ARTIFICIAL INTELLIGENCE AND MASS INCARCERATION

Leah Rosenbloom

The Workshop School (USA)

leah.rosenbloom@gmail.com

ABSTRACT

Artificial intelligence (AI) is now common throughout the criminal justice system. Police use predictive algorithms to target locations and individuals for surveillance. Judges use risk assessment algorithms to determine whether defendants should be granted bail or parole. Prosecutors use the results of forensic analysis algorithms to accuse and convict defendants of crimes, including those punishable by death. These algorithms are considered intellectual property and are closed off from public scrutiny.

In this paper, we explore the impact of AI on mass incarceration. A comprehensive survey of existing practice reveals the ways in which algorithms perpetuate systemic injustice and violate defendants' legal rights. We argue the need for solutions that integrate technical and legal perspectives, including novel ways to shift the focus of the algorithms from punitive to restorative practices. With due oversight, transparency, and collaboration between experts in technology, law, and government, we can leverage existing algorithms to combat systemic injustice.

KEYWORDS: artificial intelligence, predictive policing, risk assessment, machine testimony, restorative justice, technology and the law.

1. INTRODUCTION

While existing literature concerning algorithms in criminal justice applications is extensive, research is scattered between the scientific and legal communities. In order to form a complete picture of the impact of AI on mass incarceration, which touches problems in machine learning, data science, law, and governance, it is necessary to integrate these perspectives. To that end, this paper explores sources, issues, and proposed solutions in each area of research.

Several ethical concerns emerge from the survey of existing literature. First is the issue of data that reflect existing racial and socio-economic bias in the criminal justice system. Data science experts have proposed statistical models that remove racial bias from the data, which leads us to consider the implications of "objective" black-box algorithms operating within contexts of deeply entrenched bias. We argue that the use of black-box solutions to racial discrimination encourages law enforcement and judiciaries to defer their responsibilities, preferring automatic arrests and convictions over critical consideration.

Without transparency and oversight, it is impossible to examine the underlying mechanisms that process and objectify the data. Furthermore, even if the underlying data is scrubbed clean, the "correctness" metrics and implementation of the algorithms may still reflect systemic bias. Comprehensive solutions to problems with algorithms in criminal proceedings must include technical, practical, and legal components. We introduce novel, integrated analyses for each application of artificial intelligence in the criminal justice system: predictive policing, risk assessment, and machine testimony. Our conclusion is a call to action: technologists, legal experts, and government officials alike have a responsibility to face—and hopefully fix—these issues.

2. MACHINE LEARNING

Artificial intelligence can be reduced to a machine's ability to learn (Russell and Norvig, 1995). Machine learning is defined as the ability to process input data such that the categorization of new data is correct to some degree of approximation. While a specific analysis of closed-source algorithms is regrettably impossible, all machine learning algorithms must necessarily "learn" from pre-existing data. Therefore, we can use our understanding of the data to draw conclusions about the effectiveness of these algorithms in practice.

2.1. How machine learning works

Learning algorithms break down into two stages: a learning or "training" stage, and an extrapolation stage. In the learning stage, the machine classifies human-configured data into categories, which are either human or machine-generated. We can visualize this process as points in space, where data with similar contexts and outcomes are plotted closer together. The machine then draws boundaries around close clusters of points, segmenting the space into regions. In the case of predictive policing, the algorithms classify existing data on policing. This might include the type and location of reported crimes, existing patrol routes and routines, or background information on individual suspects and arrests. If the goal of the algorithm is to locate future crimes, the algorithm might draw boundaries around groups of points with high crime rates.

Once the machine has processed the training data, it is ready to classify new data in the extrapolation stage. It plots the new point into the existing space, and the point takes on the outcome of the spatial region in which it is plotted. For example, in the case of risk assessment, the new data point is a new defendant who may share similar attributes and criminal history with existing defendants. The algorithm places the defendant among defendants with similar histories, and projects the new defendant's risk based on the previous defendants' outcomes.

2.2. Measuring correctness

The accuracy of the algorithm in the extrapolation stage reduces to the correctness of the boundaries drawn around the initial training data set. The more voluminous, diverse, and correctly classified the training data set, the better the algorithm will do. If the training data are malformed—if points are incorrectly classified or there are not enough points in a particular area—the classification of new data can be incorrect or unpredictable. For human-verifiable problems like image recognition, it is possible to run the algorithm on new data in the extrapolation stage and directly measure the accuracy of the results.

Correctness is difficult to measure for complex human systems because the input data are generally too large to evaluate on a case-by-case basis, and they are full of error and inconsistency. For example, DNA samples can be easily contaminated, intermixed, and degraded (DiFonzo, 2005). Algorithms built on millions of these inconsistent samples are similarly inconsistent. Unlike image classification, it is not trivial for a human to step in and verify the correctness of a DNA match; analysts cannot comb through millions of samples to verify their reliability, nor go back in time and preserve evidence from the crime scene.

2.3. Existing criminal justice practices are not correct

Logically we can expect that any errors, inconsistencies, and biases in the underlying training data will carry over into the algorithms. Communities that are already hyper-targeted by law enforcement will be similarly targeted by predictive policing. Risk assessment algorithms will work better for defendants who have been treated fairly in the past. Machine analysis of forensic evidence will only yield correct results if similar evidence has been impeccably collected, stored, and analysed. We know from existing problems in all of these areas that the training data is far from accurate and unbiased. Law enforcement has a serious and long-standing problem with racist policing (Langan, 1995; Alexander, 2010; Lum & Isaac, 2016). Judges are known to make racially biased decisions about bail, sentencing, and parole (Johndrow & Lum, 2017; Goel et al., 2018). Forensic evidence is often contaminated, and analysts are known to make mistakes and collude with prosecutors to guarantee convictions (DiFonzo, 2005; Shaer, 2016; Mettler, 2017). Rather than acknowledging and examining these issues, law enforcement and legal systems continue to plow forward with the integration of machine learning into criminal justice proceedings (Mohler et al., 2015; Danner et al., 2016; Saunders et al, 2016; Kaufman et al., 2017; Winston, 2018; Human Rights Watch, 2018). The result, as we will discuss in the rest of our paper, is the blind perpetuation of injustice.

3. PREDICTIVE POLICING

The U.S. National Institute of Justice (NIJ) describes predictive policing as a law enforcement approach that “leverages computer models…for law enforcement purposes, namely anticipating likely crime events and informing actions to prevent crime” (2014). These computer models are trained on existing reports of criminal and police activity. One of the most widely-used algorithms, PredPol, uses only the “three most objective data points” of time, location, and type of previously-reported crime in each precinct’s regional area (PredPol, 2020). Others, like Chicago’s “Strategic Subjects List”, focus on identifying groups and individuals (Saunders et al., 2016). The NIJ confirms that predictive algorithms can focus on “places, people, groups, or incidents” (NIJ, 2014). Each of these models has been shown to perpetuate existing racial and socio-economic bias (Saunders et al., 2016; Lum & Isaac, 2016).

3.1. In practice, predictive policing gets personal

The Chicago Police Department (CPD) uses predictive policing to curate a “heat list” of people who are likely to be involved in violent gun crime, either as victims or perpetrators. The 2013 pilot program, which was funded by the NIJ, used an algorithm on “co-arrest networks” along with human intelligence to produce a “Strategic Subjects List” (SSL) of 426 high-risk individuals.

Saunders et al. evaluated the pilot in 2016 and noted these individuals were "not necessarily under official criminal justice supervision nor were they identified through intelligence to be particularly criminally active" (p. 349).

The list was disseminated to commanders in each police district, who decided on an individual basis how and when to use the list to inform practice. In 10 of 22 districts (45.4%), officers made contact with named individuals only if they spotted one acting suspiciously. In 7 of 22 districts (31.6%), officers made regular visits to named individuals' homes. In the remaining 5 of 22 districts (22.6%), officers used a combination approach. Otherwise, there was "no practical direction about what to do with individuals on the SSL" once they were contacted (Saunders et al., 2016, p. 356).

The study found that the pilot had no statistically significant effect on the homicide rates in Chicago (Saunders et al., 2016, p. 361). Out of 405 total homicide victims between 2013 and 2014, the pilot identified only three (0.74%). During the same time period, police made contact with almost 90% of the individuals on the list with an average of 10.72 interactions each, a 39% increase over a matched control group (p. 363). While increased police contact did not affect an individual's likelihood of being arrested for a shooting, individuals on the SSL were 2.88 times more likely to be arrested for a shooting (p. 362-363). The CPD explained this phenomenon to the study's authors by admitting that the list was used to come up with suspects for unsolved shootings (p. 365).

3.2. The positive feedback loop of racially biased policing

A subsequent study by Lum and Isaac in 2016 begins with a grim anecdotal account of the Chicago pilot. A CPD commander visits the home of a 22-year-old black man on the South Side of Chicago. The commander tells him, this is a warning: you'd better not commit any more crimes. The man is confused. He is not involved in crime. He is on a heat list.

Black and Brown people have been disproportionately targeted for arrest, incarceration, and police brutality in the United States going back hundreds of years (Alexander, 2010). One of the most well-studied examples of racial discrimination in policing is the War on Drugs. A Department of Justice report published in 1995 found that while only 16% of black people reported selling drugs, they accounted for 49% of drug distribution arrests, with similar numbers of discrepancy for drug possession (Langan, 1995, p. 3). Lum and Isaac demonstrated a similar discrepancy using data from the 2011 National Survey on Drug Use and Health. They compared the demographics of arrest records in Oakland, California to the estimated demographics of drug users in Oakland according to the survey. What they found is that while drug use was roughly evenly distributed over the population, low-income and minority neighborhoods had 200 times more arrests than their middle- and upper-class white counterparts (Lum & Isaac, 2016, p. 17).

Lum and Isaac further simulated what impact the popular predictive policing algorithm PredPol would have had on the Oakland population. PredPol claims that its omission of personal information in the training data "eliminat[es]…profiling concerns" (PredPol, 2020), but the simulation illustrated how PredPol would continue and potentially worsen the disproportionate targeting of minority communities. Most notably, the authors created a positive feedback loop with PredPol's algorithm: the more the algorithm sent police to particular neighborhoods, the more crimes would be reported in those neighborhoods, the more likely the algorithm would be to send police back to those neighborhoods on a repeated basis (Lum & Isaac, 2016, p. 18-19).

Machine learning experts have begun to address PredPol's positive feedback loop. One such study by Ensign et al. proposes filtering input data to obtain a more representative sample (2018). The authors admit, however, that their model does not address the underlying problem of biased reporting and arrests. Rather, they falsely assume that crimes identified by the police are equivalent to true crime rates (Ensign et al., 2018, p.11).

3.3. Algorithms currently facilitate bad practices

There are two fundamental issues with predictive policing. First, policing data reflects institutionalized racism (Langan, 1995; Alexander, 2010; Lum & Isaac, 2016). Another study from the same time period compared Los Angeles districts using predictive algorithms against those using traditional methods. They found no statistically significant difference in the racial and ethnic breakdown of arrests between the two (Mohler et al., 2015). This is not surprising: programmed correctly, an algorithm will echo, but not enhance or diminish, existing racism.

Companies like PredPol selling their algorithms as "objective" solutions to racial bias are misinformed. This perpetuates the existence of "colorblind" racism, whereby people are convinced they are race-blind despite overwhelming evidence to the contrary (Alexander, 2010). Reliance on these algorithms will encourage law enforcement to stop thinking about the problem of biased crime reporting and arrests, even as those same biases continue to dictate their actions. If we were to ask the police commander mentioned above—a figure with authority over an entire district on the South Side of Chicago—why he knocked on that 22-year-old black man's door, would he think critically about his behavior? Would he consider it necessary to provide independent justification? Or would he respond indignantly that the man was flagged by cutting-edge technology?

The second major issue is therefore implementation—what actually happens as a result of the algorithm's predictions. Even if the algorithms really did provide an objective analysis of future criminal activity, police commanders and officers would still be the final arbiters in deciding how the results are applied. The stated purpose of the heat list in Chicago was to deter gun crime, but in reality, the CPD used the list to harass people and produce suspects for open shootings (Danner et al., 2016). A recent civilian audit of the Los Angeles Police Department found that the department's data-driven predictive policing programs "lacked oversight and that officers used inconsistent criteria to label people as 'chronic offenders'" (Puente, 2019). Some of those programs have since been decommissioned.

In order to ensure police are acting responsibly, advocates stress the need for transparency surrounding how and when police are relying on algorithms, and for stricter regulation of predictive policing technology.

3.4. From bad to worse

Perhaps the world's most comprehensive predictive policing system is China's Integrated Joint Operations Platform (IJOP), which is widely deployed in the Xinjiang Uygur Autonomous Region. IJOP is a data-driven system that receives constant, real-time input from the following "sensors":

> CCTV cameras, some of which have facial recognition or infrared capabilities…entertainment venues, supermarkets, schools…"wifi sniffers," which collect the unique identifying addresses of computers, smartphones, and other

> networked devices…license plate numbers and citizen ID card numbers from some of the region's countless security checkpoints and "visitors' management systems" in access-controlled communities. (Human Rights Watch, 2018)

In addition to the sensors, IJOP also receives data on criminal history and prior police contact, purchase history and financial records, family planning, legal records, and religious practices, including whether or not the person is an Uyghur. IJOP then issues a daily forecast to law enforcement, including the names of people to investigate further. Unnamed sources report that IJOP is also able to produce a "round-up" list of people to detain immediately. Some of the people flagged are "detained and sent to extralegal 'political education centers' where they are held indefinitely without charge or trial, and can be subject to abuse" (Human Rights Watch, 2018).

3.5. Algorithms can be used to facilitate good practices

Police contact is associated with negative mental and physical health consequences (Sewell & Jefferson, 2016). PredPol, which claims to "help protect one out of every 33 people in the United States" (PredPol, 2020), is still subject to a positive feedback loop that increases police contact in targeted areas. The "heat list" model, which explicitly targets individuals, was shown to steeply increase police contact with those individuals (Saunders et al., 2016). Communities that already experience high levels of police contact will experience even more contact in areas that employ either model. Increased policing will likely contribute to the further deterioration of community-police relations, and help to perpetuate a cycle of violence, poverty, and crime.

While predictive policing algorithms are currently employed to inform policing, it is possible to use the same algorithms for community healing and restorative justice. These algorithms reveal bias: we can use them to identify communities that are likely to have broken relationships with law enforcement. Police can wield predictive policing algorithms to confirm their bias and decide where to focus patrol and arrests, or they can harness those same algorithms to face their bias and decide where to focus outreach, mediation, and social service referrals. These algorithms are already deeply embedded in global policing systems; rather than work to patch them up or decommission them completely, the path of least resistance and greatest efficacy is to re-purpose them for methods that can repair the harm of dehumanizing police practices (Marshall, 1999). Until we see movement towards restorative justice, predictive policing, and policing in general, will continue to plague communities in need.

4. RISK ASSESSMENT

After someone is arrested, a judge determines the conditions of that person's release. Typically, the judge makes some kind of "risk assessment" to determine how likely the defendant is to commit more crimes. These assessments can influence bail, sentencing, and parole. While a judge gets the final say, risk estimates have been increasingly performed by machine learning algorithms.

Similarly to how predictive policing algorithms run on biased arrest data, risk assessment algorithms run on biased arrest data *and* biased judicial data. The data used in risk assessment, however, is uniquely biased by selective outcome representation; if the defendant in the input data set was not released on bail, there is no way to determine whether or not they would have committed an offense if they had been released. This would suggest that if a particular group was disproportionality arrested and detained, or detained for inconsistent reasons, the algorithm would be more unpredictable and less accurate for that group.

4.1. A case study of selective unpredictability

An evaluation of the Virginia Pretrial Risk Assessment Instrument (VPRAI) found evidence of the algorithm's unpredictability for People of Color (Danner et al., 2016). While they did not find race to be a statistically significant predictor of risk, they did find a statistically significant difference in the predictive ability of the algorithm based on race, "with the model performing better for Whites" (p. 8). The authors attributed this disparity in part to the inclusion of risk factors that could "over-classify the risk" for People of Color, and found that if they "weighted, summed, and collapsed [the risk factors] into risk levels, the difference…is no longer statistically significant" (p. 8). It is unclear what specific operations they performed to "collapse" the risk factors, and whether or not these mitigations are used in practice.

The study also found that weighting certain risk factors led to "overclassifying pretrial failure risk for females" (p. 9). Despite the unexplained unpredictability for marginalized groups, the authors conclude that VPRAI is race and gender neutral (p. 8-9). Similar assessments and a possible explanation for this discrepancy, which we discuss later in this section, were outlined by data scientists Johndrow and Lum (2017).

4.2. Risk assessment algorithms are unregulated and regionally inconsistent

A comprehensive review of risk assessment algorithms suggests that while algorithmic results are not free of data bias, they can be "more accurate and less biased than clinical decision making" (Goel et al., 2018, p. 2). The review cites "extremely vague" legal requirements for risk prediction testimony, which have led to traditional verdicts steeped in judicial bias. These vague requirements were upheld for risk assessment algorithms by the Wisconsin Supreme Court, which rejected the transparency concerns raised in *Wisconsin v. Loomis* (Goel et al., 2018, p. 17).

Even with added accuracy and reduced bias overall, these algorithms pose a threat to fair legal proceedings. Without transparency and regulation, we cannot know where the algorithms work and where they do not. We are left in the dark until the failures surface, at which point people have already been locked up, denied bail, and over-sentenced. The Defender Association of Philadelphia, which represents approximately 70% of the people arrested in Philadelphia, testified in Harrisburg that the Pennsylvania Sentencing Commission's algorithm correctly identifies a "risky" defendant "only 52% of the time"—hardly better than a coin toss (Defender Association of Philadelphia, 2018).

One such incorrect assessment was made for Defender Association Bail Navigator LaTonya Myers, who was incarcerated as a juvenile. Myers spoke at the hearing about being the victim of domestic violence. She described the night she stepped in to defend her mother from her mother's live-in boyfriend, who was assaulting her. Her mother's boyfriend called the police and the police took both women into custody, where they were held in separate cells. The police offered Myers a deal—they would let her and her mother go if Myers agreed to probation. Myers knew she wasn't guilty, but she took the deal anyway because she was scared and alone.

Myers then described her life in and out of the criminal justice system all the way through her twenties, when she became involved in criminal justice advocacy and defense. Even though Myers had long ceased to be of concern to parole officers and was by all human accounts a model citizen, the algorithm still classified her as a "risky" defendant.

Myers cited concerns that these algorithms encourage a judge to "overlook individual circumstances and experiences, and preclude the possibility for personal growth and rehabilitation" (Defender Association of Philadelphia, 2018).

4.3. Removing race from the input data

One proposed solution to algorithmic bias is to eliminate race as a variable. Johndrow and Lum show it is possible to do this by first identifying variables that "encode" for race, then creating a transformed set of variables that are mutually independent of the race-encoding variables (Johndrow & Lum, 2017, p. 3). For example, the new algorithm might notice that People of Color are more likely to be re-arrested for a particular crime, and would adjust re-arrest rates to correct for that bias. The model was found to somewhat equalize the predictive ability of a sample algorithm with respect to race, for an overall predictive accuracy that is close to the same as the unadjusted model (p. 16-17).

In order to justify the use of racially-independent training data, the authors argue that "the most reasonable approach is to treat all races as though they are the same with respect to recidivism" (Johndrow & Lum, 2017 p. 4). Like most risk assessment algorithms, however, the "accuracy" of the model is still defined as the algorithm's ability to predict whether a defendant will be re-arrested. The authors admit that re-arrest is a racially biased measure of criminal behavior, and that People of Color are disproportionately stopped, arrested, and incarcerated (p. 3).

A truly "accurate" risk assessment model would therefore have to predict that People of Color would be at higher risk for re-arrest—not because they are inherently prone to criminal behavior, but because they are disproportionately subjected to police attention and incarceration. An algorithm with re-arrest as its correctness metric can only measure the risk of re-arrest; it cannot measure or quantify a defendant's risk of criminal behavior. The removal of race as an input variable is a step toward colorblind data, but it cannot account for the racial bias inherent in conviction and arrest.

4.4. Rethinking risk

A better standard might be to consider the predictors of recidivism, such as income insecurity, education, mental health, and drug addiction (Makarios et al., 2010). Rather than classifying people as "high risk" for re-arrest, the algorithm might classify people as being "high risk" for unemployment, depression, or relapse. This would help judges make recommendations or choose programs for defendants that benefit them and reduce their long-term risk of recidivism.

5. MACHINE TESTIMONY

Prosecutors have become increasingly reliant on algorithms that classify forensic evidence, especially DNA, to secure convictions. Unlike more traditional methods, where an analyst might compare two forensic samples in a lab, machine learning algorithms allow analysts to compare samples against millions of other samples in a DNA database, and obtain the probabilistic estimates of various matches. Problems with traditional forensics carry over into the millions of samples in DNA databases. For instance, samples can easily become contaminated, intermixed, or decomposed, which leads to false positive matches (DiFonzo, 2005). There is also growing evidence of dishonest actors in crime labs, who have rushed testing, intentionally contaminated samples, faked results, and colluded with prosecutors to guarantee convictions (Shaer, 2016; Mettler, 2017). Rogue actors could further compromise results if they were to exploit flexibility or vulnerability in the algorithm's input parameters.

There is one problem unique to forensic algorithms that poses a grave threat to defendants' right to a fair trial. Traditionally, analysts and experts would be able to testify to each step of forensic analysis in detail. If a particular chemical reagent in a DNA experiment was called into question by the defense, an expert would be able to draw on direct knowledge of the reagent to confirm or refute concerns about its reliability. Forensic analysts that handle biological and chemical samples are understandably educated in biology and chemistry; they are not educated in machine learning, and cannot attest to the reliability of machine learning tests. Moreover, even machine learning experts cannot attest to the reliability of these tests, because the details of the algorithm are obscured behind copyrights and corporate policy. Without the source code, it is impossible for defendants to hear, understand, and question the evidence against them.

5.1. Crime labs are a mess

The scope and volume of problems with crime labs are well summarized in "The Crimes of Crime Labs" (DiFonzo, 2005). While forensic analysis works well under perfect conditions, there are a myriad of real-life conditions that hinder correct analysis. For example, the sample must be adequately sized, isolated, collected, and maintained. This is difficult to achieve in practice due to the mixing of evidence at crime scenes. Once a sufficient sample is obtained, DiFonzo describes the analysis itself as "slapdash…often performed by untrained, underpaid, overworked forensic technicians" (DiFonzo, 2005, p. 2). He cites a lack of oversight on education, certification, and lab accreditation. Mishandling and incorrect classification of historical samples undoubtedly influences the accuracy of any machine learning algorithm trained on those samples.

Even when technicians are educated and proficient, crime labs as institutions are often closely associated with police departments and prosecutors. They have a history of dishonesty and corruption, from faking the results of drug tests (Mettler, 2017) to compensating labs for DNA analysis that ends in conviction (Shaer, 2016). These bad-faith convictions would have an even worse result on training data: whereas the honest mishandling of evidence might cause the algorithm to behave erratically, the dishonest mishandling of evidence could bias the algorithm towards false positives, continuing the cycle of unjust conviction.

DiFonzo also highlights long-standing issues with the gross misrepresentation of statistical evidence in court, citing in particular an example in which the prosecution claimed the probability of a match between crime scene DNA and DNA from database subjects was one in 694,000, when in reality it was independently determined to be one in eight (DiFonzo, 2005, p. 5). This false claim and many others led to false convictions, and the crime lab responsible was subsequently shut down. There have been hundreds of similar cases, including "'perjury by expert witnesses, faked laboratory reports, and testimony based on unproven techniques'" (p.

5). The public is largely unaware of the unreliability of DNA testing and analysis, which becomes crucially important in cases where a DNA match is the only piece of accusatory evidence.

5.2. Forensic algorithms are a mess inside of a black box

The public is even less aware of the reliability of algorithms that perform forensic analysis, and the results are easier to obfuscate. In the United States, the use of closed-source software in criminal trials potentially violates several laws. One such law, the Confrontation Clause of the Sixth Amendment of the Constitution, guarantees defendants the right to "be confronted with the witnesses against him" (U.S. Const. amend. VI). In support of the defendant in *California v. Johnson*, attorneys at the American Civil Liberties Union (ACLU) argued that the "witness" to accusatory DNA evidence included the designers of DNA analysis software TrueAllele, the system's programmers, and the code itself (Kaufman et al., 2017, p. 21). Failure to produce the code—the specific procedure by which the DNA was matched—was therefore failure to produce a complete witness for the defense to confront.

Black-box witness testimony calls into question the overall "fairness" of the trial. The Due Process clause of the Fourteenth Amendment has been interpreted to include the defendant's right to perform "adversarial testing" on any evidence—that is, both sides should be able to examine the evidence and reach an independent conclusion (Kaufman et al., 2017, p. 21). In the case of an algorithm, adversarial testing might include tweaking input data and parameters to reveal bias or inconsistency, verifying the legitimacy of each function, and documenting bugs or vulnerabilities. All of these tests are impossible to perform without the code itself.

5.3. Criminal justice proceedings should not be hidden from public scrutiny

The general public has a right to "observe and evaluate the workings of the criminal justice system" (Kaufman et al., 2017, p. 21). In the United States, this is guaranteed by the First Amendment, which includes the right of the public to "petition the government for a redress of grievances" (U.S. Const. amend. I). Logically and according to Constitutional interpretation, this gives us the right to identify and confront our grievances before we suggest solutions. In a stand-alone paragraph, ACLU attorneys summarize the legitimate demand for civil rights in the digital age: "Algorithms used to produce evidence introduced to prove the guilt of a criminal defendant fit well within the broad reach of the First Amendment right of access" (Kaufman et al., 2017, p. 32). Many countries have laws that require the transparency of criminal proceedings. Until source code access is granted, people all over the world will be subject to illegal black-box convictions.

Regardless of legal basis, let us briefly consider the "fairness" of a trial that allows black-box evidence. One party, the government, has a great deal of power and resources, including this mysterious black box. The other party, the defendant, has relatively little power, few resources, and no way to understand what is in the government's secret box. The government pulls a name out of the box and declares the defendant guilty, the punishment for which can include death. Government prosecutors and Silicon Valley giants would have us believe this trial is "fair", and that the conviction is reasonable without a shadow of a doubt. In actuality, this black-box standard of evidence echoes courtrooms of the Dark Ages.

6. CONCLUSION

As we continue to integrate "intelligent" machines into society, it is worth considering what it is we actually want from these machines. Algorithms are currently employed to aggressively digest and maintain the status quo. Despite what proponents say, they do not offer any real solutions to the underlying problems of systemic bias and corruption. While the algorithms themselves are similarly unlikely to worsen bias, they create a facade of objectivity and fairness that discourages people from facing reality. Still, we believe it is possible to use these algorithms to change society for the better. We propose the following changes to the current use of machine learning algorithms in the criminal justice system.

1. **Use machine learning as a tool to understand systemic bias.** Instead of striving to remove bias in data, effectively obscuring and ignoring the root causes, we could use it to better understand the root causes. We will not be able to effectively address systemic racism until we can perform an intersectional analysis of where and how it occurs. Machine learning can help us do that.

2. **Shift the focus of implementation of machine learning results from punitive to restorative practices.** Once we determine where the bias is, we can start to address it with practices that heal, rather than practices that divide.

3. **Law enforcement, crime labs, and courtrooms should create positions for people who understand machine learning.** We need experts to help determine the accuracy and reliability of the algorithms currently used throughout the criminal justice system. These experts could also curate input data, testify clearly about the inner-workings of the algorithms in court, and act as a liaison between criminal justice offices and tech companies.

4. **Machine learning and data science researchers and developers should take responsibility for the impact of their creations.** Research and development is advancing at an unprecedented rate. It is up to the technical experts to help explain and evaluate the impact of new technology on human systems. This help could include application testing and analysis, research collaboration with legal experts, and recommendations for government policy.

5. **Government officials should regulate the use of artificial intelligence in the criminal justice setting.** We cannot count on tech companies to regulate themselves. We need transparency and government oversight in order to fully understand what is happening inside of these algorithms. This is necessary for everyone, but especially for defendants, who are being arrested, convicted, and sentenced to prison based on evidence they can neither see nor contend with.

By collectively shifting our frame of reference on machine learning in criminal justice applications, we can work toward addressing the problem of systemic injustice rather than perpetuating or ignoring it. Each stakeholder has an important role, and it will take all of us to create a better future.

ACKNOWLEDGMENTS

Thank you to Vera Eidelman, Andrea Woods, and Daniel Kahn Gillmor for encouraging this work and providing excellent ideas and expertise. You taught me a lot. Thank you to the Speech, Privacy, and Technology Project at the ACLU for the partial funding of this work in 2017. Thank you to Anna Lysyanskaya for recommending me for the ACLU program. Thank you to William Fleischman for encouraging me to write this paper, and for helping me with the citations. And finally, a big thank you to all of my family for your feedback and suggestions.

REFERENCES

Alexander, M. (2010). *The New Jim Crow: Mass Incarceration in the Age of Colorblindness.* New York, NY: The New Press.

Danner, M., VanNostrand, M. & Spruance, L. (2016). Race and gender neutral pretrial risk assessment, release recommendations, and supervision: VPRAI and PRAXIS revised. *Luminosity, Inc.* Retrieved from https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/race-and-gender-neutral-pretrial-risk-assessment-release-recommendations-and-supervision.pdf

Defender Association of Philadelphia (2018). Press Release: Risk Assessment. Retrieved from https://www.philadefender.org/risk-assessment/

DiFonzo, J. (2005). The Crimes of Crime Labs. *Hofstra Law Review*, 34(1) 1-12. Retrieved from https://scholarlycommons.law.hofstra.edu/cgi/viewcontent.cgi?article=2372&context=hlr

Ensign, E., Friedler, S., Neville, S. Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research Conference on Fairness, Accountability, and Transparency*, 81, 1-12. Retrieved from https://arxiv.org/pdf/1706.09847.pdf

Goel, S., Shroff, R., Skeem, J, & Slobogin, C. (2018). The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment. *Social Science Research Network* [SSRN]. Retrieved from https://ssrn.com/abstract=3306723

Human Rights Watch (2018). China: Big Data Fuels Crackdown in Minority Region. Retrieved from https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region

Johndrow, L. & Lum, K. (2017). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1), 189-220. Retrieved from https://arxiv.org/pdf/1703.04957.pdf

Kaufman, B., Buskey, B., Goodman, R., Eidelman, V., Woods, A., & Bibring, P. (2017). Brief of *Amici Curiae* In support of Defendant*. American Civil Liberties Union* [ACLU], Case No. F071640. Retrieved from https://www.aclu.org/sites/default/files/field_document/2017-09-14_billy-ray-johnson_amicus-full_accepted.pdf

Langan, P. (1995). *The Racial Disparity in U.S. Drug Arrests.* Washington, DC: Bureau of Justice Statistics, U.S. Department of Justice. Retrieved from https://bjs.gov/content/pub/pdf/rdusda.pdf

Lum, K. & Isaac, W. (2016). To predict and serve? *Significance*, vol. 13, no. 5, October 2016. http://doi.org/10.1111/j.1740-9713.2016.00960.x

Makarios, M., Steiner, B., & Travis III, L. (2010). Examining the Predictors of Recidivism Among Men and Women Released from Prison in Ohio. *Criminal Justice and Behavior*, 37(12), 1377-1391. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0093854810382876

Marshall, T. (1999). *Restorative Justice: An Overview.* London: Home Office. Retrieved from http://www.antoniocasella.eu/restorative/Marshall_1999-b.pdf

Mettler, K. (2017). How a lab chemist went from 'superwoman' to disgraced saboteur of more than 20,000 drug cases. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2017/04/21/how-a-lab-chemist-went-from-superwoman-to-disgraced-saboteur-of-more-than-20000-drug-cases

Mohler, G., Short, M., Malinowski, S., Johnson, M., Tita, G., Bertozzi, A., & Brantingham, P. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1-12. Retrieved from https://escholarship.org/content/qt1br4975j/qt1br4975j.pdf

National Institute of Justice [NIJ] (2014). *Predictive Policing*. Washington, DC: National Institute of Justice. Retrieved from https://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/Pages/welcome.aspx

PredPol (2020). Overview. Retrieved from https://www.predpol.com/about/

PredPol (2020). Proven Crime Reduction Results. Retrieved from https://www.predpol.com/results/

Puente, M. (2019). LAPD to scrap some crime data programs after criticism. *Los Angeles Times.* Retrieved from https://www.latimes.com/local/lanow/la-me-lapd-predictive-policing-big-data-20190405-story.html

Russell, S. & Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*. Prentice-Hall. Retrieved from https://pdfs.semanticscholar.org/bef0/731f247a1d01c9e0ff52f2412007c143899d.pdf

Saunders, J., Hunt, P., & Hollywood, J. (2016). Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology*, 12(3), 347-371. Retrieved from https://link.springer.com/article/10.1007/s11292-016-9272-0

Sewell, A. & Jefferson, K. (2016). Collateral Damage: The Health Effects of Invasive Police Encounters in New York City. *Journal of Urban Health*, 93(1), 42-67. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4824697/

Shaer, M. (2016). The false promise of DNA testing. *The Atlantic*, June 2016. Retrieved from https://www.theatlantic.com/magazine/archive/2016/06/a-reasonable-doubt/

U.S. Const. amend. I, VI. Retrieved from https://www.law.cornell.edu/constitution/index.html