Limitations of mitigating judicial bias with machine learning

Machine-learning algorithms trained with data that encode human bias will reproduce, not eliminate, the bias, says **Kristian Lum**.

Growing public awareness of race- and class-based inequities in mass incarceration have led to urgent calls for criminal justice reform in the United States. One potential driver of such inequalities is demographically disparate treatment by the courts. Under the traditional system, decisions that have profound effects on the defendant's liberty and quality of life — including whether and how high to set bail, conditions of release, and sentencing — have been determined largely at the discretion of the judge. This leaves these decisions vulnerable to even the most well-intended judge's implicit biases. To combat the inevitability of human decision-makers introducing their own biases into the decision-making process, court systems are adopting machine-learning tools to advise and inform judicial decision-making.

Underlying the belief that machine-learning tools will create an equitable, objective, and fair criminal justice system is the assumption that data represents truth and that machine-learning models — because they are based on data — will necessarily make objective and fair predictions. This is simply not accurate.

Data that measures human behavioural attributes and social phenomena are the product of a complex interaction between those attributes or phenomena and the process by which they are recorded. If the recording process allows for human discretion in deciding which information to record and how to record it, the data ceases to be an objective measure and instead encodes the interaction between the attribute or phenomenon of interest and the human recorder's perception or knowledge of it. The data itself then reflects the human biases introduced by the collection process.

For example, consider arrest data. This data is often used as a measure of an individual's criminality — individuals with more arrest records are considered to be more criminal than those with fewer. Whether an individual is arrested is the result of actions taken by many human decision-makers, each with their own biases: victims and witnesses (via their decision to

report the crime), police leadership (via their strategic deployment decisions), and patrol officers (via their on-the-ground decisions regarding whom to investigate and arrest, and the charges under which the arrested person will be booked).

Take the specific example of patrol officers. Patrol officers have wide discretion regarding which individuals to investigate on their patrol, whether a suspect's observed behaviour constitutes a crime, and how that crime should be recorded. This offers an opportunity for the officer's biases to seep into the data. If an officer is (even subconsciously) more likely to investigate an individual who fits their preconceptions about the demographic characteristics of a criminal, people who fit this description will be more likely to appear in the arrest data than people with the same level of criminal behaviour who do not fit this profile. Similarly, if an officer's perception of the severity of the crime is (even subconsciously) influenced by the demographic characteristics of the suspect, this bias may be reflected in the booking charge in the form of more serious charges against certain defendants, even for the same criminal behaviour.

Risk assessment models will only be as objective as the data used to train them.

In an effort to reduce the subjectivity of judicial decision-making and bring about a more evidence-based process, many courts are turning to 'risk assessment' tools that use information about a defendant (for example, criminal history, age, booked charges, and so on) to calculate a predicted risk score that advises the judge on, among other things, the defendant's risk of future criminal behaviour. The formula used to calculate this risk score is typically derived, at least in part, from a machine-learning model that is trained using re-arrest as an outcome variable. Machine-learning algorithms learn patterns in the data they are given and

inherit the biases encoded in their training data. If the training data is generated as the product of decisions that are vulnerable to human bias, the models and their predictions will be similarly vulnerable and subject to the same bias. Therefore, because re-arrest is a statistically biased measure of criminality, machine-learning-derived predictions of re-arrest will be biased predictions of future criminality.

Further, in some cases one of the inputs to the risk score formula is the booking charge. Because this is largely determined at the discretion of the arresting officer, this offers additional opportunities for human bias to enter into the equation and systematically and subtly tip the scales against demographic groups (consciously or not) perceived to be more criminal. In trying to rid the system of human bias, these tools trade the potentially biased decision-making of one individual (the judge) for that of another (the police), now filtered through an automated process that gives the risk score prediction a veneer of objectivity and, ostensibly, the backing of science.

Despite all of this, incorporating risk scores derived from arrest data into the decision-making process may be preferable to that which results from judicial decisions that are not informed by such data. Even if imperfect, risk assessment tools may move us towards better and more just outcomes for some. However, we have to recognize the limitations of the tools. Risk assessment models will only be as objective as the data used to train them. If we continue to increase our reliance on these tools without first addressing the disparate policing of poor and minority communities and other root causes of statistical bias in the training data, we risk automating the type of human bias that these tools were developed to eliminate. □

*Kristian Lum is lead statistician at the Human Rights Data Analysis Group, 109 Bartlett Street Suite 204, San Francisco, California 94110, USA.
e-mail: kl@hrdag.org*

Competing interests

The author declares unpaid advice to the New York Legal Aid Society and the San Francisco Public Defender's Office.