Understanding risk assessment instruments in criminal justice

Alex Chohlas-Wood Friday, June 19, 2020

Algorithmic tools are in widespread use across the criminal justice system today. Predictive policing algorithms, including PredPol and HunchLab, inform police deployment with estimates of where crime is most likely to occur.[1] Patternizr is a pattern recognition tool at the New York Police Department that helps detectives automatically discover related crimes.[2] Police departments also use facial recognition software to identify possible suspects from video footage.[3] District attorneys in Chicago and New York have leveraged predictive models to focus prosecution efforts on high-risk individuals.[4] In San Francisco, the district attorney uses an algorithm that obscures race information from case materials to reduce bias in charging decisions.[5] [6]

Risk assessment instruments

One class of algorithmic tools, called risk assessment instruments (RAIs), are designed to predict a defendant's future risk for misconduct. These

predictions inform high-stakes judicial decisions, such as whether to incarcerate an individual before their trial. For example, an RAI called the Public Safety Assessment (PSA) considers an individual's age and history of misconduct, along with other factors, to produce three different risk scores: the risk that they will be convicted for any new crime, the risk that they will be convicted for a new violent crime, and the risk that they will fail to appear in court.[7] A decision-making framework translates these risk scores into release-condition recommendations, with higher risk scores corresponding to stricter release conditions. Judges can disregard these recommendations if they seem too strict or too lax. Other RAIs influence a wide variety of judicial decisions, including sentencing decisions and probation and parole requirements.

Algorithmic RAIs have the potential to bring consistency, accuracy, and transparency to judicial decisions. For example, Jung et al. simulated the use of a simple checklist-style RAI that only considered the age of the defendant and their number of prior failures to appear.[8] The authors noted that judges in an undisclosed jurisdiction had widely varying release rates (from roughly 50% to almost 90% of individuals released). The authors found that if judges had used their proposed checklist-style model to determine pretrial release, decisions would have been more consistent across cases, and they would have detained 30% fewer defendants overall without a corresponding rise in pretrial misconduct. Other studies have found additional evidence that statistical models consistently outperform unaided human decisions.[9] In contrast to the opacity of traditional human decision-making, the transparent nature of a checklist-style model, like the one proposed by Jung et al., would also allow courts to openly describe how they calculate risk.[10] These benefits—along with a general belief that important decisions should be anchored in data—have compelled many jurisdictions across the country to implement RAIs.

The COMPAS RAI

In parallel with their expansion across the country, RAIs have also become increasingly controversial. Critics have focused on four main concerns with RAIs: their lack of individualization, absence of transparency under trade-secret claims, possibility of bias, and questions of their true impact.[11] A 2016 Wisconsin Supreme Court case, *Loomis v. Wisconsin*, grappled with many of these issues. The petitioner, Eric Loomis, made several arguments against the use of an RAI called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) in his sentencing decision.[12]

First, Loomis contended that his sentence was not individualized. Instead, he claimed it was informed by historical group tendencies for misconduct, as assessed by COMPAS. The court disagreed, arguing that the judge's decision was not solely determined by COMPAS, avoiding Loomis' individualization concerns. Although the court made this distinction, it is worth noting that both humans and algorithms learn from historical behavior. A risk prediction for a given individual—whether from a judge or an RAI—is, as a result, anchored in the historical behavior of similar individuals.

Second, Loomis argued that the company that created COMPAS declined to release enough details on how the algorithm calculated his risk score, preventing him from scrutinizing the accuracy of all information presented at his sentencing. Many RAIs can explain exactly how they arrive at their decisions, an advantage over traditional human decision-making. However, commercial vendors that sell RAIs often hide these details behind trade-secret claims.[13] While the court did not strictly agree with Loomis—arguing that it was sufficient to observe the inputs and outputs of COMPAS—there are compelling reasons for transparency and interpretability in such high-stakes contexts.

For example, although Loomis did not know the full structure of the model, he knew that it incorporated gender as a factor, and he argued that this was discrimination. The court disagreed, emphasizing that including gender in the model helped increase its accuracy. This follows the fact that, given

similar criminal histories, recidivism rates are statistically lower for women than for men.[14] Either way, Loomis' knowledge of the model's use of gender allowed him to challenge its inclusion, an example of how transparency in RAIs can help stakeholders better understand this high-stakes decision-making process.

Potential discrimination and RAI problems

Other charges of discrimination have been levied against RAIs (and machine learning algorithms in general), noting that they can perpetuate and exacerbate existing biases in the criminal justice system.[15] Perhaps the most notable claim appeared in a 2016 ProPublica article about the use of COMPAS alongside pretrial detention decisions in Broward County, Florida.[16] The article concluded that COMPAS was biased because it performed worse on one measure of performance (false positive rates) for Black individuals when compared to white individuals. However, other researchers have noted a substantial statistical flaw in ProPublica's findings: They can be mathematically explained by differences in underlying offense rates for each race without requiring a biased model.[17] When researchers apply a traditional measure of model fairness—whether individuals with the same risk score re-offend at the same rate, regardless of race—evidence of racial discrimination disappears.[18]

Even still, a lack of evidence does not guarantee that discrimination is absent, and these claims should be taken seriously. One of the most concerning possible sources of bias can come from the historical outcomes that an RAI learns to predict. If these outcomes are the product of unfair practices, it is possible that any derivative model will learn to replicate them, rather than predict the true underlying risk for misconduct. For example, though race groups have been estimated to consume marijuana at roughly equal rates, Black Americans have historically been convicted for marijuana possession at higher rates.[19] A model that learns to predict convictions for marijuana possession from these historical records would unfairly rate Black

Americans as higher risk, even though true underlying rates of use are the same across race groups. Careful selection of outcomes which reflect true underlying crime rates may avoid this issue. For example, a model that predicts convictions for violent crime is less likely to be biased, because convictions for violent crime appear to mirror true underlying rates of victimization.[20]

> *"[A] lack of evidence does not guarantee that discrimination is absent, and these claims should be taken seriously."*

Many would argue that a pure focus on algorithmic behavior is too limited; that the more important question is how RAIs influence judicial decisions in practice, including any difference in impacts by race. To illustrate this point, it is useful to think of two possible extremes. We may not be as concerned about an inaccurate RAI if it is categorically ignored by judges and does not affect their behavior. On the other hand, a perfectly fair RAI may be cause for concern if it is selectively used by judges to justify punitive treatment for communities of color.

Though many studies have simulated the impact of RAIs, research on their real-world use is limited. A study of RAIs in Virginia between 2012–2014 suggests that pretrial misconduct and incarceration can both be reduced at the same time.[21] Another study examined the 2014 implementation of a PSA in Mecklenburg County, North Carolina, and found that its implementation coincided with higher release rates, while rates of pretrial misconduct went unchanged.[22] A third study scrutinized the implementation of RAIs across Kentucky between 2009–2016, finding limited evidence that the tool reduced incarceration rates.[23] The study did find that a judge's use of an RAI did not unevenly impact outcomes across race groups.

Recommendations

Anybody, including executive, planning, management, analysis, and software development staff, considering the use of algorithms in criminal justice—or any impactful context more broadly—should heed these concerns when planning policies that leverage algorithms, particularly those steering criminal justice decisions.

First, policymakers should preserve human oversight and careful discretion when implementing machine learning algorithms. In the context of RAIs, it is always possible that unusual factors could affect an individual's likelihood of misconduct. As a result, a judge must retain the ability to overrule an RAI's recommendations, even though this discretion may reduce accuracy and consistency. One way to balance these competing priorities is to require a detailed explanation anytime a judge deviates from an RAI recommendation. This would encourage judges to consciously motivate their decision and would discourage arbitrary deviations from an RAI's recommendations. In general, humans should always make the final decision, with any deviations requiring an explanation and some effort by the judge.

> *"[P]olicymakers should preserve human oversight and careful discretion when implementing machine learning algorithms."*

Second, any algorithm used in a high-stakes policy context, such as criminal sentencing, should be transparent. This ensures that any interested party can understand exactly how a risk determination is made, a distinct advantage over human decision-making processes. In this way, transparency can help establish trust, and is an acknowledgement of the role these tools play in consequential, impactful decisions.

Third, algorithms, and the data used to generate their predictions, should be

carefully examined for the potential that any group would be unfairly harmed by the outputs. Judges, prosecutors, and data scientists should critically examine each element of data provided to an algorithm—particularly the predicted outcomes—to understand if these data are biased against any community. In addition, model predictions should be tested to ensure that individuals with similar risk scores reoffend at similar rates. Finally, the use of interpretable models can help demonstrate that the scores generated by each model appear to be fair, and largely conform to domain expertise about what constitutes risk.

Fourth, data scientists should work to build next-generation risk algorithms that predict reductions in risk caused by supportive interventions. For example, current RAIs only infer the risk of misconduct if an individual is released without support. They do not consider the influence of supportive interventions—such as court-date text-message reminders—even though they may have a tampering effect on an individual's risk for misconduct. Imagine an individual who is predicted by a traditional RAI to have a low likelihood of court appearance if they are released without support. With only this rating, a judge would likely choose to incarcerate the individual to ensure they appear in court. However, with next-generation RAIs, a judge might also see that text-message reminders substantially increase the likelihood of the individual's appearance. With this additional information, the judge may instead choose to release the individual and enroll them in reminders. Next-generation risk algorithms that estimate the impact of supportive interventions could encourage judges and other decision-makers to avoid the considerable social and financial costs of punitive action in favor of more humane alternatives.

Finally—and perhaps most important—algorithms should be evaluated as they are implemented. It is possible that participants in any complicated system will react in unexpected ways to a new policy (e.g., by selectively using RAI predictions to penalize communities of color). Given this risk, policymakers should carefully monitor behavior and outcomes as each new

algorithm is introduced and should continue routine monitoring once a program is established to understand longer-term effects. These studies will ultimately be key in assessing whether algorithmic innovations generate the impacts they aspire to achieve.

RAIs are only one algorithmic tool in consideration today. Separate challenges surround the use of other algorithms. Most notably, criminal justice agencies must explain how they plan to protect individual privacy and liberty in their use of facial recognition, public DNA databases, and other new forms of surveillance. But if used appropriately and carefully, algorithms can substantially improve impactful decisions, making them more consistent and transparent to any interested stakeholder. As with any new policy or practice, these efforts must include continued evaluation and improvement to ensure that their adoption generates effective and fair outcomes over time.