MTLR MICHIGAN TECHNOLOGY LAW REVIEW

Anti-Discrimination Laws and Algorithmic Discrimination

Machine algorithms can discriminate. More accurately, machine algorithms can produce discriminatory outcomes.

It seems counterintuitive to think that dispassionately objective machines can make biased choices, but it is important to remember that machines are not completely autonomous in making decisions. Ultimately, they follow instructions written by humans to perform tasks with data provided by humans, and there are many ways discriminations and biases can occur during this process. The training data fed to the machine algorithm may contain inherent biases, and the algorithm may then focus on factors in the data that are discriminatory towards certain groups. For example, the natural language processing algorithm "word2vec" learns word associations from a large corpus of text. After finding a strong pattern of males being associated with programming and females being associated with homemakers in the large text datasets fed to it, the algorithm came up with the analogy: "Man is to Computer Programmer as Woman is to Homemaker."

Such stereotypical determinations are among the many discriminatory outcomes algorithms can produce. The European Union (EU), out of fear of these outcomes leading to discriminatory effects produced by decision-making algorithms, included Article 22 when enacting the General Data Protection Regulation, which gives people "the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Although what constitutes "solely automated processing" is debatable, the EU's concern of algorithmic discrimination is evident.

In the United States (U.S.), instead of passing laws that specifically target algorithmic discrimination, such concerns are handled largely under regular anti-discrimination laws, such as the Equal Protection Clause, the Civil Rights Act of 1964, the Fair Housing Act, etc. However, these traditional anti-discrimination laws are difficult to apply to algorithms that produce discriminatory outcomes.

The anti-discrimination laws in the U.S. have a long history of focusing on two modes of discrimination – disparate treatment and disparate impact. The prohibition on disparate treatment forbids unequal treatments based on protected characteristics (race, gender, age, etc.), while the prohibition on disparate impact forbids inadequately justified behaviors that have disproportionate adverse effects on people with protected characteristics. Neither of these prohibitions can effectively regulate algorithmic discrimination.

Laws that prohibit disparate treatment, such as the Civil Rights Act of 1964, emphasize the motive and intent to discriminate, to treat people with protected characteristics differently. It advocates for a commitment to neutrality, a "blindness" to the protected characteristics, in the decision-making process. Translated to the algorithmic context, it seems to require algorithms to not consider these protected characteristics as weighted factors when making decisions. An algorithm might be fed training data containing strong stereotypical associations involving protected characteristics that indicate past biases and discriminations, such as the association between females and homemaker in the word2vec case, or an association between Black Americans and marijuana use in the historical criminal data fed to a pretrial risk assessment algorithm. The disparate treatment prohibition requires the algorithm to "look past" these stereotypical patterns and make neutral decisions that are not influenced by factors like gender or race.

The first obstacle with this requirement is that it is often hard or even impossible to tell what factors the algorithm considers when making decisions. For example, pattern matching, a machine learning process with a variety of applications (web search engines, feature detection, etc.), is so complex that researchers and developers do not always understand which factors the machine considers to be meaningful patterns, or how significantly these factors affect the machine's decision making process. It would be difficult to decide if a pattern matching algorithm is "blind" for disparate treatment purposes, when we do not even know what and how the algorithm "sees".

On the other hand, there are active efforts in technological research that try to prevent an algorithm from "seeing" the protected characteristics. These efforts are called the "fairness through unawareness" approach in machine learning literature. However, this approach "is widely considered naïve" "in the machine learning community". Because the machine can process an enormous number of variables, characteristics, and patterns, it can easily correlate a protected characteristic with an unprotected characteristic, such as correlating male with taller body height and female with shorter body height. To prevent a machine algorithm that has been trained with data containing strong associations between males and violent crimes from using gender as a decision-making factor can still result in the algorithm determining that taller people tend to commit violent crimes. This determination is not only imprecise, but also still biased. Realistically speaking, these "taller people" are more likely to be men than women. Hence, "even if developers deliberately avoid using variables for protected classes, such systems can still produce a disparate impact if they use variables that are correlated with both the output variable the system is trying to predict and a variable for protected-class status."

Whether a legal claim exists based on such disparate impact is also unclear. A disparate impact claim requires a showing that a facially neutral policy has a disproportionately adverse effect on members of a protected class that is not justified by business necessity. This standard is clearly created with human decision-making in mind. When we do not understand which factors an algorithm considers when processing data, and how these factors influence decision-making, we do not know where to start with the justification analysis. In addition, to mitigate disparate impact on members of protected classes, many machine learning algorithms actively utilize the protected characteristics captured in the training data to create better trained machine learning models. Ironically, such utilization of protected characteristics might be deemed as disparate treatment under current anti-discrimination laws, since the prohibition on disparate treatment requires a "blindness" and neutrality to these protected characteristics.

Find More

As machine learning algorithms with decision-making capacities evolve, they will become even more complex, and be applied to many more fields, including autonomous driving, housing, criminal justice reform, finance and investment, legal research, etc. Issues with algorithmic biases and discriminations will arise more frequently with this increase in complexity and widespread application, and our laws need to be prepared for it. It is perhaps time for laws that specifically target algorithms. In April, 2019, the Algorithmic Accountability Act of 2019 was proposed by Senators Cory Booker and Ron Wyden. The proposed act "direct[s] the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments." This proposed act "is the first federal legislative effort to regulate AI systems across industries in the United States, and it reflects a growing and legitimate concern regarding the lawful and ethical implementation of AI."

** Kay Li is a Managing Executive Editor on the Michigan Technology Law Review.*

E-mail: **mich.tech.law.rev@umich.edu**

Find More