Jenn Halen
Whitepaper Submission
Trustworthy Algorithmic Decision-Making

Fairness and Algorithmic Decision-Making

Algorithmic decision-making applications will require individual consumers and, at some point, the general public to believe that the technology's decisions are sound, and while many factors may contribute to that belief, the idea that the algorithm will produce "fair" results will weigh heavily on many people's minds. As with many instances that pit one person's interests against another's, notions of fairness are highly contextual and conflicting, particularly in a time when so much of the public debate about any issue is highly polarized and politically charged. Algorithmic fairness is and will be entangled in contentious debates about expertise (who should decide what fairness means?) and power (who can decide?).

Fairness can be defined in numerous ways, but a framework that is particularly prevalent in the underlying ideological arguments about the subject is that of *is* versus *ought*.

One individual may believe that a fair outcome is the natural result when each individual is unencumbered by outside forces, either for good or ill. However, another may believe that a fair outcome is exactly the opposite; that is, fairness occurs when competing sides are made equal to one another. In short, the former believes fairness as what is (i.e., an objective truth that exists without interference), while the latter believes that fairness is what ought to be (that outside intervention should occur if the nature of the dispute or playing ground has put opposing sides on unequal footing).

This foundational view must be addressed in any system, including in software systems, before other considerations are made for any specific decision-making context. For example, if an algorithmic recommendation system were to be implemented by a university to recommend (or, potentially, determine) student admission, each metric, its corresponding weight, and the outcome of those combined measures, could give a different version of fairness than another system. In a simple example, say that there are two students being analyzed, but only one admissions spot is open. If each student's primary metrics were the same (high GPA, high test scores, similar extracurricular activities, etc.), a decision might come down to other features, but there are many attributes that could be characterized as a positive or a negative attribute (either explicitly by software engineers or implicitly by a dataset).

If both students have the same GPA, say a 3.7 out of 4, but student A rose steadily throughout the previous 3 and a half years, while student B maintained a steady 3.7 for their entire high school career, who is the better high school student? Or a different question that may be more relevant for the issue of college admittance; who will be the better college student? Both of these perspectives stem from a view of what is "fair." Is the student who has maintained consistency a better learner, or is that recognition better suited for a student who has illustrated that they have the drive to improve?

Mathematically, Student A has established an upward trend that may be indicative of further growth, so that may warrant either a higher weight or even a separate variable that measures improvement. Furthermore, the context in which Student A's grades rose may also be taken into consideration and even valued more if they have a background that may have deterred their achievement, such as if they had a low family income that required them to work full time and go to school. This view corresponds, at least in part, with the perspective that if the world had been different, if Student A had been given equal footing with their peers, if the world were as it ought to be, they would be the better student. And perhaps an algorithm that weighs all variables appropriately, or machine learning software that can analyze vast amounts of data to better determine upward trends, could find that and could predict that Student A would make a stellar college student. The point is that this outcome takes a different view than a side-by-side snapshot of the students. At present, Student A and Student B are exactly equal, and, in side-by-side snapshots of the past, Student B had superior scores. Student B's consistency may even be considered a more admirable quality than Student A's history of improvement.

Two human school administrators might make different decisions if given this case, and so might two algorithmic recommendation systems. The value of algorithmic decision-making lies not in finding an unimpeachable right answer to what is fair, but rather in its ability to more concisely, effectively, and predictably weigh what we, as subjective humans, determine to be fair. Part of building trustworthy machine learning and artificial intelligence software must be to explicitly and transparently articulate those definitions and the benefits and pitfalls of using algorithmic decision-making tools to achieve the outcomes we seek.