**An Evaluation of Machine Learning in Criminal Law:**

Mohammed Khadadeh

Santa Clara University

mkhadadeh@scu.edu

Natalie Van Maerssen

Santa Clara University

nvanmaerssen@scu.edu

Samuel Vivian

Santa Clara University

svivian@scu.edu

**Abstract**

With machine learning solutions finding their way into all sectors, including criminal justice, fairness when it comes to these algorithms has become paramount. The purpose of this paper is to look into whether machine learning algorithms are ready to take on the role of deciding whether or not a defendant should be imprisoned pretrial. We look at how the relationship between AI and Law has evolved over history, and why its integration is desirable. We then look at data from arrests made in traffic to discuss the problems with proxy data which is so often used to train these defendant risk assessment tools like COMPAS. We then conclude with a discussion of the problems that arise when AI is used for criminal law to explain why we strongly believe that AI is not yet ready to take on criminal justice.

**Introduction**

Based on our current system of law and how advanced AI currently is, we do not believe AI is prepared to integrate with the law - at least not yet. Most of these algorithms today use proxy data such as arrest info because court data is extremely hard to find. For example, “Some

states share information about defendants, but rarely collect and share information about judges and other court actors. Most jurisdictions do not collect important information about bail hearings, including the amount of bail the prosecution and defense requested, the arguments that the attorneys made, the reasons why a judge imposed bail, or whether a defendant’s ability to pay was assessed at any point in the process. [2]” So as it is, it is extremely challenging to accurately and correctly train AI because the information needed to do so is limited. But even assuming that information is available, we have to consider algorithmic fairness.

Algorithmic fairness has been defined in many ways, usually in terms of satisfying certain fairness criteria. An important note to make here is that it is impossible to fulfill some of the criteria together [3]. Depending on the fairness criterion an algorithm is trying to satisfy, it can be evaluated mathematically to measure its degree of fairness. Usually, these evaluations are done in the form of probability equations. To fulfill the fairness criterion of “statistical parity,” for example, the algorithm’s results must satisfy:

Pr[Y=y|G = g] = Pr[Y=y]

where Y is the algorithm’s output and G is a set of groups, such as race or gender [6]. In this case, G could be sex, where 1 = female and 0 = male. In this case, statistical parity is saying that the probability of a candidate getting assigned a specific result y given their sex should be equal to the probability of them being assigned that same result without taking their sex into account [6]. We will discuss the fairness criteria we will be using for AI and law further in this paper.

**Why we would want AI for Law**

The main reason AI could be beneficial to the legal system is to achieve efficiency and uniformity. Human bureaucracy is slow and, with AI, these issues could be solved almost instantaneously. This means there would be no need to train people to do all of the jobs involved, no need to monitor their performance, no need to provide safe workspaces and health insurance and as a result, the cost reduction would be enormous. If it were possible to eliminate bias from AI it would also be preferable in that respect as humans often display discrimination and cognitive bias. Although it would be very difficult it may be possible to code in some level of empathy and case-by-case understanding to an AI. If this were the case AI would become far preferable to humans having the empathy of a human without any of the bias [1]. In theory an algorithm in and of itself could be made without any bias as long as the humans who created it were able to do so without inadvertently including their own biases.

With this theoretically very-possible algorithm, we can now have a perfectly unbiased system in which the algorithm is able to process the data perfectly without subconsciously projecting its own bias on its decisions as all humans often do. Decisions would be made with incredible efficiency in a completely unbiased way. This would then be the perfect solution for solving issues of prejudice in the legal system, at least insofar as the decision making surrounding sentencing goes. There would be no inherent bias in the legal system as with this utopic algorithm, histories of systematic prejudice and bias would be irrelevant, and only the actions of an individual would be grounds for incarceration. At the current time, this is not the case.

**AI meets the Law**

The field of AI is between fifty and seventy years old [5], but the intersection between machine learning and criminal law is much newer. Countries like France, Germany, and the Netherlands have a civil law system, meaning they are rule-based, whereas countries like the U.S. use common law, reaching decisions by analyzing precedents set by previous cases. So, in the Netherlands for example, each crime is given a specific punishment. Extenuating circumstances may be taken into account in trial, but focus is placed on the previously determined punishment for that crime. In the U.S. though, more focus is placed on precedents set by earlier cases. So if crime X is punishable by up to 20 years in prison, that doesn’t necessarily mean whoever commits crime X will go to prison for 20 years.

In the early 1980s, researchers began experimenting with civil law AI. One of the earliest examples of this was the section of the British Nationality Act developed with logic programming by Marek Sergot, Robert Kowalski, and others at Imperial College London. In 1984 however, writer Anne Gardner wrote her doctoral dissertation on the issues that could arise when the programmer ran out of rules [5]. Though civil systems focus on rules, they still take precedents into account. But how could precedents be translated into code? This turned out to be one of the many issues faced by the team at Imperial College London. In fact, their biggest challenges included “the open textured nature of legal predicates and the difﬁculties in modeling negation, exceptions, and counterfactual conditionals” [5]. In other words, legal precedents are not black and white – they are very convoluted, and thus incredibly challenging to translate into code. Negations, exceptions, and counterfactual conditionals – conditionals with false ‘if’ statements – are also very difficult to properly teach a machine.

These challenges helped shift the focus towards common law systems, which emphasized case decisions over rules and therefore seemed much more complex for machine learning. HYPO was one of the first successful CBR (case-based reasoning) systems, and “one of the pioneering systems in CBR in general” [5]. Named for its original purpose – creating hypotheticals – HYPO evolved quickly and ultimately had huge influences on groundbreaking projects like CABARET, an AI used for tax law that became the first truly hybrid RBR and CBR system. Since then, work on hybrid RBR and CBR systems has skyrocketed, and the field of AI and Law in general has blossomed into an enormous international community [5].

However, the challenges remain - especially with regards to fairness. COMPAS is an algorithm used as a risk-assessment tool (RAT) before trials. It is designed to determine how risky a defendant is - that is, whether they are likely to reoffend or commit other crimes, are they likely to skip their court hearings if they’re released on bail, etc. COMPAS takes over 100 factors into account, including criminal history, age, housing and employment status, and sometimes gender or race to return a number 1 through 10. This number dictates how risky the defendant is supposed to be, and, though not used directly in the trial or sentencing, it is definitely taken into consideration [6]. Since RATs oftentimes take gender and race into consideration, they have been known to clash with the Fifth and Fourteenth Amendments, which guarantee equal protection under due process of law in both federal and state governments [4]. Analyzing data from courts in Broward County, FL, revealed that black defendants are far more likely to be labeled risky than white defendants, and among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be labeled as risky. [6]” Evidence also shows that RATs sometimes overestimate women’s likelihood to reoffend compared to men’s [4].

Title VI of the Civil Rights Act of 1964 stipulates that “all programs or activities that receive federal funding may not perpetrate or perpetuate discrimination on the grounds of race, color, or national origin. [4]” In other words, any time a court uses information from a RAT, it has to make sure the defendant is not being discriminated against based on race - which COMPAS seems to do. The Equal Protection Clause of the Fourteenth Amendment makes it unconstitutional to discriminate based on gender, which COMPAS also seems to do [4]. So, it seems COMPAS is not only algorithmically unfair - it can even yield unconstitutional results rooted in gender and race discrimination. One of the biggest criticisms of COMPAS is its racial inequality, evident by a much higher rate of false positives with black defendants than with white defendants [6]. Ultimately, while algorithms like COMPAS and HYPO are quickly evolving and improving, they are currently not balanced enough to accurately and fairly sentence defendants, especially since court information is so challenging to find.

**Empirical Analysis of Arrest Data**

It is extremely difficult to get access to court record data for this kind of study, especially when the study is trying to put the court system under scrutiny [2]. We have chosen to analyze arrest data because it is often used as proxy data for machine learning solutions to the question of pretrial detainment. In fact, COMPAS uses prior arrests and arrests of friends and family members as proxies for determining a risk score [3]. Using proxy data for training is usually chosen when the real data is difficult to acquire. Court data is almost impossible to acquire if the people working on the machine learning solution do not work for the court [2]. The problem with using arrest data is that the choices that lead to arresting someone are very different from the choices that are necessary to detain them as they await their trial. This introduces measurement bias into algorithms like COMPAS which primarily rely on proxy data [3].

The following tables show arrests made by officers pulling over cars in California. Each table shows the data for a specific city. Data for these tables comes from databases containing millions of traffic stops in the United States. The databases were made public by E. Pierson et al. [7]. While many other cities from California had their databases made available, only Oakland, San Diego, San Francisco, San Jose, and Stockton showed the race of the driver along with whether an arrest was made. Note that the Arrest Rate is calculated as Pr[Driver is arrested | Driver is pulled over, Driver’s race].

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 613 | 8099 | 7.57% |
| black | 11442 | 78925 | 14.50% |
| hispanic | 2610 | 26257 | 9.94% |
| other | 321 | 4498 | 7.14% |
| white | 1204 | 15628 | 7.70% |

Table 1: Oakland arrests made in traffic based on race, between 2013 and 2017

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 309 | 32541 | 0.95% |
| black | 857 | 42705 | 2.01% |
| hispanic | 1714 | 117083 | 1.46% |
| NA | 4 | 1234 | 0.32% |
| other | 160 | 27238 | 0.59% |
| white | 1771 | 162226 | 1.09% |

Table 2: San Diego arrests made in traffic based on race, between 2014 and 2017

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 1221 | 157684 | 0.77% |
| black | 3540 | 152196 | 2.33% |
| hispanic | 2284 | 116014 | 1.97% |
| other | 910 | 106858 | 0.85% |
| white | 3970 | 372318 | 1.07% |

Table 3: San Francisco arrests made in traffic based on race, between 2007 and 2016

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 705 | 16062 | 4.39% |
| black | 1334 | 13538 | 9.85% |
| hispanic | 7381 | 79885 | 9.24% |
| NA | 489 | 5485 | 8.92% |
| other | 725 | 11523 | 6.29% |
| white | 1729 | 26341 | 6.56% |

Table 4: San Jose arrests made in traffic based on race, between 2013 and 2018

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 52 | 3787 | 1.37% |
| black | 166 | 10870 | 1.53% |
| hispanic | 233 | 16573 | 1.41% |
| NA | 5 | 206 | 2.43% |
| other | 11 | 1640 | 0.67% |
| white | 99 | 8553 | 1.16% |

Table 5: Stockton arrests made in traffic based on race, between 2012 and 2016

If we were to combine all the previous tables, we would get this table of sums:

|  |  |  |  |
| --- | --- | --- | --- |
| Subject Race | Arrests Made | Pulled Over | Arrest Rate |
| asian/pacific islander | 2900 | 218173 | 1.33% |
| black | 17339 | 298234 | 5.81% |
| hispanic | 14422 | 355812 | 4.05% |
| NA | 498 | 6925 | 7.19% |
| other | 2127 | 151757 | 1.40% |
| white | 8773 | 585066 | 1.50% |

Table 6: Sums of arrests made and people pulled over from Tables 1-5.

Table 6 has multiple problems. At first glance, summing up the data from Tables 1-5 shows the difference of arrest rates between white and black drivers is only 4 percentage points, which does not seem like much. However, the data is skewed due to Simpson’s Paradox, the idea that trends in data between different subgroups change when the data is aggregated [3]. Each of the subgroups - the cities - in this case has its data collected over a different period of time. Some of the cities also do better than others. In San Diego, black and white arrest rates do not differ by more than 1%. The same cannot be said for Oakland, where black people are arrested at a rate of 14.50% while white people are arrested at almost half that rate.

Another issue with this data is representation bias. These arrests are only being made when a car is stopped in traffic. The cars to stop are completely up to the police officer, and because of that, sometimes racism from the police can seep into the data. Consider the number of cars pulled over in San Jose from Table 4. Hispanic people were pulled over 79885, which is more than twice the amount of cars pulled over from any other race on the list. Even so, their rate of arrest is very close to black people, whose number of cars pulled over is just 13538. Compare this data with population demographics from the US Census Bureau [8]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Subject Race | Drivers Pulled Over (Over the 6 year period from Table 4) | Average number of drivers pulled over per year | Population estimate [8] | Percent of population pulled over per year (on average) |
| asian/pacific islander | 16062 | 2677 | 366824 | 0.73% |
| black | 13538 | 2256.33 | 30654 | 7.36% |
| hispanic | 79885 | 13314.17 | 326974 | 4.07% |
| other | 11523 | 1920.5 | 31676 | 6.06% |
| white | 26341 | 4390.17 | 265667 | 1.65% |

Table 7: Comparing the data from Table 4 and the US Census

While the US Census data is not from the same year range as the data from Table 4, it still holds population data from a close time period. The numbers here show a large disparity in the average number of cars being pulled over per year given the race of the driver. That is, Pr[Being pulled over | Race] is widely different between hispanic, black, and white people. The representation bias in the arrest data in Table 4 is very large, as a Hispanic driver is more than twice as likely to be pulled over than a white driver, and a black driver is more than four times as likely to be pulled over than a white driver.

The data gathered above also has a temporal bias. Tables 1,2,4, and 5 are results of data collected over a timespan of no more than 6 years, while Table 3’s results come from 10 years worth of data. The rates in Table 3 cannot be held with the same weight as those from Tables 1,2,4, and 5. The temporal bias makes it even less helpful to create an aggregate table like Table 6, because the data was collected over different time periods, each with its different sets of regulations and cultural biases.

While there are many other biases in this data we could discuss, the point of this analysis is to realize how fallible this kind of arrest data is. Traffic stops are one of the most common ways for people to interact with the police [7], and if arrests are being used as proxies for pretrial detainment algorithms, it makes sense that these arrests could skew that data. That is without even beginning to talk about racial biases inherent in some other kinds of arrests. The data has so many possibilities to be skewed that using it to train a machine learning algorithm like COMPAS to make decisions on very loosely related matters can lead to devastating results. COMPAS is meant to analyze the risks of not detaining a defendant before their trial, and using biased arrest data as a proxy will make sure that the bias seeps into COMPAS. Consider the following table of risk scores given to different ethnicities according to COMPAS [9]:

|  |  |  |  |
| --- | --- | --- | --- |
| Ethnicity | Risk of Failure to Appear | Risk of Recidivism | Risk of Violence |
| African-American | 4.62 | 6.31 | 4.77 |
| Arabic | 2.68 | 3.84 | 2.88 |
| Asian | 2.5 | 2.57 | 2.05 |
| Caucasian | 3.13 | 3.59 | 2.6 |
| Hispanic | 2.67 | 3.31 | 2.57 |
| Native American | 3.19 | 5.29 | 3.67 |
| Oriental | 2.46 | 3.23 | 2.54 |
| Other | 2.07 | 2.87 | 2.59 |

Table 8: Average risk scores (on an integer scale of 1 to 10; 1 being low risk and 10 being high risk) given to members of different ethnicities by COMPAS.

Table 8 shows just how badly skewed COMPAS got because of being trained on biased proxies. African-Americans had an average risk score at least one point higher than any other group. Considering the fact that COMPAS ranks on an integer scale, this would mean that African-Americans would on average score in the medium and high risk ranges, (5-10) whereas most of the other groups would score in the low and medium risk ranges (1-5). Native American risk of recidivism also seems uncharacteristically high for where the rest of the data is. The biases inherent in the data make it very difficult to come up with a fair and accurate algorithm.

**The Problems with AI and Law**

There are many issues that come with integrating AI and law. These issues range from the reliability and accuracy of the algorithms to issues surrounding the ways in which society could hypothetically integrate or attempt to integrate AI into the judicial system in the future. Prior to understanding these issues, it is important to understand the difference between equitable and codified law. Codified law is the quantifiable rule-based side of the law, with strict black and white rules and data which lends itself well to machine learning. Equitable law takes into account case-by-case scenarios and allows for human understanding and emotion to influence decisions surrounding incarceration and other legal punishment [1].

If the algorithm needs to be fair but also treat individuals on a case-by-case basis, it needs to conform to the “conditional statistical parity” criterion. Conditional statistical parity states that people in any group g*,* from the set of groups G*,* should have likely probabilities of being assigned a specific outcome Y = y, given a set of factors that are deemed to be legitimate to be used for that decision L [3]. This can be represented mathematically as:

Pr[Y=y|L,G=g] = Pr[Y=y|L] [6].

In the case of the data we have analyzed previously, COMPAS’s risk scores would be the random variable Y. COMPAS takes in 137 features to come up with its risk scores, but these are not publicly available [3]. What is known is that COMPAS uses arrest data as a proxy, so arrest data would be considered a “legitimate factor” in the set of factors L. We do not have access to other COMPAS data, so we will consider previous arrests as the only legitimate factor here. Finally, in this case, the set G corresponds to the different races. The condition can be rewritten as:

Pr[Risk score > r|Number of previous arrests, Race] = Pr[Risk score > r|Number of previous arrests]

The point of this would be to say that race will not have an effect on the probability that someone gets a specific risk score, because that risk score is reliant only on legitimate factors. However, analysis of arrest data shows that the data is heavily skewed by race. For example, Table 1 shows that in Oakland, California, black people are twice as likely to get arrested as white people. With that racial bias baked into the data, considering the number of previous arrests as legitimate is comparable to saying that race is a legitimate factor in producing these risk scores. When a machine learning solution takes biased data in, it will make sure that all results are produced with that same bias.

There are two ways to ensure fairness in this scenario. The first is to ensure that the training data is not biased, which is almost impossible given the racial stereotypes still present today in the law. The second is to accept lower fairness scores. As long as Pr[Risk score > r|Number of previous arrests, Race] for each race is no more than x different than another race, the algorithm can be considered “fair enough.” However, a false prediction (That is, giving someone a score that is not indicative of their unknown true risk) is very costly in this scenario. A false positive means the defendant’s record is permanently stained by an unjust imprisonment. A false negative means that a dangerous person is sent back into society where they could cause more damage. If an algorithm is tasked with such a burden, “fair enough” would be satisfied when the difference x is very close to zero.

The biggest issues aside from those regarding fairness, are the issues that come from further implementation of AI, increasing the extent of its own implementation. That is, as AI is included more into the legal system it will emphasize codified justice over equitable justice and slowly change the laws and eventually also the societal values surrounding the balance therein. This will then lead to a further increase in the implementation of AI and the cycle continues. For example in the case of a murder, a human judge might reduce the sentence or negate it entirely if it could be proved to be accidental or in self defence. AI and the resulting codified justice favors a more direct cause and effect law system. Murder in cold blood is treated the same as murder in in self defence-- murder is murder and the result is the given sentence for murder, whether that be indefinite incarceration or death.

It is also quite possible that early success in the implementation of AI-- that is seemingly satisfying, accurate, and unbiased convictions with a smooth integration-- could lead to false trust in its validity, safety, and long term success, leading to an increase in the velocity of this implementation. In addition, Re et al. have analyzed the theoretical issues of implementing AI in the future in our society. The first proposed issue is that of Incomprehensibility. Due to the fact that a machine learning or neural net’s decision-making process is essentially a black box, no one would have any logically grounded reason or any causal inferences as to why a given decision would be made. As such, decisions about incarceration would be made without any provided reasoning which could lead to a reduced sense of accountability in our legal system. This could become extremely frustrating to the general populous and would make it very difficult to argue for or against the decisions that the algorithm was making and therefore argue for or against the use of said machine learning in the legal system. The second major issue is that of datafication. As a result of a lack of incomprehensibility, the datafied nature of an AI-driven legal system would shield it from legitimate criticism due to a complete lack of the ability to analyze the decision-making process. Because of this, inherent bias in the algorithms would be left unchecked causing many of the same problems that the current human lead judicial system already has. As stated earlier, it could even cause increased biases in the system. Datification could also lead to a strong emphasis on the codified side of law [1].

The next issue is disillusionment. Companies who create these algorithms might end up being motivated by only profit and not consider other issues. They could also criticize the human system leading to a further increase in the implementation of AI leading to more of the other issues. Over time this could cause an erosion of confidence in our legal system. In a world where judges are openly criticized, the type of person who aspires to become a judge may change leading to differently motivated judges. This could also weaken the power of practicing attorneys and slowly humans would be removed altogether in a cyclical fashion. Finally, this could lead to alienation. As human participation decreases and the system becomes unbalanced in favor of codified justice, there may be an extreme lack of civic engagement irreparably ruining the desired balance of codified and equitable justice [1].

**Conclusion**

As things stand today implementing artificial intelligence and machine learning into the judicial system is dangerous in many ways. The problems with the current technology is that the potential problems and issues far outweigh the benefits. While AI in law would no doubt cause a marked increase in efficiency, the other major benefit would be the elimination of bias. As it stands, machine learning needs a large amount of preexisting data for training. The issue is that most if not all of available preexisting data is contaminated by an endless history of human bias and prejudice. This essentially means that at best, an AI trained on this data would be equally biased to the current human system if not far more biased. A machine learning algorithm will compound bias after using its own decisions to further train itself in a cycle of increasing biases [3]. In addition we could have any combination of the issues that come with the implementation and integration of AI in the legal system, including but not limited to: a lack of equitable justice in favor of codified justice, incomprehensibility, datafication, disillusionment and alienation.

All in all, with current technologies and human biases, an implementation of AI in the legal system would result solely in an increase in efficiency and likely a decrease in accuracy paired with a compiled increase in biased decisions, which would lead to the eventual collapse of the legal and governmental system as we know it. While the current system is far from perfect, we recommend that the technology surrounding AI needs to advance significantly before it can be safely implemented in this way. In addition it would be ideal that humans as a whole make significant progress dealing with and moving past their own prejudice as well so that the data that these algorithms would be based upon could have less inherent bias. This is by no means to say that AI should never be used in the judicial system, however it is a warning-- extreme care and analysis must be used when and if the time comes.

**References**

[1] R. Re and A. Solow-Niederman, "Developing Artificially Intelligent Justice", *Stanford Technology Law Review*, vol. 22, no. 2, pp. 242-289, 2019. Available: https://ssrn.com/abstract=3390854.

[2] C. Barabas, J. Rubinovitz, C. Doyle and K. Dinakar, "Studying Up: Reorienting the study of algorithmic fairness around issues of power", *FAT\* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 167-176, 2020. Available: https://dl.acm.org/doi/abs/10.1145/3351095.3372859.

[3] N. Mehrabi, F. Morstatter, N. Saxena and K. Lerman, "A Survey on Bias and Fairness in Machine Learning", *arXiv.org*, pp. 1-31, 2020. Available: https://arxiv.org/abs/1908.09635.

[4] M. Sun and M. Gerchick, "The scales of (algorithmic) justice", *AI Matters*, vol. 5, no. 2, pp. 30-40, 2019. Available: 10.1145/3340470.3340478.

[5] E. Rissland, K. Ashley and R. Loui, "AI and Law: A fruitful synergy", *Artificial Intelligence*, vol. 150, no. 1-2, pp. 1-15, 2003. Available: 10.1016/s0004-3702(03)00122-x.

[6] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel and A. Huq, "Algorithmic Decision Making and the Cost of Fairness", *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, 2017. Available: 10.1145/3097983.3098095.

[7] E. Pierson et al., "A large-scale analysis of racial disparities in police stops across the United States", *Nature Human Behaviour*, 2020. Available: 10.1038/s41562-020-0858-1 [Accessed 21 May 2020].

[8] U.S. Census Buereau, "U.S. Census Buereau QuickFacts: San Jose city, California", U.S. Census Buereau, Washington, D.C., 2020.

[9] J. Larson, S. Mattu, L. Kirchner and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm", *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed: 21 May 2020].