the consequences are dire, leading to teachers being fired.

There are also reasons to worry about scoring criminal defendants rather than relying on a judge's discretion. Consider the data pouring into the algorithms. In part, it comes from police interactions with the populace, which is known to be uneven, often race-based. The other kind of input, usually a questionnaire, is also troublesome. Some of them even ask defendants if their families have a history of being in trouble with the law, which would be unconstitutional if asked in open court but gets embedded in the defendant's score and labelled "objective".

[The popularity of these predictive algorithms] relies on the notion they are objective, but the algorithms that power the data economy are based on choices made by fallible human beings. And, while some of them were made with good intentions, the algorithms encode human prejudice, misunderstanding, and bias into automatic systems that increasingly manage our lives.

Scenario 2: Automation in Hiring

Finding work used to be largely a question of whom you knew. Companies like Kronos brought science into corporate human resources in part to make the process fairer. The hiring business is becoming automated, and many of the new programs include personality tests. Such tests now are used on 60 to 70% of prospective workers in the US.

Defenders of the tests note that they feature lots of questions and that no single answer can disqualify an applicant. Certain patterns of answers, however, can and do disqualify them. And we do not know what those patterns are. We're not told what the tests are looking for. The process is entirely opaque. What's worse, after the model is calibrated by technical experts, it receives precious little feedback.

Sports provide a good contrast here. Most professional basketball teams employ data geeks, who run models that analyse players by a series of metrics, including foot speed, vertical leap, free-throw percentage, and a host of other variables. Teams rely on these models when deciding whether or not to recruit players. But if, say, the Los Angeles Lakers decide to pass on a player because his stats suggest that he won't succeed, and then that player subsequently becomes a star, the Lakers can return to their model to see what they got wrong. Whatever the case, they can work to improve their model.

Naturally, many hiring models attempt to calculate the likelihood that a job candidate will stick around. Evolv, Inc, helped Xerox scout out prospects for its call centres, which employ more than 40,000 people. The model took into account some of the metrics you might expect, including the average time people stuck around on previous jobs. But they also found some intriguing correlations. People the system classified as "creative types" tended to stay longer at the job, while those who scored high on "inquisitiveness" were more likely to set their questioning minds towards other opportunities.

But the most problematic correlation had to do with geography. Job applicants who lived farther from the job were more likely to churn. This makes sense: long commutes are a pain. But Xerox managers noticed another correlation. Many of the people suffering those long commutes were coming from poor neighbourhoods. So Xerox, to its credit, removed that highly correlated indicator of churn from its model. The company sacrificed a bit of efficiency for fairness.

What is AI?

Point-based algorithms for making decisions are a rudimentary form of Artificial Intelligence. Kaplan and Haenlein recently defined AI as

> a system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation.

Our example of a point-based algorithm for college admission is used to predict who will most likely succeed at a particular college or university. Similarly, each of the other scenarios that you read involved simple predictive algorithms.

EU Ethics Guidelines for Trustworthy AI

We hear frequently of the challenges with AI and even simple predictive algorithms share the same issues. Recently, the EU came up with a set of guidelines to help diminish the chance of ethical issues in such algorithms. As you go through these guidelines,

- Initial reactions to the guidelines?
- How are they compatible or incompatible with biblical ethics / a Christian worldview?
- How would the guidelines would have benefited the developers of the previous algorithms that you considered to help them avoid bias or other pitfalls? Name the guideline and articulate how it would have helped.

1. Human agency and oversight

   Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. Humans are involved in the design and governance of the system including when and where the system is used.

2. Technical robustness and safety

   Including resilience to attack and security, fall back plan and general safety, accuracy, reliability with a range of inputs and reproducibility of results with same inputs. The system is well-tested during creation and reviewed during use.

3. Privacy and data governance

   Including respect for privacy, quality and integrity of data, and access to data. Data must be ethically cleaned before it is used.

4. Transparency

   The data sets on which the decision making is based must be well-documented with regard to origin and alteration (such as in cleaning). Decisions made by AI must be understandable and able to be traced. Users should have access to the decision-making process and be aware that the decision was made, at least in part, but a computer.

5. Diversity, non-discrimination and fairness

   Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation. Diversity should be encouraged and designers should consider carefully whether past positive results (success of people with certain gender, ethnicity, personality etc.) were related or unrelated to the success. Avoid propagation of past bias. Systems should be user-centric and accessible to the widest range of users. People who will be affected by the system should be involved throughout the design of the system.

6. Societal and environmental wellbeing

   Including sustainability and environmental friendliness, social impact, society and democracy. Supply chain for the system should be sustainable. Its use should impact social relationships in a negative way. It should enhance, rather than detract from, a democratic society.

7. Accountability

   This necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use. Potential impacts should be minimized and when encountered, they should be addressed and there should be a policy for making amends to the injured party. The system should be adjusted to eliminate the possibility of this happening in the future. When a conflict of interest arises, the tradeoffs made should be acknowledged.