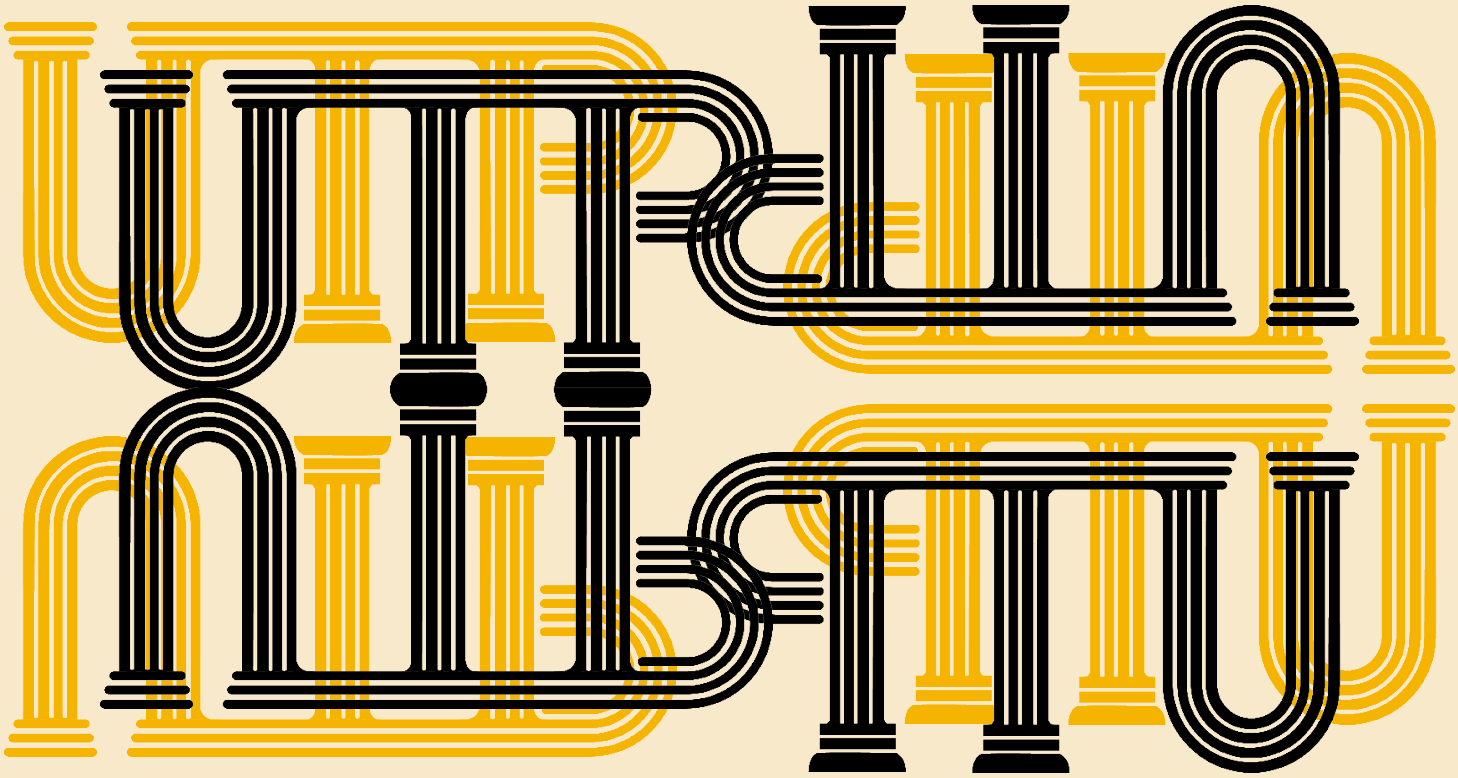
UOFT POLICYTHON
FEBRUARY 2024



ENCODING JUSTICE: PREDICTIVE ALGORITHMS AND PREDICTIVE JUSTICE IN THE CRIMINAL JUSTICE SYSTEM

NATALIE OULIKHANIAN

ENCODING JUSTICE: PREDICTIVE ALGORITHMS AND PREDICTIVE JUSTICE IN THE CRIMINAL JUSTICE SYSTEM

Decisions concerning criminal justice have historically been reserved for humans: their innate complexity and high-stakes outcomes have severely limited the exploration of alternative forms of decision-making. However, as machines have grown exceedingly proficient in some of the tasks society has historically deemed too complex for anything short of a human mind, the narrative surrounding their use in the criminal justice system has started to change dramatically. Legal systems around the world have adopted the use of algorithms, especially those involving 'AI,' to aid in their most complex judicial decisions by automation. However, while the application of AI in judicial decisions strives to resolve the most pressing issue inherent in human adjudication—cognitive bias—many of these technologies instead inadvertently hinder the legal rights of defendants. By reflecting historical bias in their decisions and limiting the ability for defendants to access the rationale behind a decision, many risk being deprived of their rights to due process and freedom from discrimination.

Prominent in criminal justice today are recidivism prediction tools that use AI which despite having been implemented in several countries are especially prominent in the United States. As a country with a higher prison inmate population than any other in the world, the assessment of recidivism, the likelihood a defendant will reoffend, has been a particularly crucial metric for American judges (Hao 2019). America has adopted an algorithmic tool called the Correctional Offender Management Profiling for Alternative Solutions (COMPAS) to help. This algorithm outputs a score based on characteristics like a convict's gender, age, and criminal record to gauge the likelihood that they will reoffend (Dressel et al. 2021). A defendant's risk of recidivism acts as a direct proxy for a range of decisions surrounding probation, surveillance, sentencing, or rehabilitation. Consequently, any failure to accurately evaluate this risk has impacts for several aspects of a defendant's life. Although the use of COMPAS ideally produces objective and unbiased predictions of recidivism, the technology is not considered any more accurate than human judgement from individuals with little to no criminal understanding, often rendering arbitrary and discriminatory decisions (Dressel et al. 2018).

Unlike other applications of predictive models any degree of objectivity is an especially challenging goal for recidivism predictors. The root of COMPAS' inaccuracy derives from the same reason that predictive models in different fields have reached near total accuracy: its relationship with its training data. While the employment of large-scale data analysis may lead one to assume a degree of impartiality, predictions are only reflections of the data upon which they rely. If the data

input is not unbiased to begin with, the result of an algorithm is predestined to be biased in some way. For fields like image detection or speech recognition, where algorithms have shown extremely high rates of accuracy, training data used have much more straightforward connections to the intended outcome. As a result, they are able to effectively recognise new patterns unexplored by humans and form quick and accurate connections between input and outcome. However, recidivism is uniquely a human issue, where perceptions and behavior surrounding crime are constantly in flux and many factors are a matter of individual interpretation.

More specifically, training data for recidivism algorithms is deeply historical: it reflects decisions and attitudes relevant to the people or time from which its data is collected. If data is inherently subjective, it can hardly be used as an effective source of prediction in this context. Relying on historical data, it is exceptionally difficult for an algorithm like COMPAS to generate a standard for the causes of recidivism that would eliminate the errors already endemic to humans. While outputs like the number of arrests may appear to demonstrate correlations between the attributes of someone arrested and their likelihood of recidivism, it is impossible to ever know if the connection is accurate: the number of arrests of people with a particular attribute can be biased and can change depending on differences in how closely a particular group is surveilled and policed, and what socio-economic conditions motivate its members' behavior. It is insufficient to rely on these outcomes and contrast them with defendant attributes since any outcome will ultimately reflect the political attitudes and decisions of its time. Consequently, most decisions made by the algorithm are fixated on certain time frames and contexts not on the basis of their relevance, but rather the accessibility of data relevant to them.

The influence of biased data is already evident in decisions regarding marginalized communities; COMPAS' data has exhibited racial bias, reflecting the higher rates of negative outcomes, like arrests, that racialized groups have faced historically. As a prime example, one study showed that Black Americans are nearly twice as likely as their White counterparts to be classified as high-risk by COMPAS, however are not proportionally likely to reoffend (Angwin et al. 2016). Although the algorithm excludes discriminatory variables like race, the algorithm can still form correlations with other criteria that are included, like socioeconomic status, to produce discriminatory predictions (Starr 2014). Eager applications of recidivism prediction technologies therefore risk reviving or exacerbating old prejudices in the criminal justice system. If COMPAS remains uncontrolled, the algorithm can continue to amplify and perpetuate embedded biases, hurting communities and potentially generating more bias-tainted data to fuel its inaccuracies.

Furthermore, while the current

COMPAS algorithym is dangerously inaccurate, adjustments to accomodate biased data would not necessarily equate to an increase in equity. In criminal justice, both defendants and judges are entitled to the total rationale behind a judicial decision, however, the COMPAS algorithm, as it currently stands, is entirely proprietary. This means that information about how the algorithm works, including the specifics of what data is being used or how variables are weighted, is entirely inaccessible to those outside of its development circle (Rudin, Wang, and Coker 2020). Although general information about what data COMPAS collects from defendants is publicly available, the specific way these inputs are treated in the algorithm's internal workings is entirely hidden.

Justice demands more than accuracy from these algorithms; it requires a commitment to upholding the public's right to a trustworthy institution that is transparent, accountable, and fair in its decisions. The legal system has been considered by many scholars a system rooted in prediction: laws and judicial decisions are derived from assumptions about how individuals and entities will act. All are measures to either promote or restrict a predicted course of action. The adoption of prediction tools, like those excelling in other industries, is therefore no surprise in the world of criminal law. However, with the visible increase of inaccuracies made by algorithms in the criminaljustice system, like COMPAS in the United States, addressing prediction cannot be treated with the same solution everywhere. The eagerness to apply technological solutions that excel in different cases into new fields like law has let significant room for preventable human right violations. While it is certain the legal sphere is expected to be entirely transformed, any change must be introduced with consideration of what has insulated the criminal justice system from other technological changes in the past: the complexity of human behaviour and politics.

REFERENCES

Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. "Machine Bias." ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Dressel, Julia, and Hany Farid. (2018). "The Accuracy, Fairness, and Limits of Predicting Recidivism." Science Advances 4. https://doi.org/10.1126/sciadv.aao5580.

Dressel, Julia, and Hany Farid. (2021). "The Dangers of Risk Prediction in the Criminal Justice System." MIT Case Studies in Social and Ethical Responsibilities of Computing. https://doi.org/10.21428/2c646de5.f5896f9f.

Hao, Karen. 2019. "AI Is Sending People to Jail-and Getting It Wrong." MIT TechnologyReview. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/.

Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. "The Age of Secrecy and Unfairness in Recidivism Prediction." 2.1 2, no. 1. https://doi.org/10.1162/99608f92.6ed64b30.

Starr, Sonja B. 2014. "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination." Stanford Law Review 66, no. 4: 803–72. https://www.jstor.org/stable/24246717.org/10.1162/99608f92.6ed64b30.